Chapter 5

Differences-in-Differences

MASTER KAN: If while building a house, a carpenter strikes a nail and it proves faulty by bending, does the carpenter lose faith in all nails and stop building? So it is with empirical work.

Kung Fu, Season 1, Episode 7

Our Path

Credible instrumental variables and dramatic policy discontinuities can be hard to find; you'll need other 'metrics tools in your kit too. The *differences-in-differences* (DD) method recognizes that in the absence of random assignment, treatment and control groups are likely to differ for many reasons. Sometimes, however, treatment and control outcomes move in parallel in the absence of treatment. When they do, the divergence of a post-treatment path from the trend established by a comparison group may signal a treatment effect. We demonstrate DD with a study of the effects of monetary policy on bank failures during the Great Depression. We also revisit the MLDA.

5.1 A Mississippi Experiment

On the eve of the largest economic downturn in American history—the Great Depression—economic spirits ran high in the halls of high finance.

Caldwell and Company's slogan "We Bank on the South" reflected the confidence of a regional financial empire. Based in Nashville, Caldwell ran the largest Southern banking chain in the 1920s, and owned many nonbanking businesses as well. Rogers Caldwell, known as the J. P. Morgan of the South, lived large on an estate that housed his stable of prize-winning thoroughbreds. Alas, in November of 1930, mismanagement and fallout from the stock market crash of October 1929 brought the Caldwell empire down. Within days, Caldwell's collapse felled closely tied banking networks in Tennessee, Arkansas, Illinois, and North Carolina. The Caldwell crisis was a harbinger of a surge in bank failures across the country.

Banking is a business built on confidence and trust. Banks lend to businesses and property owners in the expectation that most loans will be paid off when they come due. Depositors trust they'll be able to withdraw their funds on demand. This confidence notwithstanding, banks hold less cash than needed to pay all depositors, because most deposits are out on loan. The resulting maturity mismatch poses no problem in normal times, when few depositors make withdrawals on any given day.

If confidence falters, the banking system breaks down. In the 1930s, when your bank went out of business, your savings very likely disappeared with it. Even if your bank's mortgage and loan portfolios looked safe, you wouldn't have wanted to be the last depositor to try to get your money out. Once other depositors are seen to withdraw in panic, you'd do well to panic too. That's how a bank run starts.

Caldwell's demise shook depositor confidence throughout the American South and precipitated a run on Mississippi banks in December 1930. Deposits in Mississippi fell slowly at first, but on December 19, the floodgates opened when savers panicked. On that day, the Mississippi state Banking Department closed three banks. Two more banks ceased operations the day after, and another 29 folded in the next six months. The regional panic of 1930 was one of many more to come. In 1933, the year Depression-era bank failures peaked, more than 4,000 banks failed nationwide.

Economists have long sought to understand whether and how monetary policy contributed to the Great Depression, and whether more aggressive monetary intervention might have stemmed the financial collapse and economic free fall seen in those dark days. Depression-era lessons may help us understand the present. Although financial markets today are more sophisticated, the pillars of finance remain much as they were: banks borrow and lend, typically at different maturities, and bet on being able to raise the cash (known in banking jargon as "liquidity") needed to cover liabilities as they come due.

We're unlucky enough to live in economically interesting times. The year 2008 saw the U.S. financial system shaken by a collapse in the market for mortgage-backed securities, followed by a European sovereign debt crisis beginning in late 2009. Carmen Reinhart and Kenneth Rogoff have recently chronicled financial crises since the fourteenth century, arguing they share a common anatomy. The apparent similarity of such episodes makes you wonder whether they can be avoided, or at least whether their effects can be mitigated. In their masterful 1963 monetary history of the United States, Milton Friedman and Anna Schwartz convinced many economists that a proper understanding of the effects of monetary policy is the key to answering this question.[1]

One Mississippi, Two Mississippi

Policymakers facing a bank run can open the flow of credit or turn off the tap. Friedman and Schwartz argued that the Federal Reserve (America's central bank) foolishly restricted credit as the Great Depression unfolded. Easy money might have allowed banks to meet increasingly urgent withdrawal demands, staving off depositor panic. By lending to troubled banks freely, the central bank has the power to stem a liquidity crisis and obviate the need for a bailout in the first place.

But who's to say when a crisis is merely a crisis of confidence? Some crises are real. Bank balance sheets may be so sickened by bad debts that no amount of temporary liquidity support will cure 'em. After all, banks

don’t lose their liquidity by random assignment. Rather, bank managers make loans that either fail or are fruitful. Injecting central bank funds into bad banks may throw good money after bad. Better in such cases to declare bankruptcy and hope for an orderly distribution of any remaining assets.

Support for bad banks also raises the specter of what economists call *moral hazard*. If bankers know that the central bank will lend cheaply when liquidity runs dry, they needn’t take care to avoid crises in the first place. In 1873, *The Economist*’s editor-in-chief Walter Bagehot described the danger this way:

> If the banks are bad, they will certainly continue bad and will probably become worse if the Government sustains and encourages them. The cardinal maxim is, that any aid to a present bad Bank is the surest mode of preventing the establishment of a future good Bank.[2]

Bagehot was a professed Social Darwinist, believing that evolutionary principles applied in social affairs just as in biology. Which policy stance is more likely to speed a happy ending to an economic downturn, liquidity backstopping or survival of banking’s fittest? As always, masters of ’metrics would like to settle this question with a randomized trial. We have a grant proposal to fund such a bank liquidity experiment under review; we’ll surely blog the results if it comes through. In the meantime, we must learn about the effects of monetary policy from the history of banking crises and policy responses to them.

Fortunately for this research agenda, the U.S. Federal Reserve System is organized into 12 districts, each run by a regional Federal Reserve Bank. Depression-era heads of the regional Feds had considerable policy independence. The Atlanta Fed, running the Sixth District, favored lending to troubled banks. By contrast, the St. Louis Fed ran the Eighth District according to a philosophy known as the Real Bills Doctrine, which holds that the central bank should restrict credit in a recession. Especially happily for research on monetary policy, the border between

the Sixth and Eighth Districts runs east-west smack through the middle of the state of Mississippi (District borders were determined by population size in 1913, at the birth of the Federal Reserve System). This border defines a within-state natural experiment from which we can profit.

