

Empirical Tools of Public Finance

3

3.1 The Important
Distinction Between
Correlation and Causality

3.2 Measuring Causation with
Data We'd Like to Have:
Randomized Trials

3.3 Conclusion

Once again, we return to your days as an employee of your state's Department of Health and Human Services. After doing the careful theoretical analysis outlined in the previous section, you are somewhat closer to making a meaningful contribution to the debate between the governor and the secretary of Health and Human Services. You can tell them that a reduction in TANF benefits is likely, but not certain, to raise labor supply among single mothers, and that the implications of this response depend on their concerns about equity versus efficiency. Yet these politicians don't just want to know that TANF reductions *might* raise labor supply, nor are they interested in the graphical calculations of the social welfare effects of lower benefits. What they want is numbers.

To provide these numbers, you now turn to the tools of **empirical public finance**, the use of data and statistical methodologies to measure the impact of government policy on individuals and markets. Many of these tools were developed more recently than the classical analyses of utility maximization and market equilibrium that we worked with in the last chapter. As a result, they are also more imperfect, and there are lively debates about the best way to approach problems like estimating the labor supply response of single mothers to TANF benefit changes.

In this chapter, we review these empirical methods. In doing so, we encounter the fundamental issue faced by those doing empirical work in economics: disentangling causality from correlation. We say that two economic variables are **correlated** if they move together. But this relationship is **causal** only if one of the variables is *causing* the movement in the other. If, instead, there is a third factor that causes both to move together, the correlation is not causal.

This chapter begins with a review of this fundamental problem. We then turn to a discussion of the "gold standard" for measuring the causal effect of an intervention (*randomized trials*) where individuals are randomly assigned to receive or not receive that intervention. While such randomized trials are much more common in medicine than in public finance, they provide a

empirical public finance The use of data and statistical methods to measure the impact of government policy on individuals and markets.

correlated Two economic variables are correlated if they move together.

causal Two economic variables are causally related if the movement of one causes movement of the other.

benchmark against which other empirical methods can be evaluated. We then discuss the range of other empirical methods used by public finance economists to answer questions such as the causal impact of TANF benefit changes on the labor supply of single mothers. Throughout, we use this TANF example, using real-world data on benefit levels and single-mother labor supply to assess the questions raised by the theoretical analysis of the previous chapter.

3.1

The Important Distinction Between Correlation and Causality

There was once a cholera epidemic in Russia. The government, in an effort to stem the disease, sent doctors to the worst-affected areas. The peasants of a particular province observed a very high correlation between the number of doctors in a given area and the incidence of cholera in that area. Relying on this fact, they banded together and murdered their doctors.¹

The fundamental problem in this example is that the peasants in this town clearly confused *correlation* with *causality*. They correctly observed that there was a positive association between physician presence and the incidence of illness. But they took that as evidence that the presence of physicians *caused* illness to be more prevalent. What they missed, of course, was that the link actually ran the other way: it was a higher incidence of illness that caused there to be more physicians present. In statistics, this is called the *identification problem*: given that two series are correlated, how do you identify whether one series is causing another?

This problem has plagued not only Russian peasants. In 1988, a Harvard University dean conducted a series of interviews with Harvard freshmen and found that those who had taken SAT preparation courses (a much less widespread phenomenon in 1988 than today) scored on average 63 points lower (out of 1,600 points) than those who hadn't. The dean concluded that SAT preparation courses were unhelpful and that "the coaching industry is playing on parental anxiety."² This conclusion is another excellent example of confusing correlation with causation. Who was most likely to take SAT preparation courses? Those students who needed the most help with the exam! So all this study found was that students who needed the most help with the SAT did the worst on the exam. The course did not cause students to do worse on the SATs; rather, students who would naturally do worse on the SATs were the ones who took the courses.

Another example comes from medical evaluation of the benefits of breast-feeding infants. Child-feeding recommendations typically include breast-feeding beyond 12 months, but some medical researchers have documented increased rates of malnutrition in breast-fed toddlers. This has led them to conclude that breast-feeding for too long is nutritionally detrimental. But the

¹ This example is reproduced from Fisher (1976).

² *New York Times* (1988).

misleading nature of this conclusion was illustrated by a study of toddlers in Peru that showed that it was those babies who were already underweight or malnourished who were breast-fed the longest.³ Increased breast-feeding did not lead to poor growth; children's poor growth and health led to increased breast-feeding.

The Problem

In all of the foregoing examples, the analysis suffered from a common problem: the attempt to interpret a correlation as a causal relationship without sufficient thought to the underlying process generating the data. Noting that those who take SAT preparation courses do worse on SATs, or that those infants who breast-feed longest are the least healthy, is only the first stage in the research process, that of documenting the correlation. Once one has the data on any two measures, it is easy to see if they move together, or *covary*, or if they do not.

What is harder to assess is whether the movements in one measure are *causing* the movements in the other. For any correlation between two variables A and B , there are three possible explanations, one or more of which could result in the correlation:

- ▶ A is causing B .
- ▶ B is causing A .
- ▶ Some third factor is causing both.

Consider the previous SAT preparation example. The fact is that, for this sample of Harvard students, those who took an SAT prep course performed worse on their SATs. The interpretation drawn by the Harvard administrator was one of only many possible interpretations:

- ▶ SAT prep courses worsen preparation for SATs.
- ▶ Those who are of lower test-taking ability take preparation courses to try to catch up.
- ▶ Those who are generally nervous people like to take prep courses, and being nervous is associated with doing worse on standardized exams.

The Harvard administrator drew the first conclusion, but the others may be equally valid. Together, these three interpretations show that one cannot interpret this correlation as a causal effect of test preparation on test scores without more information or additional assumptions.

Similarly, consider the breast-feeding interpretation. Once again, there are many possible interpretations:

- ▶ Longer breast-feeding is bad for health.
- ▶ Those infants who are in the worst health get breast-fed the longest.
- ▶ The lowest-income mothers breast-feed longer, since this is the cheapest form of nutrition for children, and low income is associated with poor infant health.

³ Marquis et al. (1997).

Once again, all of these explanations are consistent with the observed correlation. But, once again, the studies that argued for the negative effect of breast-feeding on health *assumed* the first interpretation while ignoring the others.

The general problem that empirical economists face in trying to use existing data to assess the causal influence of one factor on another is that one cannot immediately go from correlation to causation. This is a problem because for policy purposes what matters is causation. Policy makers typically want to use the results of empirical studies as a basis for predicting how government interventions will affect behaviors. Knowing that two factors are correlated provides no predictive power; prediction requires understanding the causal links between the factors. For example, the government shouldn't make policy based on the fact that breast-feeding infants are less healthy. Rather, it should assess the true causal effect of breast-feeding on infant health, and use that as a basis for making government policy. The next section begins to explore the answer to one of the most important questions in empirical research: How can one draw causal conclusions about the relationships between correlated variables?

3.2

Measuring Causation with Data We'd Like to Have: Randomized Trials

One of the most important empirical issues facing society today is understanding how new medical treatments affect the health of medical patients. An excellent example of this issue is the case of estrogen replacement therapy (ERT), a popular treatment for middle-aged and elderly women who have gone through menopause (the end of menstruation).⁴ Menopause is associated with many negative side effects, such as rapid changes in body temperature (“hot flashes”), difficulty sleeping, and higher risk of urinary tract infection. ERT reduces those side effects by mimicking the estrogen produced by the woman's body before the onset of menopause.

There was no question that ERT helped ameliorate the negative side effects of menopause, but there was also a concern about ERT. Anecdotal evidence suggested that ERT might raise the risk of heart disease, and, in turn, the risk of heart attacks or strokes. A series of studies beginning in the early 1980s investigated this issue by comparing women who did and did not receive ERT after menopause. These studies concluded that those who received ERT were at no higher risk of heart disease than those who did not; indeed, there was some suggestion that ERT actually *lowered* heart disease.

There was reason to be concerned, however, that such a comparison did not truly reflect the causal impact of ERT on heart disease. This is because women who underwent ERT were more likely to be under a doctor's care, to

⁴ For an overview of ERT issues, see Kolata (2002).

lead a healthier lifestyle, and to have higher incomes, all of which are associated with a lower chance of heart disease (the third channel previously discussed, where some third factor is correlated with both ERT and heart disease). So it is possible that ERT might have raised the risk of heart disease but that this increase was masked because the women taking the drug were in better health otherwise.

