36 Critical Thinking about Research

CHAPTER OUTLINE AND LEARNING OBJECTIVES

36.1 Selection Bias *p. 723*
Give some examples of studies that might suffer from selection bias.

36.2 Causality *p. 724*
Understand the difference between correlation and causation.

36.3 Statistical Significance *p. 731*
Understand how researchers decide whether their results are meaningful.

36.4 Regression Analysis *p. 732*
Understand how regression analysis can be used for both estimation and testing.

Throughout this book we have highlighted the many areas in which economists use data and statistical methods to answer questions that are important to households, businesses, and government policymakers. Some of these questions are narrow: What happens to the sales of ketchup when the manufacturer raises its price? How much will charging a small fee for a vaccine in a developing country affect vaccination rates? Others are much broader: Will a large unexpected fall in housing prices have a substantial effect on household consumption? What happens to employment if we raise the minimum wage? These are all questions that we can begin to answer with economic theory, as you have seen in this text. To get quantitative answers to questions like these, we need to use statistical methods to look at real world data. In this chapter we provide an introduction to the tools economists and other social scientists use to look at data. We will focus both on the standard techniques used and on some of the most common pitfalls associated with using data to answer complex questions.

The statistical tools that economists use to analyze issues are an important part of the discipline. If you go on in economics, you will learn much more about these tools. For those of you who do not continue to study economics, we hope the introduction here will allow you to be a more discriminating consumer of the economic research that you see described in the media and elsewhere.

The techniques you will learn in this chapter are used in many fields other than economics. Psychology, political science, some historical research, some sports research, and some medical research also use these techniques.

Selection Bias

36.1 LEARNING OBJECTIVE

Give some examples of studies that might suffer from selection bias.

We all know that people slow down physically as they age. World records in track and field are not set by 45-year-olds. Few professional baseball players continue to play after the age of 45. Yet consider this: In the 2013 Chicago marathon, the average time of the 30- to 39-age group for men was 4 hours and 17 minutes, which was essentially the same as the average time of the 40- to 49-age group for men of 4 hours and 18 minutes. What do we make of this? Should we conclude, for example, that in the marathon there is essentially no slowing down in the 10-year age interval between the mid-30s and the mid-40s?

Or say you came across a study that randomly sampled 1,000 70-year-old men and 1,000 90-year-old men and measured their bone density. Can we compare the average bone density of the 70-year-olds to that of the 90-year-olds to estimate how much bone density on average declines with age?

The answer to both questions is no. In both cases, there is a substantial likelihood of **selection bias**, which results in unreliable answers. There are many aged 30–39 ham-and-eggers running the Chicago marathon, but many fewer casual runners aged 40–49. Many people aged 30–39 run for fun, to impress a friend, or to pay off a bet. Many of these casual runners have probably selected out by age 40. Moreover, one reason people select out or stop running is that they discover they are not good runners. As a result, the average runner left in the age 40- to 49-interval is likely to be a better runner than the average runner in the age 30- to 39-interval. It is thus not surprising that there is little change in the average times between runners in the two age intervals, but this says nothing about how fast a *particular* runner slows down with age. We are in some sense comparing apples and oranges in looking at the two groups.

selection bias Selection bias occurs when the sample used is not random.

Selection bias would also exist in the bone density study. There are fewer 90-year-old men than there are 70-year-olds. Those with lower bone density at age 70 are more likely to have fallen, broken a hip and passed away. The men left in the population age by 90 disproportionately will thus consist of those who at younger ages had higher bone densities. So it would not be sensible to compare the average bone density of the two samples. This comparison would tell us nothing about how bone density changes with age for a particular person.

The type of selection bias in these two examples is also called **survivor bias**, for obvious reasons. The more fit in the populations have survived, so there is a bias in comparing younger and older groups. Similar problems arise in financial markets when we make inferences about corporate returns in the general market from a population of firms that has survived in the market for a long period. Firms that survive are typically different, generally more successful, than the average firm. Apple, which has survived for a number of years, is surely different in its ability to deliver innovative products that people want than the average company.

survivor bias Survivor bias exists when a sample includes only observations that have remained in the sample over time making that sample unrepresentative of the broader population.

The problem of selection bias pervades many studies in economics (and other disciplines). In recent years there has been considerable interest in trying to understand and improve educational outcomes in the United States. In many areas, charter schools have grown up in part to experiment with alternative educational methods. Charter schools are publicly funded, providing free education to their students, but operate independently of the traditional school district and thus have more autonomy in terms of choices around teacher selection, school hours, and pedagogy. Naturally, there has been considerable interest in how these different charter schools are performing. It might occur to you that one way to answer this question is to compare the scores of students in charter versus traditional schools in an area on the common mastery tests now given across all schools in the United States. Indeed, we see comparisons of this sort often in local newspapers, but here too in making this comparison, you would be running into the problem of selection bias.

Where does the bias come in here? In most charter systems, students are randomly chosen to attend the school. You might think this would eliminate the selection bias problem. Unfortunately, that random choice does not fully eliminate the problem. In most charter systems, to be chosen in the lottery you must apply in the first place. Families who apply to a lottery for a charter school may well differ considerably from those who do not apply. Those differences—more attention to education, more organizational skills, and so on—are both likely to matter to educational performance and will be hard for us to observe. In other words, children who apply to a charter school may do better on the mastery tests than the average child, even if he or she does not get chosen to attend the charter! As we will see in a later section, there are ways around this selection problem but they require some ingenuity.