Masters Gary Richardson and William Troost analyzed Mississippi's monetary two-step.[3] As we might expect from their differing approaches to monetary policy, the Atlanta and St. Louis Feds reacted very differently to the Caldwell crisis. Within 4 weeks of Caldwell's collapse, the Atlanta Fed had increased bank lending by about 40% in the Sixth District. In the same period, bank lending by the St. Louis Fed in the Eighth District fell almost 10%.

The Richardson and Troost policy experiment imagines the Eighth District as a control group, where policy was to do little or even restrict lending, while the Sixth District is a treatment group, where policy was to increase lending. A first-line outcome is the number of banks still operating in each District on July 1, 1931, about 8 months after the beginning of the crisis. On that day, 132 banks were open in the Eighth District and 121 were open in the Sixth District, a deficit of 11 banks in the Sixth District. This suggests easy money was counterproductive. But look again: the Sixth and Eighth Districts were similar but not identical. We see this in the fact that the number of banks operating in the two districts differed markedly across districts on July 1, 1930, well before the Caldwell crisis, with 135 banks open in the Sixth District and 165 banks open in the Eighth. To adjust for this difference across districts in the pre-treatment period, we analyze the Mississippi experiment using a tool called differences-in-differences, or DD for short.

Parallel Worlds

Let Y_{dt} denote the number of banks open in District d in year t, where the subscript d tells us whether we're looking at data from the Sixth or Eighth District and the subscript t tells us whether we're looking at data from 1930 (before the Caldwell crisis) or 1931 (after). The DD estimate

(δ_{DD}) of the effect of easy money in the Sixth District is

$$\begin{aligned}\delta_{DD} &= (Y_{6,1931} - Y_{6,1930}) - (Y_{8,1931} - Y_{8,1930}) \\ &= (121 - 135) - (132 - 165) \\ &= -14 - (-33) = 19. \qquad (5.1)\end{aligned}$$

Instead of comparing the number of banks open in the Sixth and Eighth Districts after Caldwell, DD contrasts the change in the number of banks operating in the two districts.

Comparing changes instead of levels adjusts for the fact that in the pre-treatment period, the Eighth District had more banks open than the Sixth. To see this, note that we can produce the same DD bottom line this way:

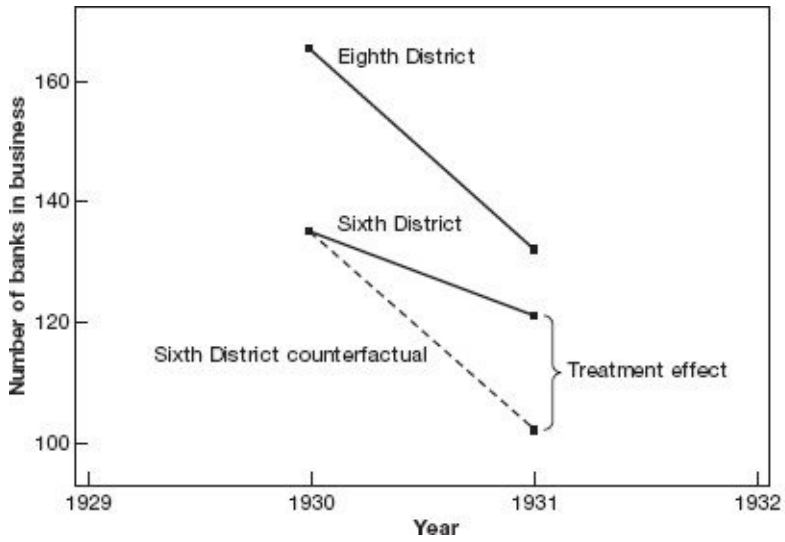
$$\begin{aligned}\delta_{DD} &= (Y_{6,1931} - Y_{8,1931}) - (Y_{6,1930} - Y_{8,1930}) \\ &= (121 - 132) - (135 - 165) \\ &= -11 - (-30) = 19. \qquad (5.2)\end{aligned}$$

This version of the DD calculation subtracts the pre-treatment difference between the Sixth and Eighth Districts from the post-treatment difference, thereby adjusting for the fact that the two districts weren't the same initially. DD estimates suggest that lending to troubled banks kept many of them open. Specifically, the Atlanta Fed appears to have saved 19 banks—more than 10% of those operating in Mississippi's Sixth District in 1930.

DD logic is depicted in Figure 5.1, which plots the number of banks in the Sixth and Eighth Districts in 1930 and 1931, with data from the two periods connected by solid lines. Figure 5.1 highlights the fact that while banks failed in both Federal Reserve Districts, they did so much more sharply in the Eighth.

FIGURE 5.1

Bank failures in the Sixth and Eighth Federal Reserve Districts



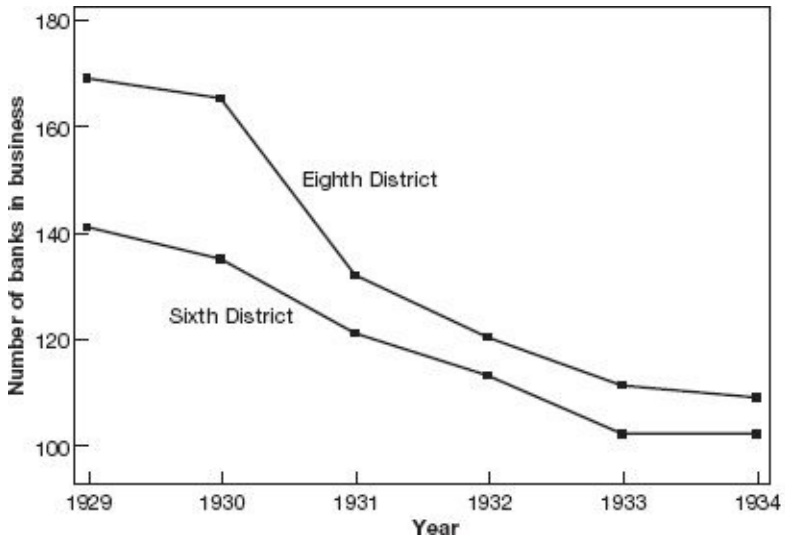
Notes: This figure shows the number of banks in operation in Mississippi in the Sixth and Eighth Federal Reserve Districts in 1930 and 1931. The dashed line depicts the counterfactual evolution of the number of banks in the Sixth District if the same number of banks had failed in that district in this period as did in the Eighth.