Randomized Trials as a Solution

How can researchers address this problem? The best solution is through the gold standard of causality: **randomized trials**. Randomized trials proceed by taking a group of volunteers and *randomly* assigning them to either a **treatment group**, which gets the medical treatment, or a **control group**, which does not. Effectively, volunteers are assigned to treatment or control by the flip of a coin.

To see why randomized trials solve our problem, consider what researchers would ideally do in this context: take one set of older women, replicate them, and place the originals and the clones in parallel universes. Everything would be the same in these parallel universes except for the use of ERT. Then, one could simply observe the differences in the incidence of heart disease between these two groups of women. Because the women are precisely the same, we know by definition that any differences would be causal. That is, there is only one possible reason why the set of women assigned ERT would have higher rates of heart disease, since otherwise both sets of women are the same.

Unfortunately, we live in the real world and not some science-fiction story, so we can't do this parallel universe experiment. But, amazingly, we can approximate this alternative reality through the randomized trial. This is because of the definition of randomization: assignment to treatment groups and control groups is not determined by anything about the subjects, but by the flip of a coin. As a result, the treatment group is identical to the control group in every facet but one: the treatment group gets the treatment (in this case, the ERT).

The Problem of Bias

We can rephrase all the studies we have discussed so far in this chapter in the treatment/control framework. In the SAT example, the treatment group was those who took preparatory classes and the control group was those who did not. In the breast-feeding example, the treatment group was those infants who breast-fed more than a year and the control group was those who did not. In the ERT studies that occurred before randomized trials, those who got ERT were the treatment group and those who did not were the control group. Even in the Russian doctor example, the treatment group was those areas where doctors were sent and the control group was those areas where doctors were not sent. Virtually any empirical problem we discuss in this course can be thought of as a comparison between treatment and control groups.

randomized trial The ideal type of experiment designed to test causality, whereby a group of individuals is randomly divided into a treatment group, which receives the treatment of interest, and a control group, which does not.

treatment group The set of individuals who are subject to an intervention being studied.

control group A set of individuals comparable to the treatment group that is not subject to the intervention being studied.

bias Any source of difference between treatment and control groups that is correlated with the treatment but not due to the treatment.

We can therefore always start our analysis of an empirical methodology with a simple question: Do the treatment and control groups differ for any reason *other* than the treatment? All the earlier examples involve cases in which the treatment groups differ in consistent ways from those in the control groups: those taking SAT prep courses may be of lower test-taking ability than those not taking the courses; those breast-fed longest may be in worse health than those not breast-fed as long; those taking ERT may be in better health than those not taking ERT. These non-treatment-related differences between treatment and control groups are the fundamental problem in assigning causal interpretations to correlations.

We call these differences **bias**, a term that represents any source of difference between treatment and control groups that is *correlated* with the treatment but is *not due* to the treatment. The estimates of the impact of SAT prep courses on SAT scores, for example, are *biased* by the fact that those who take the prep courses are likely to do worse on the SATs for other reasons. Similarly, the estimates of the impact of breast-feeding past one year on health are *biased* by the fact that those infants in the worst health are the ones likely to be breast-fed the longest. The estimates of the impact of ERT on heart disease are *biased* by the fact that those who take ERT are likely in better health than those who do not. Whenever treatment and control groups consistently differ in a manner that is correlated with, but not due to, the treatment, there can be bias.

By definition, such differences do not exist in a randomized trial, since the groups do not differ in any consistent fashion, but rather only by the flip of a coin. Thus, randomized treatment and control groups cannot have consistent differences that are correlated with treatment, since there are no consistent differences across the groups other than the treatment. As a result, *randomized trials have no bias*, and it is for this reason that randomized trials are the gold standard for empirically estimating causal effects.

Quick Hint The description of randomized trials here relies on those trials having fairly large numbers of treatments and controls (large *sample sizes*). Having large sample sizes allows researchers to eliminate any consistent differences between the groups by relying on the statistical principle called the *law of large numbers*: the odds of getting the wrong answer approaches zero as the sample size grows.

Suppose that a friend says that he can flip a (fair, not weighted!) coin so that it *always* comes up heads. This is not possible; every time a coin is flipped, there is a 50% chance it will land tails up. So you give him a quarter and ask him to prove it. If he flips just once, there is a 50% chance he will get heads and claim victory. If he flips twice, there is still a 25% chance that he will get heads both times, and continue to be able to claim victory; that is, there is still the possibility of getting a biased answer *by chance* when there is a very small sample.

As he flips more and more times, however, the odds that the coin will come up heads *every time* gets smaller and smaller. After just 10 flips, there is only a 1

in 1,024 chance that he gets all heads. After 20 flips, the odds are 1 in 1,048,576. That is, the higher the number of flips, the lower the odds that we get a biased answer. Likewise, if randomly assigned groups of individuals are large enough, we can rule out the possibility of bias arising by chance.

Randomized Trials of ERT

When the National Institutes of Health appointed its first female director, Dr. Bernadine Healy, in 1991, one of her priorities was to sponsor a randomized trial of ERTs. This randomized trial tracked over 16,000 women ages 50–79 who were recruited to participate in the trial by 40 clinical centers in the United States. The study was supposed to last 8.5 years but was stopped after 5.2 years because its conclusion was already clear: ERT did in fact raise the risk of heart disease. In particular, women taking ERT were observed to annually have (per 10,000 women): 7 more coronary heart diseases (both fatal and nonfatal), 8 more strokes, and 8 more pulmonary embolisms (blood clots in the lungs). In addition, the study found that women taking ERT had 8 more invasive breast cancers as well. Thus, the randomized trial revealed that the earlier ERT studies were *biased* by differences between these groups. These new findings led some doctors to question their decisions to recommend ERTs for postmenopausal women.⁵

Randomized Trials in the TANF Context

Measuring the health impacts of new medicines is not the only place where randomized trials are useful; they can be equally useful in the context of public policy. Suppose that we want to measure the causal impact of TANF on labor supply. To begin, we gather a large (e.g., 5,000 person) group of single mothers who are now receiving a \$5,000 benefit guarantee. One by one, we take each single mother into a separate room and flip a coin. If it is heads, they continue to receive a benefit guarantee of \$5,000; these mothers are the *control group* whose benefits do not change. If it is tails, then the guarantee is cut to \$3,000; these mothers are the *treatment group* who receive the experimental reduction in their benefits. After we have assigned a guarantee to all of these mothers, we follow them for a period of time and observe their labor supply differences. Any labor supply differences would have to be *caused* by the change in benefit guarantee, since nothing else differs in a consistent way across these groups.

There is a real-world randomized trial available that can help us learn about the impact of cash welfare benefits on the labor supply of single mothers. Under its Aid to Families with Dependent Children (AFDC) program in 1992, California had one of the most generous benefit guarantees in the United States, \$663 per month (\$7,956 per year) for a family of three. The

⁵ Results of the study are reported in Writing Group for the Women's Health Initiative Investigators (2002).

state wanted to assess the implications of reducing its AFDC benefit levels, in order to reduce costs. It conducted an experiment, randomly assigning one-third of the families receiving AFDC in each of four counties to the existing AFDC program, and assigning the other two-thirds to an experimental program. The experimental program had 15% lower maximum benefits, and several other provisions that encouraged recipients to work. The experiment lasted until 1998, at which point all families became subject to the 15% lower benefit.

Hotz, Mullin, and Scholz (2002) studied the effects of these benefit changes on the employment of recipients. They found that the experiment increased the employment rate of those families assigned to the experimental treatment to 49%, relative to an employment rate for the control group of 44.5%. The difference, 4.5%, is about 10% of the employment rate of the control group. It is often convenient to represent the relationship between economic variables in *elasticity* form, which in this case means computing the percentage change in employment for each percentage change in benefits. The estimated elasticity of employment with respect to benefits here is about -0.75 ; that is, a 15% reduction in the benefit guarantee resulted in a 10% increase in employment in the treatment group relative to that of the control group.

Why We Need to Go Beyond Randomized Trials

It would be wonderful if we could run randomized trials to assess the causal relationships that underlie any interesting correlation. For most questions of interest, however, randomized trials are not available. Such trials can be enormously expensive, take a very long time to plan and execute, and often raise difficult ethical issues. On the last point, consider the example of a recent trial for a new treatment for Parkinson's disease, a debilitating neurological disorder. The proposed treatment involved injecting fetal pig cells directly into patients' brains. In order to have a comparable control group, the researchers drilled holes in the heads of all 18 subjects, but put the pig cells in only 10 of the subjects.⁶ As you can imagine, there was substantial criticism of drilling holes in 8 heads for no legitimate medical purpose.