One more example may help to show the range of the issues involved. Many studies in the medical area are aimed at helping us figure out how to live longer and healthier lives. Suppose you were interested in the effect on longevity of exercise. Luckily, you found a long-term study that tracked how often people exercised over many years and found that those people who exercised more also lived longer. Should you conclude that exercise in fact increases life span? The answer is again no. In this example we are comparing a group of people who chose to exercise with a group who chose not to. The fact that a group of people do or do not choose to exercise tells us that they likely differ on many other grounds that might independently affect life span. People who choose to exercise also likely make other healthy choices, most of which will be hard for a researcher to observe. So the longevity edge might come from the fact that one group exercised, whereas the other did not, but it might equally have arisen from the fact that the first group consists of people who make healthy choices while the second does not.

A common problem in many of these cases is that we are comparing groups who not only engaged in different activities, activities whose effect we seek to measure, but people who made different choices. To the extent that those choices reflect group differences that themselves matter to the outcomes we are measuring, we bias (or distort) our results. In the last few years, economists have become increasingly sensitive to the problem of selection bias and have engaged in many creative ways to try to eliminate the bias problem. We will describe some of the solutions to the bias problem later in this chapter. For now, we hope you will look at some of those newspaper headlines with a more skeptical eye!

36.2 LEARNING OBJECTIVE

Understand the difference between correlation and causation.

Causality

As we have seen, selection bias makes it difficult to identify the effect of a treatment on a population. In other words, selection makes it hard to pin down causal effects. Identifying causality is a general issue in data analysis and goes beyond problems that arise because of selection bias. We will consider a number of causality issues in this section.

Correlation versus Causation MyLab Economics Concept Check

Most people who have blue eyes also have light colored hair. Most people who have minivans also have children. Evidence suggests that people who are obese have a disproportionate number of obese friends. What can we conclude from these facts? Do blue eyes cause blond hair? Do minivans cause people to have children? Is obesity contagious, caught from one's friends?

correlated Two variables are correlated if their values tend to move together.

When two variables tend to move together, we say they are **correlated**. If the two variables tend to move in the same direction, we say they are *positively* correlated, and if they tend to move in opposite directions, we say they are *negatively* correlated. In the examples above, the variables in each of the three sets are positively correlated; but correlation does not imply causation. It does not take a degree in biology to know that blue eyes do not *cause* blond hair. Likely evolution has selected simultaneously on these two features causing them to appear together. Dumping a bottle of peroxide on my head, although it will surely make me a blond, will not change my eye color at all! In economics, as well as in other fields, theory is often quite helpful in helping us to differentiate between correlation and causality.

Minivans and children provide another example in which we need to sort out correlation and causation. In the data we see that the majority of minivan owners have children. Clearly minivans do not often cause children to be born ("Might as well have a fourth child. We already own a minivan!"). Here it is likely the causality runs in the opposite direction. Minivans are most attractive to families with children. So having children may indeed cause people to buy a minivan. Think now about why getting the causality right matters. If Japan, for example, wants to increase its very low birth rate, giving everyone a free minivan is not likely to be effective. Minivans do not by and large cause people to want children. Knowing the relationship between minivans and children is clearly relevant to automobile manufacturers who will want to exploit this relationship by focusing their marketing campaigns on families. An increase in average family size causes a shift in the demand for large minivans but the reverse is not the case.

The most complicated of the examples is obesity. Here there are theoretical arguments that support a hypothesis of causality running in both directions. Eating and exercise are social for many people, so having obese (or conversely thin) friends may well have an effect on your own weight. But in some circles at least obesity is a social stigma, and it may well be that being obese limits one's choice of friends. Thus, it is plausible that having obese friends does increase your own chances of being obese, but it is also likely that being obese increases your chances of having obese friends.

Identifying causality is critical for much policy work. Knowing that early exposure to reading is correlated with high adult incomes is interesting. Knowing that it *causes* high incomes suggests a policy intervention. Much empirical work in economics is concerned with trying to determine causality, given how important it is for policy issues. Let us consider a few ways researchers have used to identify causation.

Random Experiments MyLab Economics Concept Check

The gold standard for empirical work is the random experiment that many of you will be familiar with from medical research. If a research team is trying to decide if a particular drug helps in treating some form of cancer, for example, a standard protocol is to randomly divide the patients afflicted with the disease into two groups, provide one group with the drug, and give the other a placebo. With a large enough group, and enough time, one should be able to tell if the drug is effective. (Of course, there is much non-human pretesting for safety reasons). Notice in this protocol that we did not select our samples by asking people to choose whether they wanted to take the drug or not (all agreed to the drug). Indeed, part of a standard medical protocol is that patients do not know which group they are in during the experiment. In this type of experiment, we have no selection issue, since there has been no user selection.

Experiments of this sort are also run in economics and are relatively prevalent in the area of economic development. To give an example, suppose we are interested in the effects of class size on educational achievement, say test scores. Comparing classes with large and small enrollments will clearly not be informative. Among other things, it is well known that classes in more affluent areas have smaller classes than those in poorer areas and that affluence will bring with it many advantages that likely lift test scores. Instead, one could run an experiment that randomly assigns students to classes that differ in enrollments and then later compare test scores for the different groups. If the assignments are truly random, there are no selection issues.