The DD tool amounts to a comparison of slopes or trends across districts. The dotted line in Figure 5.1 is the counterfactual outcome that lies at the heart of the DD research design: this line tells us what would have happened in the Sixth District had everything evolved as it did in the Eighth. The fact that the solid line for the Sixth District declines much more gradually than this counterfactual line is evidence for the effectiveness of easy money. The 19 bank failures uncovered by our DD calculation is the difference between what really happened and what would have happened had bank activity in the two districts unfolded in parallel.

The DD counterfactual comes from a strong but easily stated assumption: *common trends*. In the Mississippi experiment, DD presumes that, absent any policy differences, the Eighth District trend is what we should have expected to see in the Sixth. Although strong, the common trends assumption seems like a reasonable starting point, one that takes account of pretreatment differences in levels. With more data, the assumption can also be probed, tested, and relaxed.

FIGURE 5.2

Trends in bank failures in the Sixth and Eighth Federal Reserve Districts

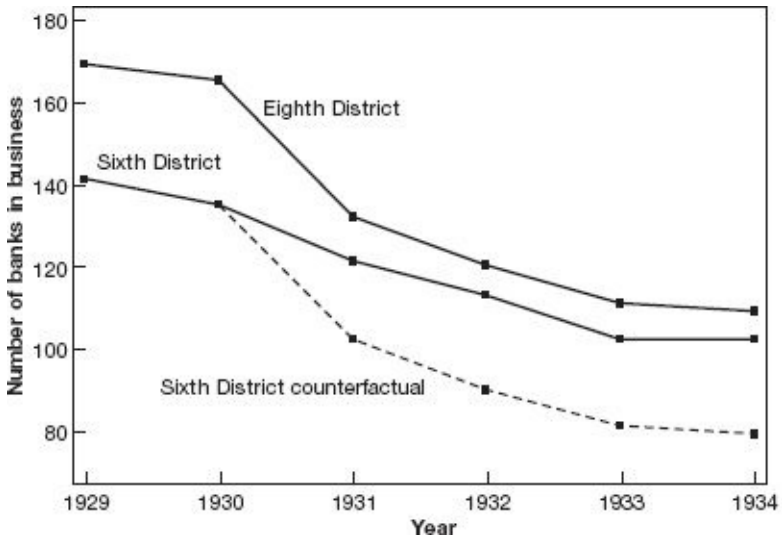


Note: This figure shows the number of banks in operation in Mississippi in the Sixth and Eighth Federal Reserve Districts between 1929 and 1934.

Figure 5.2 provides evidence on the common trends assumption for Mississippi's Federal Reserve Districts. The evidence comes in the form of a longer time series on bank activity. Before 1931, the Great Depression had not yet hit Mississippi hard. Regional Fed policies in the two districts were also similar in this more relaxed period. The fact that bank failures moved almost in parallel in the two districts between 1929 and 1930, with the number of banks declining slightly in both districts, is therefore consistent with the common trends hypothesis for untreated periods. Figure 5.3 adds the Sixth District counterfactual implied by extrapolating Eighth District trends to the Sixth District for years after 1930. The gap between actual and counterfactual Sixth District banking activity changed little through 1934.

FIGURE 5.3

Trends in bank failures in the Sixth and Eighth Federal Reserve Districts, and the Sixth District's DD counterfactual



Notes: This figure adds DD counterfactual outcomes to the banking data plotted in Figure 5.2. The dashed line depicts the counterfactual evolution of the number of banks in the Sixth District if the same number of banks had failed in that district after 1930 as did in the Eighth.

As in Figure 5.1, the relatively steep fall-off in bank activity in the Eighth District after the Caldwell collapse emerges clearly in Figures 5.2 and 5.3. But these figures document something further. Beginning in July 1931, the St. Louis Fed abandoned tight money and started lending to troubled banks freely. In other words, after 1931, Federal Reserve policy in the two districts was again similar, with both regional Feds willing to provide liquidity with a free hand. Moreover, while the Depression was far from over in 1932, the Caldwell crisis had petered out and withdrawals had returned to pre-crisis levels. Given the two regional Feds' common readiness to lend as the need arose, trends in bank activity should again have been common after 1931. The 1931–1934 data line up well with this hypothesis.

Just DDo It: A Depression Regression

The simplest DD calculation involves only four numbers, as in equations (5.1) and (5.2). In practice, however, the DD recipe is best cooked with regression models fit to samples of more than four data points, such as the 12 points plotted in Figure 5.2. In addition to allowing for more than two periods, regression DD neatly incorporates data on more than two

cross-sectional units, as we'll see in a multistate analysis of the MLDA in Section 5.2. Equally important, regression DD facilitates statistical inference, often a tricky matter in a DD setup (for details, see the appendix to this chapter).

The regression DD recipe associated with Figure 5.2 has three ingredients:

(i) A dummy for the treatment district, written $TREAT_d$, where the subscript d reminds us that this varies across districts; $TREAT_d$ controls for fixed differences between the units being compared.

(ii) A dummy for post-treatment periods, written $POST_t$, where the subscript t reminds us that this varies over time; $POST_t$ controls for the fact that conditions change over time for everyone, whether treated or not.

(iii) The interaction term, $TREAT_d \times POST_t$, generated by multiplying these two dummies; the coefficient on this term is the DD causal effect.