Moreover, even the gold standard of randomized trials has some potential problems. First, the results are only valid for the sample of individuals who volunteer to be either treatments or controls, and this sample may be different from the population at large. For example, those in a randomized trial sample may be less averse to risk or they may be more desperately ill. Thus, the answer we obtain from a randomized trial, while correct for this sample, may not be valid for the average person in the population.

A second problem with randomized trials is that of **attrition**: individuals may leave the experiment before it is complete. This is not a problem if individuals leave randomly, since the sample will remain random. Suppose, how-

attrition Reduction in the size of samples over time, which can lead to bias estimates if not random.

observational data Data generated by individual behavior observed in the real world, not in the context of deliberately designed experiments.

time series analysis Analysis of the comovement of two series over time.

⁶ Pollack (2001).

ever, that the experiment has positive effects on half the treatment group and negative effects on the other half, and that as a result the half with negative effects leaves the experiment before it is done. If we focus only on the remaining half, we would wrongly conclude that the treatment has overall positive impacts.

In the remainder of this chapter, we discuss several approaches taken by economists to try to assess causal relationships in empirical research. We will do so through the use of the TANF example. The general lesson from this discussion is that there is no way to perfectly approach the ideal of the randomized trial; bias is a pervasive problem that is not easily remedied. There are, however, methods available that can allow us to come close to the gold standard of randomized trials.

3.3

Estimating Causation with Data We Actually Get: Observational Data

In section 3.2, we showed how a randomized trial can be used to measure the impacts of an intervention such as ERT or lower TANF benefits on outcomes such as heart attacks or labor supply. As we highlighted, however, data from such randomized trials are not always available when important empirical questions need to be answered. Typically, what the analyst has instead are **observational data**, data generated from individual behavior observed in the real world. For example, instead of information on a randomized trial of a new medicine, we may simply have data on who took the medicine and what their outcomes were (the source of the original conclusions on ERT). There are several well-developed methods that can be used by analysts to address the problem of bias with observational data, and these tools can often closely approximate the gold standard of empirical trials.

This section explores how researchers can use observational data to estimate causal effects instead of just correlations. We do so within the context of the TANF example. It is useful throughout to refer to the empirical framework established in the previous section: those with higher TANF benefits are the control group, those with lower TANF benefits are the treatment group, and our concern is to remove any sources of bias between the two groups (that is, any differences between them that might affect their labor supply, other than TANF benefits differences). Thus, the major concern throughout this section is how to overcome any potential bias so that we can measure the causal relationship (if there is one) between TANF benefits and labor supply.

Time Series Analysis

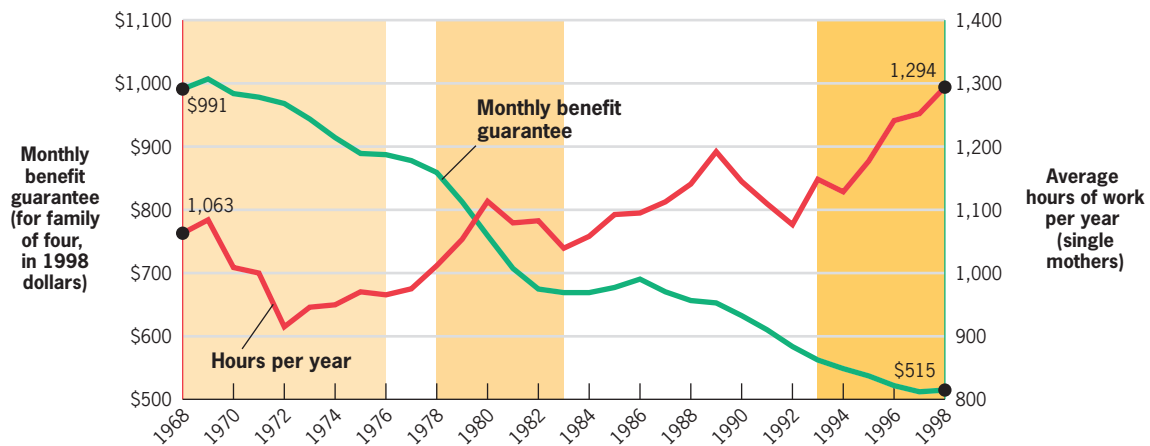
One common approach to measuring causal effects with observational data is **time series analysis**, documenting the correlation between the variables of interest over time. In the context of TANF, for example, we can gather data

over time on the benefit guarantee in each year, and compare these data to the amount of labor supply delivered by single mothers in those same years.

Figure 3-1 shows such a time series analysis. On the horizontal axis are years, running from 1968 through 1998. The right-hand vertical axis charts the average real monthly benefit guarantee for a single mother with three children (controlled for inflation by expressing income in constant 1998 dollars) available in the United States over this period. Benefits declined dramatically from \$1,000 in 1968 to \$500 in 1998, falling by half in real terms because benefit levels have not kept up with inflation. The left-hand vertical axis charts the average hours of work per year for single mothers (including zeros for those mothers who do not work). The hours worked have risen substantially, from 1,070 hours per year in 1968 to almost 1,300 in 1998. Thus, there appears to be a strong negative relationship between benefit guarantees and labor supply: falling benefit guarantees are associated with higher levels of labor supply by single mothers.

Problems with Time Series Analysis Although this time series correlation is striking, it does not demonstrate a causal effect of TANF benefits on labor supply. When there is a slow-moving trend in one variable through time, as is true for the general decline in income guarantees over this period, it is very

■ FIGURE 3-1



Average Benefit Guarantee and Single Mother Labor Supply, 1968–1998 • The left-hand vertical axis shows monthly benefit guarantee under cash welfare, which falls from \$991 in 1968 to \$515 in 1998. The right hand vertical axis shows average hours of work per year for single mothers, which rises from 1,063 in 1968 to 1,294 in 1998. Over this entire 30-year period, there is a strong negative correlation between average benefit guarantee and the level of labor supply of single mothers, but there is not a very strong relationship within subperiods of this overall time span.

Source: Calculations based on data from Current Population Survey's annual March Supplements

difficult to infer its causal effects on another variable. There could be many reasons why single mothers work more now than they did in 1968: greater acceptance of women in the workplace; better and more options for child care; even more social pressures on mothers to work. The simple fact that labor supply is higher today than it was thirty years ago does not prove that this increase has been caused by the steep decline in income guarantees.

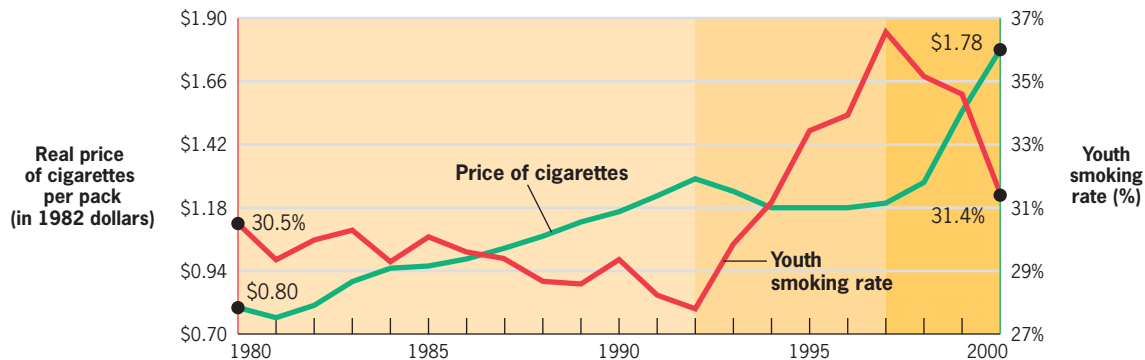
This problem is highlighted by examining subperiods of this overall time span. From 1968 through 1976, benefits fell by about 10% (from \$990 to \$890 per month), yet hours of work also fell by about 10% (from 1070 hours to 960 hours), whereas a causal effect of benefits would imply a rise in hours of work. From 1978 through 1983, the period of steepest benefits decline, benefits fell by almost one-quarter in real terms (from \$858 to \$669 per month), yet labor supply first increased, then decreased, with a total increase over this period of only 2%. The subperiods therefore give a very different impression of the relationship between benefits and labor supply than does the overall time series.