Although random experiments are common, especially in the medical field, they are not always possible to carry out. Suppose we are interested in the link between smoking and cancer. We can, of course, take a large group of mice, randomly divide them into two groups, expose one but not the other to smoke, and see whether the two groups differ in their cancer incidence. As long as our sample is reasonably large, we should be able to see a difference in cancer rates if a causal relationship between smoking and cancer exists. We have done a random experiment just as we described. Notice how this experiment differs from just comparing cancer rates of smokers to non-smokers. People who smoke have *chosen* to smoke and may well have made a number of other choices that could be unhealthy. As hard as we try to control for those smoker/nonsmoker differences, our ability to do so is limited.

For the mice there are no choice problems to worry about. If we find that smoking causes cancer in mice; it remains to determine whether the same holds for people. Clearly, we cannot force a randomly chosen group of people to smoke and then see if their cancer rates differ from that of a control group. For many of the questions economists are interested in, it is difficult to use random experiments. Randomly exposing groups of people to something that is potentially harmful is unethical and would not pass a human subjects protocol review. Even if we are looking at interventions that have only potential benefits and no costs, we still face the problem that the randomly chosen subjects we start out with may decide not to join the study or to leave the experiment early. If this happens, the groups left will no longer be random. When there is some discretion among subjects to either take up a treatment offer or to continue in a treatment over time, selection bias will again potentially creep in to our experiment. What do we do under these circumstances?

Consider a university that has admitted 200 at-risk students from households with low incomes. It has a summer program before college begins to better help prepare such students for college life. The university wants to know if this program improves a student's four-year college performance, say measured by a student's four-year GPA. How might it proceed?

Assume that the university randomly samples 100 of the 200 at-risk students and invites them to attend the summer program at no cost. Say 60 accept the offer and take the summer program. After four years the GPAs of all the 200 students are collected and we learn that the average GPA of the 60 students who took the program was higher than the average GPA of the 140 students who did not take the course. Could you conclude from this that the program had a positive effect? No. Once again, we have a selection issue. While the 100 students offered the program were indeed a random sample, the 60 students who took up the offer were not. Maybe the 60 were on average less talented than the 40 who refused the offer and felt the need to take the program, whereas the more talented 40 did not. Or maybe the 60 were on average more serious students or more organized. However the bias runs, we cannot assume that the 60 students who accepted are a random sample of the 200 initial students. Here we are not even sure if those who accept are better or worse than the non-accepters. That is, we do not know the direction of the bias.

The group of the 100 students initially drawn and invited to join the program was random by design. So after four years we can compare the average GPA of the 100 students who were offered the program to the average GPA of the 100 students who were not. If the program has a positive effect, the first average GPA should be greater than the second. You might think this is an odd process for testing the efficacy of the summer program. After all, 40 of the students whose scores we are looking at did not take the program! If they did not take the program, why are they included in the average GPA along with actual program-takers? We include all students who were made the program offer in our test sample to avoid selection bias. This procedure, which is also used in medical experiments that have patient drop-outs, is called **intention to treat**, but proceeding this way does have a cost.

intention to treat A method in which we compare two groups based on whether they were part of an initially specified random sample subjected to an experimental protocol.

Suppose only 10 students of the 100 took up our offer. In this case, we are comparing two random groups of students, one of which has no one taking the program and the other with 10 in the program and 90 not. With so many non-takers, it will be hard to find any gains from the program. If instead all 100 students invited to the program actually enrolled, clearly we would have more confidence that we could find an effect from the program if any existed. Whatever the case, we need to compare the performance of the 100 offered students with that of the 100 non-offered to avoid selection bias. Notice that intention to treat makes it harder to find results from a treatment and in this sense is a conservative statistical technique. The *Economics in Practice* box describes an experiment run by the U.S Department of Housing and Urban Development (HUD) using randomly assigned housing vouchers to examine the effects of community on household well being. Here the method of intention to treat is used.

Regression Discontinuity MyLab Economics Concept Check

In many situations economists do not answer their empirical questions with random experiments, but rather try to make inferences from market data, data that come out of the everyday transactions and choices individuals make. Using market data has a number of advantages: These data reflect real choices made in everyday life by households. Much of the data are collected as a matter of course by either government or business and so are easily available to researchers, but the fact that the data reflect individual choices, done in relatively uncontrolled settings, makes the identification of causality especially difficult. Carefully designed experiments, on the other hand, are expensive. There are a number of procedures that researchers have used to try to make progress in this area.

The United States has more prisoners per capita than any other OECD country,[1] with roughly two million incarcerated. Many of those released from prison are re-arrested within a short period of time. How does what happens while someone is in prison affect the likelihood they will be re-arrested? Do the conditions in prison affect recidivism rates?[2] Arguments about

[1] The OECD is the Organisation for Economic Co-operation and Development. It consists largely of developed world countries, heavily weighted toward Europe.

[2] This discussion is based on M. Keith Chen and Jesse Shapiro, "Do Harsher Prison Conditions reduce Recidivism? A Discontinuity-based Approach." *American Law and Economics Review* June 2007.

ECONOMICS IN PRACTICE

Moving to Opportunity

It is well known that children who grow up in high-poverty areas on average end up as adults with lower educational attainments, poorer health, lower income levels, and a higher likelihood of being incarcerated at some point in their lives. To what extent are these results attributable to the neighborhoods in which these children grow up, and, relatedly, how much could they be changed by a locational change?