We think of the Caldwell-era experimental treatment as provision of easy credit in the face of a liquidity crisis, so $TREAT_d$ equals one for data points from the Sixth District and zero otherwise. The bank failure rate slowed after 1931 as the Caldwell crisis subsided. In the 1930s, however, there were no zombie banks: dead banks were gone for good. The Caldwell-era failures resulted in fewer banks open in the years 1932–1934 as well, even though the St. Louis Fed had by then begun to lend freely. We therefore code $POST_t$ to indicate all the observations from 1931 onward. Finally, the interaction term, $TREAT_d \times POST_t$, indicates observations in the Sixth District in the post-treatment period. More precisely, $TREAT_d \times POST_t$ indicates observations from the Sixth District in periods when the Atlanta Fed's response to Caldwell mattered for the number of active banks.

Regression DD for the Mississippi experiment puts these pieces

together by estimating

$$Y_{dt} = \alpha + \beta\, TREAT_d + \gamma\, POST_t + \delta_{rDD}(TREAT_d \times POST_t) + e_{dt} \qquad (5.3)$$

in a sample of size 12. This sample is constructed by stacking observations from both districts and all available years (6 years for each district). The coefficient on the interaction term, δ_{rDD}, is the causal effect of interest. With only two periods, as in Figure 5.1, estimates of δ_{DD} and δ_{rDD} coincide (a consequence of the properties of dummy variable regression outlined in the appendix to Chapter 2). With more than two periods, as in Figure 5.2, estimates based on equation (5.3) should be more precise and provide a more reliable picture of policy effects than the simple four-number DD recipe.[4]

Fitting equation (5.3) to the 12 observations plotted in Figure 5.2 generates the following estimates (with standard errors shown in parentheses):

$$Y_{dt} = 167 - \underset{(8.8)}{29}\, TREAT_d - \underset{(7.6)}{49}\, POST_t + \underset{(10.7)}{20.5}\,(TREAT_d \times POST_t) + e_{dt}.$$

These results suggest that roughly 21 banks were kept alive by Sixth District lending. This estimate is close to the estimate of 19 banks saved using the four-number DD recipe. The standard error for the estimated δ_{rDD} is about 11, so 21 is a marginally significant result, the best we can hope for with such a small sample.

Let's Get Real

The Atlanta Fed very likely saved many Sixth District banks from failure. But banks are not valued for their own sakes. Did the Atlanta Fed's policy of easy money support real economic activity, that is, non-bank businesses and jobs? Statistics on business activity within states are scarce for this period. Still, the few numbers available suggest the Atlanta Fed's bank liquidity backstopping generated real economic benefits. This is documented in Table 5.1, which lists the ingredients for a simple DD analysis of Federal Reserve liquidity effects on the number of active wholesalers and their sales.

TABLE 5.1

Wholesale firm failures and sales in 1929 and 1933

	1929	1933	Difference (1933–1929)
Panel A. Number of wholesale firms			
Sixth Federal Reserve District (Atlanta)	783	641	−142
Eighth Federal Reserve District (St. Louis)	930	607	−323
Difference (Sixth–Eighth)	−147	34	181
Panel B. Net wholesale sales ($ million)			
Sixth District Federal Reserve (Atlanta)	141	60	−81
Eighth District Federal Reserve (St. Louis)	245	83	−162
Difference (Sixth–Eighth)	−104	−23	81

Notes: This table presents a DD analysis of Federal Reserve liquidity effects on the number of wholesale firms and the dollar value of their sales, paralleling the DD analysis of liquidity effects on bank activity in Figure 5.1.

DD estimates for Mississippi wholesalers parallel those for Mississippi banks. Between 1929 and 1933, the number of wholesale firms and their sales fell in both the Sixth and Eighth Districts, with a much sharper drop in the Eighth District, where more banks failed. In the 1920s and 1930s, wholesalers relied heavily on bank credit to finance inventories. The estimates in Table 5.1 suggest that the reduction in bank credit in the Eighth District in the wake of Caldwell brought wholesale business activity down as well, with a likely ripple effect throughout the local economy. Sixth District wholesalers were more likely to have been spared this fate. Cooked with only a four-number DD recipe, however, the evidence for a liquidity treatment effect in Table 5.1 is weaker than that produced by the larger sample for bank activity.

The Caldwell experiment offers a hard-won lesson in how to nip a banking crisis in the bud. Perhaps the governor of the St. Louis Fed, seeing a more modest collapse in the Sixth District than in the Eighth, had absorbed the Caldwell lesson by the time he reversed course in 1931. But the palliative power of monetary policy in a financial crisis was understood by national authorities only much later. In their memoirs, Milton Friedman and his wife Rose famously recounted:

> Instead of using its powers to offset the Depression, [the Federal Reserve Board in Washington, D.C.] presided over a decline in the

> quantity of money by one-third from 1929 to 1933. If it had operated as its founders intended, it would have prevented that decline and, indeed, converted it into the rise that was called for to accommodate the normal growth in the economy.[5]

Which isn't to say that the problem of financial crisis management has since been nailed. Today's complex financial markets run off the rails for many reasons, not all of which can be contained by the Fed and its printing presses. That hard lesson is being learned by the monetary authorities of our day.

5.2 Drink, Drank, …

> SHEN: Are you willing to die to find the truth?
>
> PO: You bet I am! … Although, I'd prefer not to.
>
> *Kung Fu Panda 2*

With the repeal of federal alcohol Prohibition in 1933, U.S. states were free to regulate alcohol. Most instituted an MLDA of 21, but Kansas, New York, and North Carolina, among others, allowed drinking at 18. Following the twenty-sixth amendment to the constitution in 1971, which lowered the voting age to 18 in response to agitation sparked by the Vietnam War, many states reduced the MLDA. But not all: Arkansas, California, and Pennsylvania are among the states that held the line at 21. In 1984, the National Minimum Drinking Age Act punished youthful intemperance by withholding federal aid for highway construction from states with an age-18 MLDA. By 1988, all 50 states and the District of Columbia had opted for an MLDA of 21, though some had taken the federal highway hint more quickly than others.