A particularly instructive example about the limitations of time series analysis is the experience of the 1993–1998 period. In this subperiod, there is both a sharp fall in benefits (falling by about 10%, from \$562 to \$515 per month) and a sharp rise in labor supply of single mothers (rising by about 13%, from 1148 hours per year to 1294 hours per year). The data from this subperiod would seem to support the notion that lower benefits cause rising labor supply. Yet during this period the economy was experiencing dramatic growth, with the general unemployment rate falling from 7.3% in January 1993 to 4.4% in December, 1998. It is also a period that saw an enormous expansion in the Earned Income Tax Credit (EITC), a federal wage subsidy that has been shown effective in increasing the labor supply of single mothers. It could be those factors, not falling benefits, that caused increased labor supply of single mothers. So once again, other factors get in the way of a causal interpretation of this correlation over time; factors such as economic growth and a more generous EITC can cause bias in this time series analysis because they are also correlated with the outcome of interest.

When Is Time Series Analysis Useful? Is all time series analysis useless? Not necessarily. In some cases, there may be sharp breaks in the time series that are not related to third factors that can cause bias. A classic example is shown in Figure 3-2. This figure shows the price of a pack of cigarettes (in constant 1982 dollars) on the right vertical axis and the *youth smoking rate*, the percentage of high school seniors who smoke at least once a month, on the left vertical axis. These data are shown for the time period from 1980 to 2000.

From 1980 to 1992, there was a steady increase in the real price of cigarettes (from 80¢ to \$1.29 per pack), and a steady decline in the youth smoking rate (from 30.5% to 27.8%). As previously noted, these changes over time need not be causally related. Smoking was falling for all groups over this time period due to an increased appreciation of the health risks of smoking, and prices may simply have been rising due to rising costs of tobacco production.

■ FIGURE 3-2



Real Cigarette Prices and Youth Smoking, 1980–2000 • The left-hand vertical axis shows the real price of cigarettes per pack, which rises from \$0.80 in 1980 to \$1.78 in 2000. The right-hand vertical axis shows the youth smoking rate (the share of high school seniors who smoke at least once a month), which fell from 1980 to 1992, rose sharply to 1997, and then fell again to 2000 to roughly its 1980 level. There is a striking correspondence between price and youth smoking within subperiods of this era.

Source: Calculations based on data on smoking from Monitoring the Future survey and on tobacco prices from the Tobacco Institute.

Then, in April 1993, there was a “price war” in the tobacco industry, leading to a sharp drop in real cigarette prices from \$1.29 to \$1.18 per pack.⁷ At that exact time, youth smoking began to rise. This striking simultaneous reversal in both series is more compelling evidence of a causal relationship than is the long, slow-moving correlation over the 1980–1992 period. But it doesn’t *prove* a causal relationship, because other things were changing in 1993 as well. It was, for example, the beginning of an important period of economic growth, which could have led to more youth smoking. Moreover, the rise in youth smoking is very large relative to the price decrease.

Fortunately, in this case, there is another abrupt change in this time series. In 1998 and thereafter, prices rose steeply when the tobacco industry settled a series of expensive lawsuits with many states (and some private parties) and passed the costs on to cigarette consumers. At that exact time, youth smoking began to fall again. This type of pattern seems to strongly suggest a causal effect, even given the limitations of time series data. That is, it seems unlikely that there is a factor correlated with youth smoking that moved up until 1992, then down until 1997, then back up again, as did price. That youth smoking follows the opposite pattern as cigarette prices suggests that price is causing these movements. Thus, while time series correlations are not very useful

⁷ The leading hypothesis for this sharp drop in prices on “Malboro Friday” (April 2, 1993) is that the major cigarette manufacturers were lowering prices in order to fight off sizeable market share gains by “generic” lower-priced cigarettes.

when there are long-moving trends in the data, they are more useful when there are sharp breaks in trends over a relatively narrow period of time.

Cross-Sectional Regression Analysis

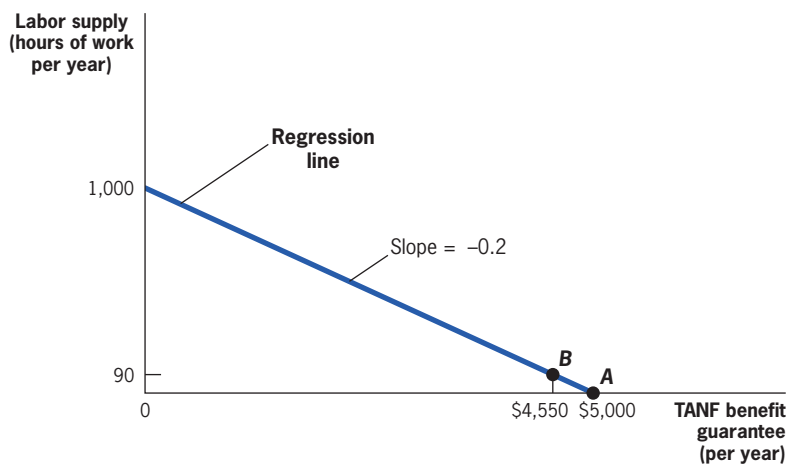
A second approach to identifying causal effects is **cross-sectional regression analysis**, a statistical method for assessing the relationship between two variables while holding other factors constant. By *cross-sectional* we mean comparing many individuals at one point in time, rather than comparing outcomes over time as in time series analysis.

In its simplest form, called a *bivariate regression*, cross-sectional regression analysis is a means of formalizing correlation analysis, of quantifying the extent to which two series covary. Returning to the example in Chapter 2, suppose that there are two types of single mothers, with preferences over leisure and food consumption represented by Figures 2-10 and 2-11 (p. ••). Before there is any change in TANF benefits, the mother who has a lower preference for leisure (Sarah in Figure 2-10) has both lower TANF benefits and higher labor supply than the mother who has a greater preference for leisure (Naomi in Figure 2-11). If we take these two mothers and correlate TANF benefits to labor supply, we would find that higher TANF benefits are associated with lower labor supply.

This correlation is illustrated graphically in Figure 3-3. We graph the two data points when the benefit guarantee is \$5,000. One data point, point *A*, corresponds to Naomi from Figure 2-11, and represents labor supply of 0 hours and an income guarantee of \$5,000. The other data point, point *B*, corresponds to Sarah in Figure 2-10, and represents labor supply of 90 hours per year and TANF benefits of \$4,550. The downward sloping line makes clear

cross-sectional regression analysis Statistical analysis of the relationship between two or more variables exhibited by many individuals at one point in time.

■ FIGURE 3-3



TANF Benefits and Labor Supply in Theoretical Example • If we plot the data from the theoretical example of Chapter 2, we find a modest negative relationship between TANF benefits and the labor supply of single mothers.

the *negative correlation* between TANF benefits and labor supply; the mother with lower TANF benefits has higher labor supply.

Regression analysis takes this correlation one step further by quantifying the relationship between TANF benefits and labor supply. Regression analysis does so by finding the line that best fits this relationship, and then measuring the slope of that line.⁸ This is illustrated in Figure 3-3. The line that connects these two points has a slope of -0.2 . That is, this bivariate regression indicates that each \$1 reduction in TANF benefits per month leads to a 0.2-hour-per-year increase in labor supply. Regression analysis describes the relationship between the variable that you would like to explain (the *dependent variable*, labor supply in our example) and the set of variables that you think might do the explaining (the *independent variables*, the TANF benefit in our example).

Example with Real-World Data The example in Figure 3-3 is made up, but we can replicate this exercise using real data from one of the most popular sources of cross-sectional data for those doing applied research in public finance: the Current Population Survey, or CPS.⁹

The CPS collects information every month from individuals throughout the United States on a variety of economic and demographic issues. For example, this survey is the source of the unemployment rate statistics that you frequently hear cited in the news. Every year, in March, a special supplement to this survey asks respondents about their sources of income and hours of work in the previous year. So we can take a sample of single mothers from this survey and ask: What is the relationship between the TANF benefits and hours of labor supply in this cross-sectional sample?

Figure 3-4 graphs the hours of labor supply per year (vertical axis) against dollars of TANF benefits per year (horizontal axis), for all of the single mothers in the CPS data set. To make the graph easier to interpret, we divide the data into ranges of TANF income (\$0 in TANF benefits; \$1–\$99 of benefits; \$100–\$250 of benefits; etc.). Each range represents (roughly) a doubling of the previous range (a logarithmic scale). For each range, we show the average hours of labor supply in the group. For example, as the highlighted point shows, single mothers receiving between \$250 and \$499 in benefits supply just over 600 hours of labor per year.