These are the very central policy questions posed by an experiment run by the U.S. Department of Housing and Urban Development in the mid-1990s and recently re-evaluated by a group of economists.[1]

The Moving to Opportunity program offered to randomly selected families living in high-poverty housing projects housing vouchers that they could use to move to lower-poverty neighborhoods. The random granting of the vouchers was a direct attempt to avoid the selection bias problems found in earlier studies of housing and later outcomes. It is easy to see that if we simply look at life outcomes for children whose families move out of high-poverty areas to those who remain in those areas we will have serious selection bias issues. Moving families likely have more access to resources—perhaps ones we cannot observe—and perhaps more initiative or organizational ability than those who stay. Those differences might well have an effect on their children's outcomes independent of the gains from the move. By randomizing the voucher choice, HUD attempted to remove the choice element. Not all families offered the vouchers moved, so the researchers used the intention-to-treat methodology described in the text to control for the potential selection bias.

Early results from the experiment found little results on the economic well-being of moving families, though there were gains in mental and physical health. A longer-term, recently completed study by some of the same authors, which looked at tax data, found substantial effects on income levels of those children who were younger than 13 years of age when their families moved, with average gains of 31 percent higher incomes for the young movers.

CRITICAL THINKING

1. Some of the same researchers whose work is described also did another study looking at the outcomes of households that moved versus those that did not in the general population. To control for selection bias, the researchers compared children of different ages within families to see how much more time in the better neighborhood influenced younger versus older children. How does this attenuate the selection bias issue?

[1] Raj Chetty, Nathaniel Hendren, Lawrence Katz, "The Effects of Exposure to Better Neighborhoods on Children: New Evidence for the Moving to Opportunity Experiment," *American Economic Review* April 2016, 855–902.

this question have been made on both sides. Some argue that harsh conditions reduce recidivism because the worse the conditions, the more incentivized released prisoners will be to stay out of prison. On the other side, harsh prison conditions may increase a taste for violence or reduce a prisoner's future labor market value. This would suggest that harsh conditions increase recidivism.

On questions like this, it is important to bring data and evidence to the table. What happens if we just compare recidivism rates in prisoners from more or less harsh prisons? Here again, identifying causality is problematic. In general, harsher prisons house more serious criminals. So, if we see more recidivism from those coming out of harsher prisons, it could well be that the recidivist traits caused the prison choice, rather than the prison type causing the recidivist traits.

regression discontinuity Regression discontinuity identifies the causal effects of a policy or factor by looking at two samples that lie on either side of a threshold or cutoff.

Chen and Shapiro used an interesting strategy, called **regression discontinuity**, to sort out causality. The design works as follows. Once an inmate is convicted and enters the federal prison system, he or she is given a security score. The score predicts the prisoner misconduct

ECONOMICS IN PRACTICE

Birth Weight and Infant Mortality

Per capita health care costs in the United States are quite high, even relative to other developed countries. Much of these expenditures are concentrated on two quite different parts of the population: The very old and the very young. A central question for public policy is how effective these high expenditures are.

Medical practice often distinguishes between babies who are below 1500 grams and those above. The former are called VLBW babies, for Very Low Birth Weight, and in most hospitals these babies receive extraordinary care at birth and immediately thereafter. This extraordinary care is expensive, and there is good evidence that the hospital bills of the VLBW babies are considerably above those for infants at higher weights. But do these expenditures help?

It is, of course, quite difficult to answer this question just by comparing outcomes, like one-year mortality rates, for babies of different birth weights. We know that low-birth-weight children are at risk. So even if we found that with treatment their mortality rates were high, it would not be informative because that result would not tell us what the mortality rate would be absent extraordinary treatment. A study by Almond, et al. using the regression discontinuity procedure we discussed in the text provided a way around this problem.[1]

As indicated, the label VLBW for babies is assigned for a birth weight less than 1500 grams. Medically, this assignment is a convention and does not reflect any threshold medical condition that turns on and off at above or below this weight. This designation does trigger a set of extraordinary treatments at most hospitals. So we have a perfect setup for a regression discontinuity study: Birth weights and medical conditions around that birth weight are continuous variables while the trigger for treatment is a fixed line. Almond et al. examined outcomes for babies on either side of the line. What did they find? If we compare babies just below the trigger line with those just above, the babies below the line had a 1 percent *lower* one-year mortality rate than the slightly heavier babies. With a base of just over 5 percent for a one-year mortality rate, this is a substantial difference. What caused the difference? The extra medical care given by virtue of the VLBW designation!

CRITICAL THINKING

1. Can you think of another medical designation for which a regression discontinuity technique might be useful?

[1] Douglas Almond, J. Doyle, A. Kowalski and H. Williams, "Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns," *Quarterly Journal of Economics*, 2010.

and security needs. There is no personal judgment in creating this score, which simply adds up points depending on the prisoner's record. The score then determines prison facility based on availability of beds. Scores above six typically go to higher-security (and typically harsher) facilities. Placement also depends on bed space, and this means that prisoners with similar scores may end up in different types of prisons. Regression discontinuity effectively compares outcomes from individuals who are close to either side of a dividing line. In this example, we are effectively comparing recidivism rates for prisoners sent to harsh versus less harsh prisons who were virtually identical on their pre-prison scores. The study in fact found that harsh prisons do not reduce recidivism, but may in fact increase it.

Similar methods have been used in other instances in which the existence of a black-and-white line based on a continuous score for individuals determines whether or not an individual is "treated." Most government programs for unemployment or insurance disability benefits have this property of setting an absolute threshold for receiving treatment, allowing a researcher to essentially use individuals very close to the threshold as a kind of control group. In the *Economics in Practice* box on this page we describe a study of birth weight and infant mortality based on this methodology.