As with much American policymaking, the interaction of federal and state law produces a colorful and oft-changing quilt of legal standards. This policy variation is a boon to masters of 'metrics: variation in state MLDA laws is easily exploited in a DD framework. In efforts to uncover

effects of alcohol policy, this framework provides an alternative to the RD approach detailed in Chapter 4.[6]

Patterns from Patchwork

Alabama lowered its MLDA to 19 in 1975, but alphabetically and geographically proximate Arkansas has had an MLDA of 21 since Prohibition's repeal. Did Alabama's indulgence of its youthful drinkers cost some of them their lives? We tackle this question by fitting a regression DD model to data on the death rates of 18–20-year-olds from 1970 to 1983. The dependent variable is denoted Y_{st}, for death rates in state s and year t. With a sample including only Alabama and Arkansas, the regression DD model for Y_{st} takes the form

$$Y_{st} = \alpha + \beta\, TREAT_s + \gamma\, POST_t + \delta_{rDD}(TREAT_s \times POST_t) + e_{st}, \quad (5.4)$$

where $TREAT_s$ is a dummy variable indicating Alabama, $POST_t$ is a dummy indicating years from 1975 onward, and the interaction term $TREAT_s \times POST_t$ indicates Alabama observations from low-drinking-age years. The coefficient δ_{rDD} captures the effect of an age-19 MLDA on death rates.

Equation (5.4) parallels the regression DD model for Mississippi's two Federal Reserve Districts. But why look only at Alabama and Arkansas? There's more than one MLDA experiment in the legislative record. For example, Tennessee's MLDA fell to 18 in 1971, then rose to 19 in 1979. A complicating but manageable consequence of differences in the timing of MLDA reductions in Alabama and Tennessee is the absence of a common post-treatment period. When combining multiple MLDA experiments in a DD framework, we swap the single $POST_t$ dummy for a set of dummies indicating each year in the sample, with one omitted as a reference group. The coefficients on these dummies, known as *time effects,* capture temporal changes in death rates that are common to all states.[7]

Our multi-MLDA regression DD procedure should also reflect the fact that there are many states driving causal comparisons. Instead of controlling only for the difference between, say, the Sixth and Eighth Federal Reserve Districts as in the Mississippi experiment of Section 5.1, or the difference between Alabama and Arkansas in the example above, the multistate setup controls for the differing death rates in each of many states. This is accomplished by introducing *state effects,* a set of dummies for every state in the sample, except for one, which is omitted as a reference group. A regression DD analysis of data from Alabama, Arkansas, and Tennessee, for example, includes two state effects. State effects replace the single $TREAT_s$ dummy included in a two-state (or two-group) analysis.

A final complication in this scenario is the absence of a common treatment variable that discretely switches off and on. The MLDA runs from age 18 to age 21, generating treatment effects for legal drinking at ages 18, 19, or 20. Masters of 'metrics simplify such things by reducing them to a single measure of exposure to the policy of interest, in this case, access to alcohol. Our simplification strategy replaces $TREAT_d \times POST_t$ with a variable we'll call $LEGAL_{st}$. This variable measures the proportion of 18–20-year-olds allowed to drink in state s and year t. In some states, no one under 21 is allowed to drink, while in states with an age-19 MLDA, roughly two-thirds of 18–20-year-olds can drink, and in states with an age-18 MLDA, all 18–20-year-olds can drink. Our definition of $LEGAL_{st}$ also captures variation due to within-year timing. For example, Alabama's age-19 MLDA came into effect in July 1975. $LEGAL_{\text{AL},1975}$ is therefore scaled to reflect the fact that Alabama's 19–20-year-olds were free to drink for only half that year.

The multistate regression DD model looks like

$$Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} + \sum_{k=\text{Alaska}}^{\text{Wyoming}} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} + e_{st}. \quad (5.5)$$

Don't let the big sums in this equation scare you. This notation describes models with many dummy variables compactly, just as in the models with college selectivity group dummies in Chapter 2. Here every state but one (the reference state) gets its own dummy variable, indexed by the subscript k for state k. The index s keeps track of the state supplying the observations. The kth state dummy, $STATE_{ks}$ equals one when an observation is from state k, meaning $s = k$, and is zero otherwise. Observations from California, for example, have $STATE_{\mathrm{CA},\,s}$ switched on, and all other state dummies switched off.

The state effects, β_k, are the coefficients on the state dummies. For example, the California state effect, β_{CA} is the coefficient on $STATE_{\mathrm{CA},s}$. Every state except the reference state, the one omitted when constructing state dummies, has a state effect in equation (5.5). Because there are so many of these, we use summation notation, $\sum_{k=\mathrm{Alaska}}^{\mathrm{Wyoming}} \beta_k STATE_{ks}$, to save writing them all out. The time effects, γ_t, are similarly coefficients on the year dummies, $YEAR_{jt}$. These switch on when observations in the data come from year j, that is, when $t = j$. We therefore also call them *year effects*. The 1975 year effect, γ_{1975}, is the coefficient on $YEAR_{1975,\,t}$. Here, too, every year in the sample except the reference year has a year effect, so we use summation notation to write these out compactly.[8]

Our multistate MLDA analysis uses a data set with 14 years and 51 states (including the District of Columbia), for a total of 714 observations. This data structure is called a *state-year panel.* The state effects in equation (5.5) control for fixed differences between states (for example, fatal car accidents are more frequent, on average, in rural states with high average travel speeds). The time (year) effects in this equation control for trends in death rates that are common to all states (due, for example, to national trends in drinking or vehicle safety). Equation (5.5) attributes changes in mortality within states to changes in $LEGAL_{st}$. As we'll see shortly, this causal attribution turns on a common trends assumption, just as in our analysis of Caldwell-induced bank

failures in the previous section.