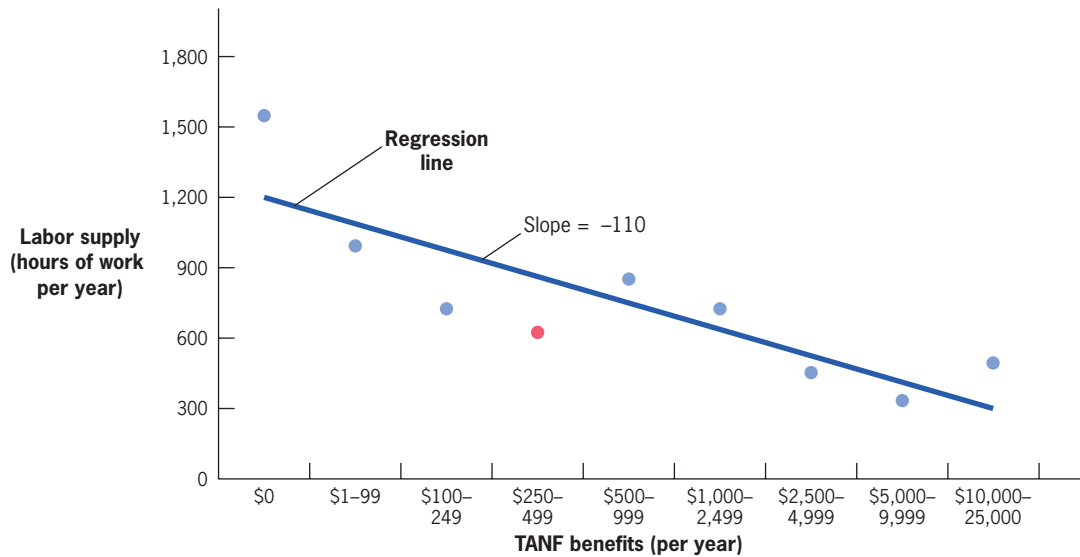
It is immediately clear from this graph that there is a *negative relationship* between TANF benefits and hours of labor supply. The single mothers at the left of the graph, where benefits are lowest, have much higher labor supply on average than those on the right of the graph, where TANF benefits are the highest. The line in Figure 3-4 formalizes this eyeball impression. This linear **regression line** shows the *best* linear approximation to the relationship between TANF benefits and labor supply that is represented by these points. Unlike the made-up example in Figure 3-3, there is no single line that fits perfectly through this set of data points; real-world data are never that neat!

regression line The line that measures the best linear approximation to the relationship between any two variables.

⁸ We discuss here only *linear* approaches to regression analysis; nonlinear regression analysis, where one fits not only lines but other shapes to the data, is a popular alternative.

⁹ Author's computations from the Current Population Survey can be found at <http://www.worthpublishers.com/gruber>.

■ FIGURE 3-4



TANF Benefit Income and Labor Supply of Single Mothers Using CPS Data • Using data from the CPS, we group single mothers by the amount of TANF income they have. Those who are receiving the lowest level of TANF income are the ones providing the highest number of work hours.

Source: Calculations based on data from Current Population Survey's annual March Supplements.

What the linear regression does is find the line that comes closest to fitting through the cluster of data points.¹⁰

This line has a slope of -110 , which indicates that each doubling of TANF benefits reduces hours of work by 110 per year (remember that each segment on the horizontal axis represents a doubling of benefits). Once again, it is convenient to represent the relationship between economic variables in elasticity form. Based on these CPS data, the mean (average number of) hours of work in our sample is 748 hours. So we know that each 100% rise in TANF benefits reduces hours of work by 15% (110 is 15% of 748), for an elasticity of -0.15 . This is a fairly *inelastic* response; there is a relatively modest reduction in hours (15%) when TANF benefits rise (by 100%).

Problems with Cross-Sectional Regression Analysis The result summarized in Figure 3-4 seems to indicate strongly that mothers who receive the largest TANF benefits work the fewest hours. Once again, however, there are several possible interpretations of this correlation. One interpretation is that higher TANF benefits are causing an increase in leisure. Another possible

¹⁰ Technically, this line is the one that minimizes the sum of squared distances of each point from the line. As a result, one major concern with linear regression analysis is “outliers”; a point that is very far from the others exerts a strong influence on this line, since we are minimizing the sum of *squared* distances, so a large distance has an exponentially large effect. For this reason, analysts often use other approaches that are less sensitive to such outlying observations.

interpretation is that some mothers have a high taste for leisure and wouldn't work much even if TANF benefits weren't available. Because TANF benefits fall as the recipient works more, mothers who take more leisure automatically get higher levels of benefits. As a result, there is a correlation between benefits and leisure (and therefore labor supply) because more leisure is causing higher TANF benefits, *not* because higher TANF benefits are causing more leisure. Thus, varying tastes for leisure cause a bias in our attempt to causally interpret the relationship between TANF benefits and labor supply. Differences in tastes for leisure are one reason why those with high and low TANF benefits are not exactly comparable; these differences in taste cause a consistent difference (bias) in labor supply among mothers with high and low TANF benefits.

This problem is most clearly illustrated in Figure 3-3, since we actually know the utility functions underlying the labor supply decisions of the two mothers represented by points *A* and *B*. The mother who works less does so because she has a higher taste for leisure, and not because her TANF benefits are higher. In fact, her higher taste for leisure is what drives her TANF benefits to be higher, because TANF benefits increase as leisure increases and hours worked decrease. Thus, the negative relationship depicted in Figure 3-3 is *not causal*; it reflects, instead, differences in the taste for leisure between the two mothers we are analyzing that are correlated with their benefit levels (bias). In other words, we haven't taken two identical mothers and assigned them different benefits, which is what causal analysis demands. Rather, we took two very different mothers and compared their benefits and labor supply, which introduces bias into the analysis.

This problem is less obvious in Figure 3-4, since we don't know the utility functions of the single mothers in the CPS. But the same problematic potential exists: maybe the mothers with low TANF income are simply those who have the lowest preference for leisure. If this is true, we can't say that each doubling of TANF income *causes* a 15% reduction in labor supply. Rather, all we can say is that each doubling of TANF income *is associated with* a 15% reduction in labor supply. It could be that other consistent differences between these low- and high-benefit groups (such as different tastes for leisure) are biasing the relationship.

Control Variables Regression analysis has one potential advantage over correlation analysis in dealing with the problem of bias: the ability to include **control variables**. Suppose that the CPS had a variable included in the data set called "taste for leisure" that accurately reflected each individual's taste for leisure. Suppose that this variable came in two categorical values: "prefers leisure" and "prefers work," and that everyone within each of these categorical values had identical tastes for leisure and work. That is, there is no bias within these groups, only across them; within each group, individuals are identical in terms of their preferences toward work and leisure.

If we had this information, we could divide our sample into two groups according to this leisure variable, and redo the analysis within each group. Within each group, different tastes for leisure cannot be the source of the relationship between TANF benefits and labor supply, because tastes for leisure

control variables Variables that are included in cross-sectional regression models to account for differences between treatment and control groups that can lead to bias.

are identical within each group. This “taste for leisure” control variable will allow us to get rid of the bias in our comparison, because within each group we no longer have a systematic difference in tastes for leisure that is correlated with benefits. Control variables in regression analysis play this role: they try to control for (take into account) other differences across individuals in a sample, so that any remaining correlation between the dependent variable (e.g., labor supply) and independent variable (e.g., TANF benefits) can be interpreted as a causal effect of benefits on work.

In reality, control variables are unlikely to ever solve this problem completely, as the key variables we want, such as the intrinsic taste for leisure in this example, are impossible to measure in data sets. Usually, we have to approximate the variables we really want, such as taste for leisure, with what is available, such as age or education or work experience. These are imperfect proxies, however, so that they don’t fully allow us to control for differences in taste for leisure across the population (e.g., even within age or education or work experience groups, there will be individuals with very different tastes for leisure). Thus, it is hard to totally get rid of bias with control variables, since control variables only represent in a limited way the underlying differences between treatment and control groups. We discuss this point in the Appendix to this chapter, which includes reference to data on our Web site, that you can use to conduct your own regression analysis.