Difference-in-Differences MyLab Economics Concept Check

Another interesting procedure to try to get a better handle on causality in social science studies is called the method of **difference-in-differences.**

difference-in-differences Difference-in-differences is a method for identifying causality by looking at the way in which the average change over time in the outcome variable is compared to the average change in a control group.

Suppose we have a community in which a small nonprofit has run a community gardening program. The group is convinced that this program increases housing values. Someone in the group suggests that they just look at what has happened to housing values in the community in the four years since the program began as a measure of the program's success. It is easy to see that this will not work. Housing prices are quite volatile, moving with the overall level of economic activity in an area. In other words, much of the fluctuations in housing prices have nothing to do with community gardens. Another suggestion might be to compare the housing prices in this community with those in a similar neighboring community without the program, but this procedure too is problematic, as no two communities are exactly alike.

The difference-in-differences method takes a third approach that melds these two ideas. In particular, we try to relate the difference in our community's housing values over time to the difference in a neighboring community's values over the same time (hence the name difference-in-differences). If all of the other factors that affect housing values are the same between the two communities (that's why we have chosen a neighboring community), then this difference-in-differences procedure will show us the effect of the program.

To be clear on what the procedure does, let *pbega* and *pbegb* denote the average housing values in communities *a* and *b* before the garden project began in community *a*. Let *penda* and *pendb* denote the average housing values after four years in the two communities. Then the effect of the garden project on housing values in community *a* is estimated as:

$$\mathit{effect} = \mathit{penda} - \mathit{pbega} - (\mathit{pendb} - \mathit{pbegb})$$

We take the difference in values in community *a* and subtract from it the difference in values in community *b*.

The difference-in-differences methodology is reasonably common in the social sciences. A classic example is presented in the *Economics in Practice* box on the next page, which looks at the effect of the minimum wage. There are pitfalls as well in doing this work, pitfalls that come in part from the difficulties of identifying an appropriate comparison group.

Consider the following example: Stimulated in part by what has been happening to the cognitive functioning of aging professional athletes, especially in football, there has been growing concern among university leaders about the long-term effects of injuries in college sports. Short of banning football, which some would advocate, there have been other suggestions to reduce the incidence of injury, notably requiring better helmets and/or eliminating kickoffs (with the ball always starting on the 20 yard line).

Suppose that several years ago the Ivy League introduced such regulations and that a researcher was interested in seeing whether the regulations in fact had reduced injuries. To test whether the regulations helped, we could compare the average number of injuries per game measured in the year before the new rules, denoted *ybeg*, with injuries in the year after the new rules were instituted, denoted *yend*. But as in the case of housing values, we cannot be sure that nothing happened in the world of Ivy League football other than the rule changes over this period. Maybe the NCAA introduced other rule changes for all the colleges in the country, including the Ivy League colleges, which were designed to lessen injuries, such as telling referees to be stricter. We need a comparison group, a second set of differences.

One possible comparison group might be the PAC-12 conference. Assume that this conference did not introduce the new rules on helmets and kickoffs. Again, we collect data on the average number of injuries per game for the same two years we used in the Ivy League case for this conference, denoted *zbeg* and *zend*. We can then compare the difference between these two values and the difference between the two Ivy League values (difference-in-differences):

$$\mathit{effect} = \mathit{yend} - \mathit{ybeg} - (\mathit{zend} - \mathit{zbeg})$$

By subtracting the PAC-12 difference from the Ivy difference we are controlling for country-wide changes that occurred during the two years. The variable *effect* is then the amount attributable to the Ivy League regulations only.

ECONOMICS IN PRACTICE

Using Difference-in-Differences to Study the Minimum Wage

There is a lively debate among economists and policymakers on the effect of the minimum wage on unemployment. Does raising the minimum wage substantially increase unemployment, particularly for the lower-skilled workers? Or can one legislate wage increases for low-wage workers without a substantial reaction from employers?

One of the first and still classic examples of the difference-in-differences technique described in the text was done by David Card and Alan Krueger in their study of state minimum wage changes.[1]

In the early 1990s New Jersey decided to raise its minimum wage. Although there is a federal minimum wage, many states adopt higher minimums for the firms employing workers within their borders. Card and Krueger decided to survey fast-food restaurants in New Jersey to determine the effect of the minimum wage increase on employment. Fast-food restaurants were an obvious target given their employment of large numbers of unskilled workers. But just looking at New Jersey would not be sufficient. Suppose employment went down after the change. One has no way of knowing what would have happened absent the rule change. After all, employment depends on a number of other factors in the economy.

Here is where difference-in-differences comes in. New Jersey is bordered by Pennsylvania, a state that made no change to its minimum wage law in the period and also has fast-food restaurants. So Card and Krueger added data from these restaurants in eastern Pennsylvania as a comparison. The key measure of effect was the difference in employment changes between New Jersey restaurants and eastern Pennsylvania restaurants over the period in question, the difference-in–differences in short. They found no effect from the law change. Not everyone writing on the topic agrees with that conclusion, but most do agree on the usefulness of the difference-in-differences technique.

CRITICAL THINKING

1. Design another experiment using difference-in-differences to understand the effect of a policy change at your college.