Estimates of δ_{rDD} in equation (5.5) suggest that legal alcohol access caused about 11 additional deaths per 100,000 18–20-year-olds, of which seven or eight deaths were the result of motor vehicle accidents. These results, reported in the first column of Table 5.2, are somewhat larger than but still broadly consistent with the RD estimates reported in Table 4.1 in Chapter 4. The MVA estimates in Table 5.2 are also reasonably precise, with standard errors of about 2.5. Importantly, as with the RD estimates, this regression DD model generates little evidence of an effect of legal drinking on deaths from internal causes. The regression DD evidence for an effect on suicide is weaker than the corresponding RD evidence in Table 4.1. At the same time, both strategies suggest any increase in numbers of suicides is smaller than for MVA deaths.

TABLE 5.2

Regression DD estimates of MLDA effects on death rates

Dependent variable	(1)	(2)	(3)	(4)
All deaths	10.80	8.47	12.41	9.65
	(4.59)	(5.10)	(4.60)	(4.64)
Motor vehicle accidents	7.59	6.64	7.50	6.46
	(2.50)	(2.66)	(2.27)	(2.24)
Suicide	.59	.47	1.49	1.26
	(.59)	(.79)	(.88)	(.89)
All internal causes	1.33	.08	1.89	1.28
	(1.59)	(1.93)	(1.78)	(1.45)
State trends	No	Yes	No	Yes
Weights	No	No	Yes	Yes

Notes: This table reports regression DD estimates of minimum legal drinking age (MLDA) effects on the death rates (per 100,000) of 18–20-year-olds. The table shows coefficients on the proportion of legal drinkers by state and year from models controlling for state and year effects. The models used to construct the estimates in columns (2) and (4) include state-specific linear time trends. Columns (3) and (4) show weighted least squares estimates, weighting by state population. The sample size is 714. Standard errors are reported in parentheses.

Probing DD Assumptions

Samples that include many states and years allow us to relax the common trends assumption, that is, to introduce a degree of nonparallel evolution in outcomes between states in the absence of a treatment effect. A regression DD model with controls for state-specific trends looks like

$$\begin{aligned} Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} \\ + \sum_{k=\text{Alaska}}^{\text{Wyoming}} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} \\ + \sum_{k=\text{Alaska}}^{\text{Wyoming}} \theta_k \left(STATE_{ks} \times t\right) + e_{st}. \end{aligned} \tag{5.6}$$

This model presumes that in the absence of a treatment effect, death rates in state k deviate from common year effects by following the linear trend captured by the coefficient θ_k.

Heretofore and hitherto we've been sayin' that DD is all about common trends. How can it be, then, that we're now entertaining models like equation (5.6), which relax the key common trends assumption? To see how such models work, consider a sample of two states: The first, Allatsea, reduced the MLDA to 18 in 1975, while neighboring Alabaster held the line at 21. As a baseline, Figure 5.4 sketches the common trends story. Deaths per 100,000 move in parallel until 1975 (most things got worse in the 1970s, so we show death rates increasing). Death rates also jump above trend in Allatsea in 1975, when that state lowered its MLDA. Given the parallelism and the timing, it seems fair to blame Allatsea's lower MLDA for this jump.

Figure 5.5 sketches a scenario with a steeper trend in Allatsea than in Alabaster. As with the data plotted in the previous figure, simple regression DD estimation in this case generates estimates implicating the MLDA (the post-minus-pre contrast in Allatsea is larger than the post-minus-pre contrast in Alabaster). In this case, however, the resulting DD estimate is spurious: the difference in state trends predates Allatsea's

MLDA liberalization and must therefore be unrelated to it.

Luckily, such differences in trend can be captured by the state-specific trend parameters, θ_k, in equation (5.6). In models that control for state-specific trends, evidence for MLDA effects comes from sharp deviations from otherwise smooth trends, even where the trends are not common. Figure 5.6 shows how regression DD captures treatment effects in the face of uncommon trends. Death rates in Allatsea increase more steeply than in Alabaster throughout the sample period. But the Allatsea increase is especially steep from 1974 to 1975, when Allatsea lowered its MLDA. The coefficient on $LEGAL_{st}$ in equation (5.6) picks this up, while the model allows for the fact that death rates in different states were on different trajectories from the get-go.