Quick Hint For many empirical analyses, there will be one clear treatment and control group, as in the ERT case. For other analyses, such as our cross-sectional TANF analysis, there are many groups to be compared with one another. A cross-sectional regression essentially compares each point in Figure 3-4 with the other points in order to estimate the relationship between TANF benefits and labor supply.

Even though the treatment/control analogy is no longer exact, however, the general intuition remains. It is essential in all empirical work to ensure that there are no factors that cause consistent differences in behavior (labor supply) across two groups and are also correlated with the independent variable (TANF benefits). When there are more than two groups, the concern is the same: to ensure that there is no consistent factor that causes groups with higher benefits to supply less labor than groups with lower benefits, other than the benefit differences themselves.

Quasi-Experiments

As noted earlier, public finance researchers cannot set up randomized trials and run experiments for every important behavior that matters for public policy. We have examined alternatives to randomized trials such as time series and cross-sectional regression analysis, but have also seen that these research methods have many shortcomings which make it hard for them to eliminate the

quasi-experiments Changes in the economic environment that create roughly identical treatment and control groups for studying the effect of that environmental change, allowing public finance economists to take advantage of randomization created by external forces.

bias problem. Is there any way to accurately assess causal influences without using a randomized trial? Is there an alternative to the use of control variables for purging empirical models of bias?

Over the past two decades, empirical research in public finance has become increasingly focused on one potential middle-ground solution: the **quasi-experiment**, a situation that arises naturally when changes in the economic environment (such as a policy change) create nearly identical treatment and control groups that can be used to study the effect of that policy change. In a quasi-experiment, outside forces (such as those instituting the policy change) do the randomization for us. Thus, if researchers can't make a comparison between treatment and control groups that is free of bias, they may be able to study the outcome of a situation in which treatment and control groups have been created naturally.

For example, suppose that we have a sample with a large number of single mothers in the neighboring states of Arkansas and Louisiana, for two years, 1996 and 1998. Suppose further that, in 1997, the state of Arkansas cut its benefit guarantee by 20%, while Louisiana's benefits remained unchanged. In principle, this alteration in the states' policies has essentially performed our randomization for us. The women in Arkansas who experienced the decrease in benefits are the treatment group, and the women in Louisiana whose benefits did not change are the control. By computing the change in labor supply across these groups, and then examining the difference between treatment (Arkansas) and control (Louisiana), we can obtain an estimate of the impact of benefits on labor supply that is free of bias.

In principle, of course, we could learn about the effect of this policy change by simply studying the experience of single mothers in Arkansas. If nothing differed between the set of single mothers in the state in 1996 and the set of single mothers in the state in 1998, other than the benefits reduction, then any change in labor supply would reflect only the change in benefits, and the results would be free of bias. In practice, such a comparison typically runs into the problems we associate with time series analysis. For example, the period from 1996 through 1998 was a period of major national economic growth, with many more job openings for low-skilled workers, which could lead single mothers to leave TANF and increase their earnings even in the absence of a benefits change. Thus, it is quite possible that single mothers in Arkansas may have increased their labor supply even if their benefits had not fallen.

Because other factors may have changed that affected the labor decision of single mothers in Arkansas, the quasi-experimental approach includes the extra step of comparing the treatment group for whom the policy changed to a control group for whom it did not. The state of Louisiana did not change its TANF guarantee between 1996 and 1998, but single mothers in Louisiana benefited from the same national economic boom as did those in Arkansas. If the increase in labor supply among single mothers in Arkansas is driven by economic conditions, then we should see the same increase in labor supply among single mothers in Louisiana; if the increase in labor supply among single mothers in Arkansas is driven by lower TANF benefits, then we would see

no change among single mothers in Louisiana. The bias introduced into our comparison of single mothers in Arkansas in 1996 to single mothers in Arkansas in 1998 by the improvement in economic conditions across the nation is *also* present when we do a similar comparison within Louisiana. In Louisiana, however, the treatment effect of a higher TANF benefit is *not* present. In this comparison, we can say that:

$$\begin{aligned} \text{Hours (Arkansas, 1998)} - \text{Hours (Arkansas, 1996)} &= \text{Treatment effect} + \text{bias} \\ &\quad \text{from economic boom} \\ \text{Hours (Louisiana, 1998)} - \text{Hours (Louisiana, 1996)} &= \text{Bias from economic} \\ &\quad \text{boom} \\ \text{Difference} &= \text{Treatment effect} \end{aligned}$$

By subtracting the change in hours of work in Louisiana (the control group) from the change in hours of work in Arkansas (the treatment group), we control for the bias caused by the economic boom and obtain a causal estimate of the effect of TANF benefits on hours of work.

Table 3-1 provides an illustrative, but hypothetical, set of numbers that we can use to analyze the results of this quasi-experiment. Suppose that the welfare guarantee was cut from \$5,000 to \$4,000 in Arkansas between 1996 and 1998. Over the same period, hours of work per year among single mothers in the state rose by from 1,000 to 1,200. The time-series estimate using the experience of Arkansas alone would be that the \$1,000 benefit reduction (20%) increased hours of work by 200 (20%). This outcome implies an elasticity of total hours with respect to benefits of -1 (a 20% benefit cut led to a 20% labor supply rise). Notice that this estimate is considerably larger than the -0.75 elasticity found in the randomized trial in California (our gold standard).

Consider now the bottom panel of Table 3-1. This panel shows that, between 1996 and 1998, there was no change in welfare benefits in Louisiana, but hours of work increased by 50 hours per year. Thus, it appears that the economic boom did play a role in the increase in hours worked by single mothers. By looking only at time series data from Arkansas, we ignore the effect of the economic boom. If we don't take this effect into account in our study, our conclusions about the effect of TANF benefits on labor supply will be biased.

A simple solution to this problem, as we have seen, is to examine the difference between the change in Arkansas and the change in Louisiana. That is, Arkansas had both a cut in welfare benefits and an economic boom, and hours of labor supply rose by 200; Louisiana had only the economic boom, and hours of labor supply rose by 50. These results suggest that the welfare benefit cut in Arkansas caused a 150 hour increase in

■ TABLE 3-1

Using Quasi-Experimental Variation

Arkansas			
	1996	1998	Difference
Benefit guarantee	\$5,000	\$4,000	-\$1,000
Hours of work per year	1,000	1,200	200
Louisiana			
	1996	1998	Difference
Benefit guarantee	\$5,000	\$5,000	\$0
Hours of work per year	1,050	1,100	50

In Arkansas, there is a cut in the TANF guarantee between 1996 and 1998 and a corresponding rise in labor supply, so if everything is the same for single mothers in both years, this is a causal effect. If everything is not the same, we can perhaps use the experience of a neighboring state that did not decrease its benefits, Louisiana, to capture any bias to the estimates.

labor supply, net of the economic changes. Once we've eliminated the bias caused by the improvement in overall economic conditions, the implied elasticity of hours with respect to welfare benefits is -0.75 , the same as that found in the California experiment. This technique is called a **difference-in-difference estimator**: Take the difference between the labor supply changes in the treatment group which experiences the change (in this case, single mothers in Arkansas) and the labor supply changes in the control group which does not experience the change, but is otherwise identical to the treatment group (in this case, single mothers in Louisiana). In this way, we can estimate a causal effect of TANF benefits changes on labor supply.

Difference-in-difference estimators try to combine time series and cross-sectional analyses to address the problems with each. By comparing the change in Arkansas to the change in Louisiana, the estimator controls for other time series factors that bias the time series analysis within Arkansas. Likewise, by comparing the change within each state, rather than just comparing the two states at a point in time, the estimator controls for omitted factors that bias cross-sectional analysis across the two states.

The cross-sectional estimate in this context would contrast Arkansas and Louisiana in 1998, when their benefits differ. In 1998, Arkansas had TANF benefits that were \$1,000 lower than Louisiana, and single mothers in Arkansas worked 100 hours more per year. Cross-sectional analysis would therefore conclude that each \$1,000 reduction in welfare benefits leads to a 100 hour increase in work, rather than the 150 hour increase that we get from difference-in-difference analysis (and that we know is true from the randomized trial).