[1] David Card and Alan Krueger, "Minimum Wage and Unemployment: A Case study of the Fast Food Industry in New Jersey and Pennsylvania," *American Economic Review*, September 1994.

This looks neat, but there are several potential pitfalls to this research plan. Most fundamentally, we have assumed that the two-year changes absent the Ivy League regulations are the same for both conferences, but PAC-12 football is not exactly like Ivy League football (ask any serious college sports fan!). Those differences may be important not only in starting levels (which is fine) but in changes over time (which is not fine). PAC-12 football is played at a higher level than the Ivy League, so it could be that the change in its injuries per game over the two years is not a good approximation of what the Ivy League change would have been absent the regulations. Perhaps the PAC-12 coaches pushed their players even harder and this led to increased injuries. If the PAC-12 is not a good comparison group, then difference-in-differences will not work in this case.

One more point to reflect on in this football example. With safer helmets it could be that the players play rougher knowing that they are better protected, and playing rougher, other things being equal, increases injuries. Regulations have the potential to affect behavior in ways not anticipated by the regulators. Some of the original work documenting this effect was done by Sam Peltzman, a Chicago economist, who found that seat-belt laws might perversely encourage people to drive faster than they did without seat belts because they felt safer.[3] In the helmet case, some gains from the physical protection of a helmet might be offset by the behavioral changes it induces in the intensity of play. Economic research is not an easy task, but we hope you can see that it encourages care and creativity!

[3] Sam Peltzman, "The effects of automobile safety regulation," *Journal of Political Economy*, August 1975. More recent work has questioned this result.

Statistical Significance

36.3 LEARNING OBJECTIVE

Understand how researchers decide whether their results are meaningful.

We all know that in tossing a coin there is a 50 percent chance we will get heads. Nevertheless, it is not true that coin tosses always alternate between heads and tails. Sometimes we get two or three heads in a row before a tail shows up. How many heads would we need to get in a row before we started to think that there was something wrong with the coin?

In the coin example we answer this question by thinking about how likely it is that a fair (or normal) coin would give us heads after heads. Two heads in a row is relatively common, happening 25 percent of the time (0.5 times 0.5). Even four in a row sometimes happens (about 6 percent of the time). However, six heads in a row happens only about one in a hundred times. At that point you may be suspicious about the coin tosser and begin to think this is not a fair coin!

In thinking about our results in empirical work in economics we use the same basic logic as we try to figure out what we can conclude from the data we have gathered and the statistical tests we have employed. The key question for the researcher is to figure out if the results he or she has found have occurred "by chance" or if they really mean something. To make that judgement researchers turn to the concept of statistical significance.

Return to the example of the summer program experiment and suppose the GPA difference observed after the program was 0.3 on a 4.0 scale. Can we conclude that the program really had a positive effect on GPA, or is 0.3 so small that it was likely due to chance? A common way of looking at this problem is to begin by assuming that the effect of whatever we are testing, here the summer program, is zero and then ask what is the probability we got the result we did if the true effect is zero. The assumption of no effect is called the *null hypothesis*. In our earlier example, our null hypothesis was that the coin was fair. Here the null hypothesis is that the summer program has no effect on GPA. What is the probability we got a difference in GPA of 0.3 if the null hypothesis is true?

The probability that one got the result that one did if the null hypothesis is true can be computed given certain statistical assumptions. It is called a **p-value.** A small p-value means that the probability is small of getting the result if the null hypothesis is true. If for the 0.3 GPA difference, the p-value was 0.02, this says that there is only a 2 percent chance of getting this value if the summer program truly has no effect on GPA. The term **statistical significance** is commonly applied to a p-value of 0.05 of less. If a p-value is less than or equal to 0.05, the results is said to be statistically significant.

p-value The probability of obtaining the result that you find in the sample data if the null hypothesis of no relationship is true.

statistical significance A result is said to be statistically significant if the computed p-value is less than some presubscribed number, usually 0.05.

Be clear on what we are doing here. We are starting from the premise that whatever effect we are trying to estimate does not exist (is zero). We collect our data and do our calculations to get a particular estimate of the effect we are interested in. We compute the p-value for this estimate, which again is the probability that the true effect is zero given the particular estimate that we obtained. If the p-value is small, usually taken to be less than or equal to 0.05, we conclude that our estimated effect is statistically significant. We have rejected the null hypothesis of no effect.

If you go on in statistics, you will learn exactly how p-values are computed. They depend on the variability of the population being analyzed. Consider the 200 at-risk students in the summer program experiment. Say that they are all identical, meaning that they will all get the same GPA at the end of four years if they don't take the summer program. If some do take the summer program, all those who do will get the same GPA, although this GPA will be different if the program does have a non zero effect on GPA.

We want to test whether the summer program effect is zero. We run the experiment discussed and get a difference of 0.3. Is this difference statistically significant? The answer is obviously yes. If the true effect were zero, everyone would get the same GPA whether they took the program or not, so the difference would be exactly zero. We in fact got a nonzero estimate, and so we are sure that the true effect is not zero. The p-value would be 0.00. In fact in this case we only need two students, one who took the program and one who did not. If the difference in the two GPAs is not zero, then the summer program has an effect. In this case there would be no need to use intention to treat—there are no selection problems because everyone is identical.