FIGURE 5.4

An MLDA effect in states with parallel trends

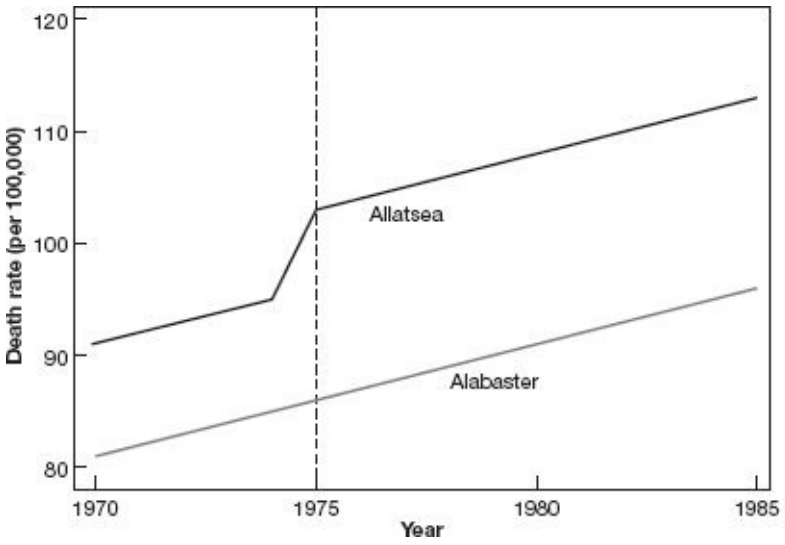


FIGURE 5.5

A spurious MLDA effect in states where trends are not parallel

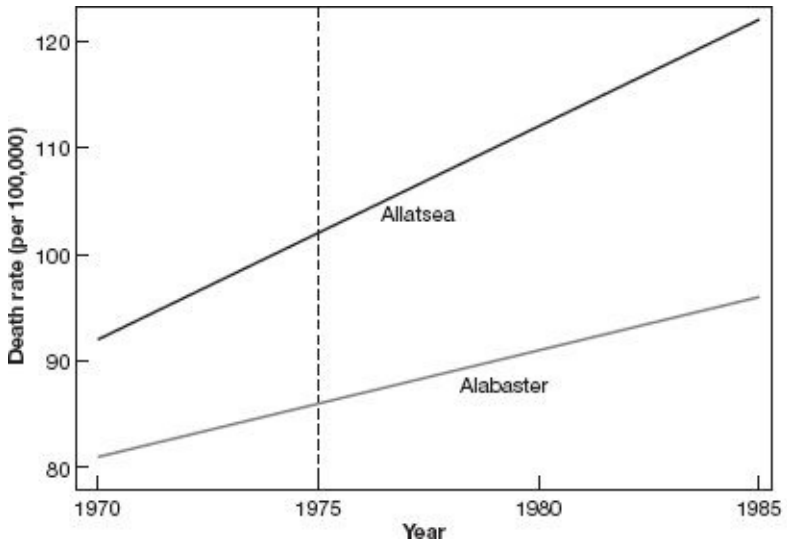
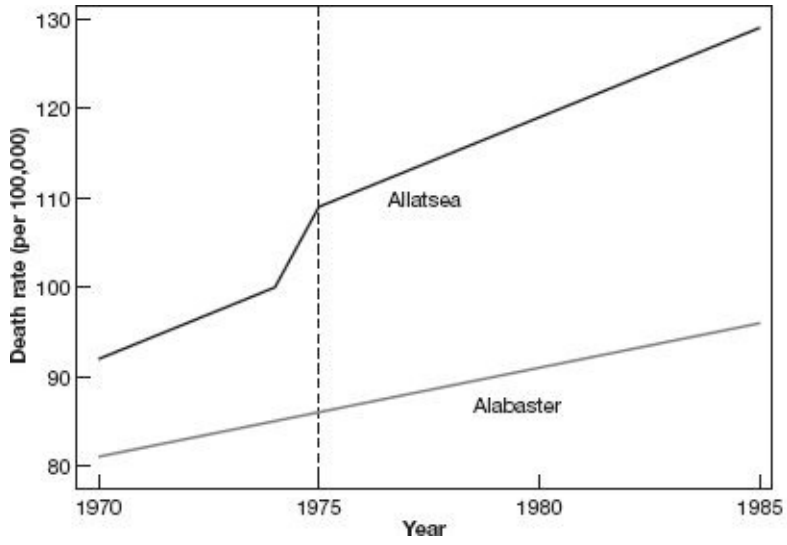


FIGURE 5.6
A real MLDA effect, visible even though trends are not parallel



Models with state-specific linear trends provide an important check on the causal interpretation of any set of regression DD estimates using multiperiod data. In practice, however, empirical reality may be considerably mushier and harder to interpret than the stylized examples laid out in Figures 5.4–5.6. The findings generated by a regression model like equation (5.6) are often imprecise. The sharper the deviation from trend induced by a causal effect, the more likely we are to be able to uncover it. On the other hand, if treatment effects emerge only

gradually, estimates of equations like (5.6) may fail to distinguish treatment effects from differential trends, with the end result being an imprecise and therefore inconclusive set of findings.

Happily for a coherent causal DD analysis of MLDA effects, introduction of state-specific trends has little effect on our regression DD estimates. This can be seen in column (2) of Table 5.2, which reports regression DD estimates of MLDA effects from the model described by equation (5.6). The addition of trends increases standard errors a little, but the loss of precision here is modest. The findings in column (2) support a causal interpretation of the more precise MLDA effects reported in column (1) of the table.

State policymaking is a messy business, with frequent changes on many fronts. DD estimates of MLDA effects, with or without state-specific trends, may be biased by contemporaneous policy changes in other areas. An important consideration in research on alcohol, for example, is the price of a drink. Taxes are the most powerful tool the government uses to affect the price of your favorite beverage. Many states levy a heavy tax on beer, which we measure in dollars per gallon of alcohol content. Beer taxes range from just pennies per gallon to more than a dollar per gallon in some Southern states. Beer taxes change from time to time, mostly increasing, much to the dismay of the Beer Institute (with a tax rate of 2 cents per gallon since 1935, Wyoming is beer bliss). It stands to reason that states might raise tax rates at the same time that they increase their MLDA, perhaps as a part of a broader effort to reduce drinking. If so, we should control for time-varying state tax rates when estimating MLDA effects.

Regression DD models that include controls for state beer taxes generate MLDA estimates similar to those without such controls. This can be seen in Table 5.3, which reports both the estimated coefficients on $LEGAL_{st}$ and the estimated coefficients on state beer taxes in models for the four death rates examined in Table 5.2. Columns (1) and (2) of Table 5.3 show beer tax and MLDA effects estimated using a single regression without controls for state-specific trends, while those in columns (3) and (4) come from another regression including controls for

state-specific trends. Beer tax effects are estimated less precisely than MLDA effects, most likely because beer taxes change less often than the MLDA. The beer tax estimates from models that include state trends are especially noisy. Still, the Beer Institute will be pleased to learn that these results don't speak in favor of further beer tax increases. We're likewise pleased to know that our MLDA estimates are robust to the inclusion of a beer tax control; we'll share a beer to celebrate!

TABLE 5.3

Regression DD estimates of MLDA effects controlling for beer taxes

	Without trends		With trends	
Dependent variable	Fraction legal (1)	Beer tax (2)	Fraction legal (3)	Beer tax (4)
All deaths	10.98	1.51	10.03	−5.52
	(4.69)	(9.07)	(4.92)	(32.24)
Motor vehicle accidents	7.59	3.82	6.89	26.88
	(2.56)	(5.40)	(2.66)	(20.12)
Suicide	.45	−3.05	.38	−12.13
	(.60)	(1.63)	(.77)	(8.82)
Internal causes	1.46	−1.36	.88	−10.31
	(1.61)	(3.07)	(1.81)	(11.64)

Notes: This table reports regression DD estimates of minimum legal drinking age (MLDA) effects on the death rates (per 100,000) of 18–20-year-olds, controlling for state beer taxes. The table shows coefficients on the proportion of legal drinkers by state and year and the beer tax by state and year, from models controlling for state and year effects. The fraction legal and beer tax variables are included in a single regression model, estimated without trends to produce the estimates in columns (1) and (2) and estimated with state-specific linear trends to produce the estimates in columns (3) and (4). The sample size is 700. Standard errors are reported in parentheses.