This cross-sectional estimate is biased by the fact that single mothers tend to work more hours in Louisiana regardless of the level of TANF benefits. This is illustrated by the fact that, when TANF benefits were identical in the two states in 1996, hours of work were higher in Louisiana. In principle, we might find control variables to account for the more hours of work in Louisiana, but in practice that is difficult. The difference-in-difference estimator suggests the best possible control: the hours of work in the *same state* before there was a benefits change. That is, by comparing the change within a state to the change within another state, the difference-in-difference estimator controls for cross-sectional differences across states that might bias the comparison.

difference-in-difference estimator The difference between the changes in outcomes for the treatment group that experiences an intervention and the control group that does not.

structural estimates Estimates of the features that drive individual decisions, such as income and substitution effects or utility parameters.

Problems with Quasi-Experimental Analysis As well as the difference-in-difference quasi-experimental approach works to control for bias, it is still less than ideal. Suppose, for example, that the economic boom of this period affected Arkansas in a different way than it affected Louisiana. If this were true, then the “bias from economic boom” terms in the previous comparison would not be equal, and we would be unable to isolate the treatment effect of higher TANF benefits by simple subtraction. Instead, we get a new bias term: the difference in the impact of the economic boom in Arkansas and Louisiana. That is, when we compute our difference-in-difference estimator we obtain:

$$\begin{aligned}
 \text{Hours (Arkansas, 1998)} - \text{Hours (Arkansas, 1996)} &= AR \text{ bias from economic boom} + \text{Treatment} \\
 \text{Hours (Louisiana, 1998)} - \text{Hours (Louisiana, 1996)} &= LA \text{ bias from} \\
 &\quad \frac{\text{economic boom}}{\text{Difference}} \\
 &= \text{Treatment effect} + \\
 &\quad (AR \text{ bias} - LA \text{ bias})
 \end{aligned}$$

Since *AR* and *LA* biases are not equal, then the estimator will not identify the true treatment effect.

With quasi-experimental studies, unlike true experiments, we can never be completely certain that we have purged all bias from the treatment-control comparison. Quasi-experimental studies use two approaches to try to make the argument that they have obtained a causal estimate. The first is intuitive: trying to argue that, given the treatment and control groups, it seems very likely that bias has been removed. The second is statistical: to continue to use alternative or additional control groups to confirm that the bias has been removed. In Chapter 14, we discuss some examples of finding such alternative or additional control groups.

Structural Modeling

The randomized trials and quasi-experimental approaches previously described have the distinct advantage that, if applied appropriately, they can address the difficult problem of distinguishing causality from correlation. Yet they also have two important limitations. First, they only provide an estimate of the causal impact of a *particular treatment*. That is, the California experiment found that cutting benefits by 15% raised employment rates by 4.5 percentage points. This is the best estimate of the impact of cutting benefits by 15%, but it may not tell us much about the impact of cutting benefits by 30%, or of raising benefits by 15%. That is, we can't necessarily *extrapolate* from a particular change in the environment to model all possible changes in the environment. These approaches give us a precise answer to a specific question, but not necessarily a general conclusion about how different changes in benefits might affect behavior.

The second limitation is that these approaches can tell us *how* outcomes change when there is an intervention, but often they cannot tell us *why*. Consider the behavior of mothers with income between \$6,000 and \$10,000 in our example from Chapter 2, and how they react to a cut in benefits under TANF. For these mothers, as we noted, there is both an income and substitution effect leading to more work; they are both poorer because benefits have fallen, and they have a higher net wage since the implicit tax rate has fallen. An experimental or quasi-experimental study of the responses of these women to the benefits reduction might show us the total effect on their labor supply, but it would tell us very little about the relative importance of these income and substitution effects.

Yet, as we will learn later in this book, we often care about the **structural estimates** of labor supply responses, the estimates that tell us about features of

reduced form estimates

Measures of the impact of a particular change on an observable variable like labor supply.

structural estimation A set of techniques for measuring structural estimates.

utility that drive individual decisions, such as substitution and income effects. Randomized or quasi-experimental estimates provide **reduced form estimates** only, the impact of one particular change on overall labor supply responses. Thus, this second disadvantage of randomized or quasi-experiments is related to the first: if we understood the underlying structure of labor supply responses, it might be possible to say more about how labor supply would respond to different types of policy interventions.

These issues have led to the vibrant field of **structural estimation**. Using this research approach, empirical economists attempt to estimate not just reduced form responses to the environment but the actual underlying features of utility functions. They do so by more closely employing the theory developed in the previous chapter to develop an empirical framework that not only estimates overall responses, but also decomposes these responses into, for example, substitution and income effects.

Structural models potentially provide a very useful complement to experimental or quasi-experimental analyses. Yet structural models are often more difficult to estimate than reduced form models because both use the same amount of information, yet structural models are trying to learn much more from that information. Consider the TANF example. The earlier analysis showed you how to derive a reduced form estimate of the impact of a change in TANF benefits. Using this same information to decompose that response into income and substitution effects is not possible employing the same simple approach. Rather, that decomposition is only possible if the researcher assumes a particular form for the utility function, as we did in Chapter 2, and then employs that assumption to decompose the overall response into its two components. If the assumption for the form of the utility function is correct, then this approach provides more information. If it is incorrect, however, then the response derived from this approach might lead one to incorrectly estimate income and substitution effects.

From the perspective of this text, reduced form estimation has one other advantage (which may be obvious after reading this section!): it is much easier to think about and explain. Thus, for the remainder of the text, we will largely rely on reduced form modeling and evidence when discussing empirical results in public finance. Yet the promise of structural modeling should not be discounted, and is a topic of fruitful future study for those of you who want to go on in economics. The lessons about empirical work learned in this book are universal for all types of studies; they provide a basis that you can take forward to more sophisticated empirical approaches such as structural modeling.

3.3**Conclusion**

The central issue for any policy question is establishing a causal relationship between the policy in question and the outcome of interest. Do lower welfare benefits *cause* higher labor supply among single mothers? Does more pollution in the air *cause* worse health outcomes? Do larger benefits for unem-

ployment insurance *cause* individuals to stay unemployed longer? These are the types of questions we will address in this book using the empirical methods described here.

In this chapter, we discussed several approaches to distinguishing causality from correlation. The gold standard for doing so is the randomized trial, which removes bias through randomly assigning treatment and control groups. Unfortunately, however, such trials are not available for every question we wish to address in empirical public finance. As a result, we turn to alternative methods such as time series analysis, cross-sectional regression analysis, and quasi-experimental analysis. Each of these alternatives has weaknesses, but careful consideration of the problem at hand can often lead to a sensible solution to the bias problem that plagues empirical analysis.

► HIGHLIGHTS

- A primary goal of empirical work is to document the causal effects of one economic factor on another, for example the causal effect of raising TANF benefits on the labor supply of single mothers.
- The difficulty with this goal is that it requires treatment groups (those who are affected by policy) and control groups (those not affected) who are identical except for the policy intervention.
- If these groups are not identical, then there can be bias—that is, other consistent differences across treatment/control groups that are correlated with, but not due to, the treatment itself.
- Randomized trials are the gold standard to surmount this problem. Since treatments and controls are identical by definition, there is no bias, and any differences across the groups are a causal effect.
- Time series analysis is unlikely to provide a convincing estimate of causal effects because so many other factors change through time.
- Cross-sectional regression analysis also suffers from bias problems because similar people make different choices for reasons that can't be observed, leading once again to bias. Including control variables offers the potential to address this bias.
- Quasi-experimental methods have the potential to approximate randomized trials, but control groups must be selected carefully in order to avoid biased comparisons.

► QUESTIONS AND PROBLEMS

1. Suppose you are running a randomized experiment and you randomly assign study participants into control and treatment groups. After making the assignments, you study the characteristics of the two groups and find that the treatment group has a lower average age than the control group. How could this arise?
2. Why is a randomized trial the “gold standard” for solving the identification problem?
3. What do we mean when we say that correlation does not imply causality? What are some of the ways in which an empirical analyst attempts to disentangle the two?
4. A researcher conducted a cross-sectional analysis of children and found that average test performance of children with divorced parents was lower than average test performance of children of intact families. This researcher then concluded that divorce is bad for children's test outcomes. What is wrong with this analysis?
5. A study in the *Annals of Improbable Research* once reported that counties with large numbers of mobile-home parks had higher rates of tornadoes than the rest of the population. The authors conclude that mobile-home parks cause tornado occurrences. What is an alternative explanation for this fact?