Now consider that there is huge variation in the population of 200 regarding what GPA they are going to achieve. Some may turn out to be stars, and some may barely make it through the four years. Whether students take the summer program or not, there will be a huge variation in GPA scores at the end of the four-year period because of the huge variation in the population. We run the experiment and get a difference of 0.3. Is this difference statistically significant? Maybe not if the variation in the population is large. The difference of 0.3 is fairly small, and it could easily be obtained by chance. It just so happened that the particular draw of 100 students led to this outcome, but it may be that a different draw would have resulted in a difference of 0.2. The p-value that is computed for the result of 0.3 would likely be very large, perhaps close to 1.00.

The intuition to take from this discussion is that one has more confidence in results obtained from populations with low variation than from those with high variation. To get potentially significant results from a high-variation population, one needs a large sample size. If we had 2,000 at-risk students, gave offers to 1,000, and got a difference of 0.3, this might be significant. When at the end we take the average of the 1,000 GPAs, the individual student characteristics tend to cancel out in a large sample size, and we can have more confidence that the difference of 0.3 is picking up the summer-program effect. When computing p-values, the size of the sample matters as well as the variation in the population.

36.4 LEARNING OBJECTIVE

Understand how regression analysis can be used for both estimation and testing.

Regression Analysis

The most important statistical tool in empirical economics is regression analysis. If you go on in economics you will see applications of regression analysis in microeconomics and macroeconomics. It can be used to forecast the effect of an increase in prices on the quantity of cat food sold in a community or the effects of a stock market decline on household consumption. Here, we provide you with a beginning sense of what regression analysis is all about.

There is evidence that the economy has an effect on votes for president in the United States.[4] If the economy is doing well at the time of the election, this may have a positive effect on votes for the incumbent-party candidate and vice versa if the economy is doing poorly. This theory suggests that many voters reward or blame the party of the president-in-office for good or bad economic performance while that president is in office. If true, this theory suggests that, all else equal, a president who presides over a strong economy will find his or her political party doing well in the next election.

How might we test this theory using regression analysis? We first need some measure of economic performance. The growth rate of the economy is one common measure of economic strength. So we can translate our theory into a more testable form: we postulate that the growth rate of the economy in the year of the election, denoted g, has a positive effect on the incumbent party's presidential vote share, denoted V. Notice here we have chosen a specific measure of performance—the growth rate—and also a time period—the year of the election. Generally speaking, when we move in economics from a theory to a practical statistical test, we will have some choices to make. In this case we have chosen to measure economic performance by the one-year growth rate.

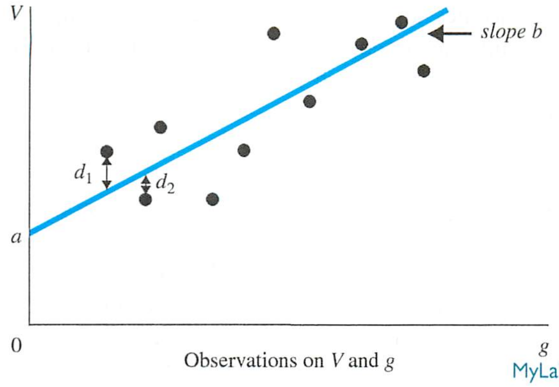
We will look at the way in which the growth rate affects the vote share. In particular, we will assume that

$$V = a + bg \tag{1}$$

If b is positive, this equation says that the growth rate has a positive effect on the vote share, as our theory states. Also, the relationship between V and g is assumed to be linear. If we take a graph with V on the vertical axis and g on the horizontal axis, as in Figure 36.1, the line is straight with intercept a and slope b. The job of regression analysis is to estimate the coefficients a and b and to see in particular if b is positive and if it "matters" in a statistical sense.

Consider how we might determine, or estimate, the values for a and b. U.S. presidential elections are held every four years, and there are data on V going back to the beginning of the country. There are also data on g going back many years. If we consider the period beginning in 1916, there have been 26 presidential elections between 1916 and 2016 so we have 26 data points, or observations, on V and g. We can plot these observations in a Figure like 36.1. In the figure we have plotted 10 hypothetical points for illustration. As drawn, the figure shows that there is a positive relationship between the vote share and the growth rate. It also shows, however, that the data points are not all on the line. If equation (1) were exact, all the points would be on the

[4] See Ray C. Fair, *Predicting Presidential Elections and Other Things*, 2nd ed. Stanford University Press, 2012, for discussion of this.



◀ FIGURE 36.1
Hypothetical plot of points between the vote share and the growth rate.

MyLab Economics Concept Check

straight line. In fact, in the real world equation (1) is not exact. There are other variables that affect votes for president. Some of these variables include other economic measures, such as perhaps inflation at the time of the election. Vote share may also be affected by foreign policy and personal characteristics and views of the people running for office. As a result, the points in the graph of the vote share on the growth rate are not exactly on the line. The job of regression analysis is to find values of *a* and *b* that provide a good fit of the data around the line. Or, in other words, to find the line that best represents the data in the figure.

How is fit determined? What do we mean by the best line? This can be seen in Figure 36.1. Draw a particular line with intercept *a* and slope *b*. For each data point compute the vertical distance between the point and the line. We have done this for the first two points in the figure, labeled d_1 and d_2. We do this for all the points, say the 26 observations between 1916 and 2016. Some values of *d* are positive and some are negative. The larger the distance above or below the line, the worse that particular point fits the line. The distances are usually called "errors" for this reason. The way the fit is determined is first to square each distance. Each squared distance is positive because the square of a negative number is positive. Then we add up all the squared distances, again in our case 26 numbers. Call this sum *SUM*. The sum is obviously a measure of fit. A small value of *SUM* means that the points are fairly close to the line, and a large value means they are not. The fact that squared distances are used means that large outliers (distances) are weighted more than small ones in computing *SUM*.