6. What are some of the concerns with conducting randomized trials? How can quasi-experiments potentially help here?
7. You are hired by the government to evaluate the impact of a policy change that affects one group of individuals but not another. Suppose that before the policy change, members of a group affected by the policy averaged \$17,000 in earnings and members of a group unaffected by the policy averaged \$16,400. After the policy change, members of the affected group averaged \$18,200 in earnings while members of the unaffected group averaged \$17,700 in earnings.
 - a. How can you estimate the impact of the policy change? What is the name for this type of estimation?
 - b. What are the assumptions you have to make for this to be a valid estimate of the impact of the policy change?
8. Consider the example presented in the Appendix to this chapter. Which coefficient estimates would be considered “statistically significant” or distinct from zero?
9. A researcher wants to investigate the effects of a public policy on housing prices and has only cross-sectional data to use for this analysis. When she performs her regression analysis, she controls for average January and July temperatures in the area. Why is she doing this?

► ADVANCED QUESTIONS

10. Researchers often use *panel data* (multiple observations over time of the same people) to conduct regression analysis. With these data, researchers are able to compare the same person over time in assessing the impacts of policies on individual behavior. How could this provide an improvement over cross-sectional regression analysis of the type described in the text?
11. Suppose that your state announced that it would provide free tuition to high-achieving students graduating from high school starting in 2007. You decide to see whether this new program induces families with high-achieving children graduating in 2007 or later to purchase new cars. To test your findings, you use a “falsification exercise”: you observe the new-car-purchasing behavior of families with children graduating in 2006. Why is this a useful exercise?
12. Your state introduced a tax cut in the year 1999. You are interested in seeing whether this tax cut has led to increases in personal consumption within the state.

You observe the following information:

Year	Consumption in your state
1994	300
1996	310
1998	320
2000	350

- a. Your friend argues that the best estimate of the effect of the tax cut is an increase in consumption of 30 units, but you think that the true effect is smaller, because consumption was trending upward prior to the tax cut. What do you think is a better estimate?
- b. Suppose that you find information on a neighboring state that did not change its tax policy during this time period. You observe the following information in that state:

Year	Consumption in neighboring state
1994	260
1996	270
1998	280
2000	300

Given this information, what is your best estimate of the effect of the tax cut on consumption? What assumptions are required for that to be the right estimate of the effect of the tax cut? Explain.

Cross-Sectional Regression Analysis

In the text, we presented a cursory discussion of cross-sectional regression analysis, and the role of control variables. In this appendix, we provide a more detailed presentation of this approach, within our TANF example.

Data For this analysis, we use data from the March 2002 Current Population Survey (CPS). From that survey, we selected all women who reported that they were unmarried and had a child younger than age 19. The total sample is 8,024 single mothers.

For this sample, we have gathered data on the following variables for each woman:

- ▶ *TANF*: Total cash TANF benefits in the previous year (in thousands).
- ▶ *Hours*: Total hours of work in the previous year, computed as reported weeks of work times usual hours per week.
- ▶ *Race*: We divide reported race into white, black, and other.
- ▶ *Age*: Age in years.
- ▶ *Education*: We use reported education to divide individuals into four groups: high school dropouts; high school graduates with no college; those with some college; and college graduates.
- ▶ *Urbanicity*: We use information on residential location to divide individuals into four groups: central city; other urban; rural; and unclear (the CPS doesn't identify location for some mothers for survey confidentiality reasons).

Regression Using these data, we can estimate a regression of the impact of welfare on hours of work of the form:

$$(1) \text{ HOURS}_i = \alpha + \beta \text{ TANF}_i + \epsilon_i$$

where there is one observation for each mother i . This is the counterpart of the regression analysis shown in Figure 3-4, but now we are using each individual data point, rather than grouping the data into categories for convenience.

In this regression, α , the constant term, represents the estimated number of hours worked if welfare benefits are zero. β is the slope coefficient, which represents the change in hours worked per dollar of welfare benefits. ϵ is the

error term, which represents the difference for each observation between its actual value and its predicted value based on the model.

The results of estimating this regression model are presented in the first column of the appendix table. The first row shows the constant term α , which is 1537: this measures the predicted hours of labor supply delivered at zero welfare benefits. The second row shows the coefficient β , which is -108 : each \$1,000 of welfare benefits lowers hours worked by 108. This is very close to the estimate from the grouped data of -110 discussed in the text.

Thus, for a mother with no welfare benefits, predicted hours of work are 1537; for a mother with \$5,000 in welfare benefits, predicted hours of work are $1537 - 5 \times 108 = 997$.

Underneath this estimate in parentheses is the estimate's *standard error*. This figure captures the precision with which these coefficients are estimated and reminds us that we have here only a statistical representation of the relationship between welfare benefits and hours worked. Roughly speaking, we cannot statistically distinguish values of β that are two standard errors below or above the estimated coefficient. In our context, with a standard error of 3.7 hours, the results show that our best estimate is that each thousand dollars of welfare lowers hours worked by 107, but we can't rule out that the effect is only 96.6 ($107 - 2 \times 3.7$) or that it is 114.4 ($107 + 2 \times 3.7$).

In the context of empirical economics, this is a *very* precise estimate. Typically, as long as the estimate is more than twice the size of its standard error, we say that it is *statistically significant*.

The final row of the table shows the R^2 of the regression. This is a measure of how well the statistical regression model is fitting the underlying data. An R^2 of 1 would mean that the data are perfectly explained by the model so that all data points lie directly on the regression line; an R^2 of

0 means that the data are not at all explained. The value of 0.095 here says that less than 10% of the variation in the data is explained by this regression model.

As discussed in the text, however, this regression model suffers from serious bias problems, since those mothers who have a high taste for leisure will have both low hours of work and high welfare payments. One approach to addressing this problem suggested in the text was including control variables. We don't have the ideal control variable, taste for leisure. We do, however, have other variables that might be correlated with tastes for leisure or other factors

■ APPENDIX 3 TABLE

Cross-Sectional Regression Analysis

	Equation (1)	Equation (2)
Constant	1537 (10)	2062 (61)
Welfare	-107 (3.7)	-93 (3.6)
White		181 (44)
Black		61 (47)
High school dropout		-756 (30)
High school graduate		-347 (25)
Some college		-232 (28)
Age		-9.3 (0.8)
Central city		-12 (30)
Other urban		34 (29)
Rural		-43 (31)
R^2	0.095	0.183

that determine labor supply: race, education, age, and urbanicity. So we can estimate regression models of the form:

$$(2) \text{ HOURS}_i = \alpha + \beta \text{ TANF}_i + \delta \text{ CONTROL}_i + \epsilon_i$$

where CONTROL are the set of control variables for individual i .

In the second column of the appendix table, we show the impact of including these other variables. When we have a categorical variable such as race (categorized into white, black, and other), we include *indicator variables* that take on a value of 1 if the individual is of that race, and 0 otherwise. Note that when we have N categories for any variable (e.g., 3 categories for race), we only include $N-1$ indicator variables, so that all estimates are relative to the excluded category (e.g., the coefficient on the indicator for “black” shows the impact of being black on welfare income, relative to the omitted group of Hispanics).

Adding these control variables does indeed lower the estimated impact of welfare benefits on labor supply. The coefficient falls to -93 , but remains highly significant. The R^2 doubles but still indicates that we are explaining less than 20% of the variation in the data.

The control variables are themselves also of interest:

- ▶ *Race*: Whites are estimated to work 181 hours per year more than Hispanics (the omitted group); blacks are estimated to work 61 hours per year more, but this estimate is only about 1.3 times as large as its standard error, so we do not call that a statistically significant difference.
- ▶ *Education*: Hours of work clearly rise with education. High school dropouts work 756 fewer hours per year than do college graduates (the omitted group); high school graduates work 347 fewer hours per year; and those with some college work 232 fewer hours per year than those who graduate from college. All of these estimates are very precise (the coefficients are very large relative to the standard errors beneath them in parentheses).
- ▶ *Age*: Hours worked decline with age, with each year of age leading to 9 fewer hours of work; this is a very precise estimate as well.
- ▶ *Location*: Relative to those with unidentified urbanicity, people in cities and rural areas work less and those in the suburbs work more, but none of these estimates are statistically precise.

Do these control variables eliminate bias in the estimated relationship between TANF benefits and labor supply? There is no way to know for sure, but it seems unlikely. The fact that this large set of controls explains only 9% more of the variation in labor supply across individuals suggests that they are unlikely to capture all the factors correlated with both labor supply and TANF benefits.