You can think of regression analysis as doing the following, although in practice finding the right line is done more efficiently:[5] Try a million different pairs of values *a* and *b*, and for each pair compute *SUM*. This gives us a million values of *SUM*. Choose the smallest value. The values of *a* and *b* that correspond to this smallest value are the best-fitting coefficients—the best-fitting intercept and slope. These estimates are called **least squares estimates** because they are the estimates that correspond to the smallest sum of the squared distances, or errors.

least squares estimates Least squares estimates are those that correspond to the smallest sum of squared distances, or errors.

In our theory we focused on the sign of the coefficient of the growth rate: does growth increase vote share? The size of the estimates is often also of interest. If, for example, the estimate of *b* were 1.0, this tells us that an increase in the growth rate of one percentage point leads to an increase in the vote share of one percentage point. This would be a nontrivial effect of the economy on voting behavior. If the estimate of *b* were instead 0.01, politicians would worry much less about how a bad economy was going to affect their votes (in practice the estimate is about 0.67).

Regression analysis is helpful in letting us test our theories. In our voting example, we are particularly interested in whether *b* is zero. If *b* is zero, this says that the growth rate has no effect on votes, and our original theory is not right. To see if we should continue to have confidence in our theory, we need to test whether *b* is zero.

How do we test whether *b* is zero? Here we go back to what we already know from our discussion of statistical significance and p-values. We first postulate the null hypothesis that *b* is in fact zero. We then use regression analysis to estimate *b*, and after this is done we compute the probability (p-value) that we would have obtained this estimate if the truth is that *b* is zero. If the p-value is low, say less than 0.05, we say that the estimate of *b* is statistically significant. We reject the null hypothesis that the growth rate does not affect the vote share, and our confidence in the theory that economic performance affects votes is bolstered.

[5] There are many statistical programs that do this calculation with one simple command, including Microsoft Excel™.

Most theories in economics are more complicated than simply one variable affecting another. In our voting example, as noted, inflation may also affect voting behavior. In this case two variables affect *V*: *g* and inflation, which will be denoted *p*. In this case we could write the voting equation as

$$V = a + bg + cp \tag{2}$$

Equation (2) has two variables that explain vote share plus a constant term. There are now three coefficients to estimate rather than two: *a*, *b*, and *c*. With more than one explanatory variable, we cannot draw a graph as we did previously. However, the fitting idea we introduced works the same way when we add variables. Given observations on *V*, *g*, and *p*, you can think of the analysis as trying a million sets of values of *a*, *b*, and *c* and choosing the set that provides the best fit. For each set of three coefficient values, the predicted value of *V* can be computed for each observation, and the distance for that observation is the difference between the predicted value of *V* and the actual value of *V*. We square this distance, do the same for all the observations, and then sum the squared distances. This gives us a value of *SUM* for the particular set of the three coefficient values. We do this a million times for a million sets of three coefficient values and choose the smallest value of *SUM*. The coefficient values that correspond to the smallest value of *SUM* are the least squares estimates of *a*, *b*, and *c*.[6] We can also test in a similar manner as discussed whether *b* and/or *c* are zero.

To conclude, regression analysis is used in many settings. In business, it is used to estimate the size of effects: How much do purchases of a good fall when prices rise? What is the effect of an increase in advertising on car sales? In public policy, magnitudes also matter and can be found using regression analysis: How much more will people use medical care if it is free, and how much will that help their health? How many lives are saved by reducing the speed limit on highways? These are all empirical questions in which regression analysis helps us to get at a magnitude with real consequences. With more data available every day, regression analysis has grown in importance.

[6] If you go on in economics, you will see that this least squares procedure has to be modified sometimes to account for various statistical problems, but the main goal of trying to find a good fit remains.

SUMMARY

36.1 SELECTION BIAS *p. 723*

1. One example of selection bias is survivor bias, where the most fit survive. This makes it difficult to compare young and old age groups.
2. Selection bias can arise if different kinds of people select into different groups, which can bias comparisons of the groups.

36.2 CAUSALITY *p. 724*

3. Correlation is not the same as causality.
4. Random experiments can sometimes be used to estimate causal effects. Intention to treat is sometimes used with random experiments to deal with limited take up in an experiment.
5. Regression discontinuity and difference-in-differences methodologies are also used to identify causality in economics.

36.3 STATISTICAL SIGNIFICANCE *p. 731*

6. An estimated effect is said to be statistically significant if the probability is small of obtaining the particular estimate when in fact the effect is zero. A probability of less than or equal to 5 percent is commonly used.

36.4 REGRESSION ANALYSIS *p. 732*

7. Regression analysis is used to estimate coefficients in equations. It is used both to obtain estimates of the magnitude of effects of various economic factors and to test alternative theories.

REVIEW TERMS AND CONCEPTS

correlated, *p. 724*
difference-in-differences, *p. 729*
intention to treat, *p. 726*
least squares estimates, *p. 733*
p-value, *p. 731*
regression discontinuity, *p. 727*
selection bias, *p. 723*
statistical significance, *p. 731*
survivor bias, *p. 723*

MyLab Economics Visit **www.pearson.com/mylab/economics** to complete these exercises online and get instant feedback. Exercises that update with real-time data are marked with ◉.