What Are You Weighting For?

The estimates of equations (5.5) and (5.6) in columns (1) and (2) of Table 5.2 give all observations equal weight, as if data from each state were equally valuable. States are not created equal, however, in at least one important respect: some, like Texas and California, are bigger than most countries, while others, like Vermont and Wyoming, have

populations smaller than those of many American cities. We may prefer estimates that reflect this fact by giving more populous states more weight. The regression procedure that does this is called *weighted least squares* (WLS). The standard OLS estimator fits a line by minimizing the sample average of squared residuals, with each squared residual getting equal weight in the sum.[9] Just as the name suggests, WLS weights each term in the residual sum of squares by population size or some other researcher-chosen weight.

Population weighting has two consequences. First, as noted in Chapter 2, regression models of treatment effects capture a weighted average of effects for the groups or cells represented in our data. In a state-year panel, these groups are states. OLS estimates of models for state-year panels produce estimates of average causal effects that ignore population size, so the resulting estimates are averages over states, not over people. Population weighting generates a people-weighted average, in which causal effects for states like Texas get more weight than those for states like Vermont. People-weighting may sound appealing, but it need not be. The typical citizen is more likely to live in Texas than Vermont, but changes in the Vermont MLDA provide variation that may be just as useful as changes in Texas. You should hope, therefore, that regression estimates from your state-year panel are not highly sensitive to weighting.

Population weighting may also increase the precision of regression estimates. With far fewer drivers in Vermont than in Texas, MVA death rates in Vermont are likely to be more variable from year to year than those in Texas (this reflects the sampling variation discussed in the appendix to Chapter 1). In a statistical sense, the data from Texas are more reliable and therefore, perhaps, worthy of higher weight. Here too, however, the case for weighting is not open and shut. As a matter of econometric theory, masters of 'metrics can claim that weighted estimates are more precise than unweighted estimates only when a number of restrictive technical conditions are met.[10] Once again, the best scenario is a set of findings (that is, estimates and standard errors) that are reasonably insensitive to weighting.

Columns (3) and (4) in Table 5.2 report WLS estimates of equations (5.5) and (5.6). These correspond to the OLS estimates shown in columns (1) and (2) of the table, but the WLS estimator weights each observation by state population aged 18–20. Happily for our understanding of MLDA effects, weighting here matters little. It would seem once again that teetotaling masters have been rewarded for their temperance.

MASTER STEVEFU: Wrap it up for me, Grasshopper.

GRASSHOPPER: Treatment and control groups may differ in the absence of treatment, yet move in parallel. This pattern opens the door to DD estimation of causal effects.

MASTER STEVEFU: Why is DD better than simple two-group comparisons?

GRASSHOPPER: Comparing changes instead of levels, we eliminate fixed differences between groups that might otherwise generate omitted variables bias.

MASTER STEVEFU: How is DD executed with multiple comparison groups and multiple years?

GRASSHOPPER: I have seen the power and flexibility of regression DD, Master. In a state-year panel, for example, with time-varying state policies like the MLDA, we need only control for state and year effects.

MASTER STEVEFU: On what does the fate of DD estimates turn?

GRASSHOPPER: Parallel trends, the claim that in the absence of treatment, treatment and control group outcomes would indeed move in parallel. DD lives and dies by this. Though we can allow for state-specific linear trends when a panel is long enough, masters hope for results that are unchanged by their inclusion.

Masters of 'Metrics: John Snow

British physician John Snow was one of the fathers of modern epidemiology, the study of how illness moves through a population. Studying an outbreak of cholera in London in 1849, Snow challenged the conventional wisdom that the disease is caused by bad air. He thought cholera might be caused by bad water instead, an idea he first laid out in his 1849 essay *On the Mode of Communication of Cholera.*

A further cholera outbreak in 1853 and 1854 claimed many lives in the London neighborhood of Soho. Snow attributed the Soho epidemic to water from a pump on Broad Street. Not afraid to give a natural experiment a helping hand, he convinced the local parish council to remove the handle of the Broad Street pump. Cholera deaths in Soho subsided soon after, though Snow noted that death rates in his Broad Street treatment zone were already declining, and that this made the data from his natural experiment hard to interpret. DD was as fickle at birth as it is today.

Snow was a meticulous data grubber, setting a standard we still aspire to meet. In an 1855 revision of his essay, Snow reported death rates by district and water source for various parts of London. He noted that many of the high-death-rate districts in South London were supplied by one of two companies, the Southwark and Vauxhall Company or the Lambeth Company. In 1849, both companies drew water from the contaminated Thames in central London. Starting in 1852, however, the Lambeth Company drew from the river at Thames Ditton, an uncontaminated water source upstream. Snow showed that between 1849 and 1854 deaths from cholera fell in the area supplied by the Lambeth Company but rose in that supplied by the Southwark and Vauxhall Company. Our Figure 5.7 reproduces Table 12 from Snow's 1855 essay.[11] This table contains the ingredients for Snow's two-period DD analysis of death rates by water source.

Appendix: Standard Errors for Regression DD

Regression DD is a special case of estimation with panel data. A state-

year panel consists of repeated observations on states over time. The repetitive structure of such data sets raises special statistical problems. Economic data of this sort typically exhibit a property called *serial correlation* (that's *serial* as in "murder," not "breakfast"). Serially correlated data are persistent, meaning the values of variables for nearby periods are likely to be similar.

We expect serial correlation in time series data like annual unemployment rates. When a state's unemployment rate is higher than average in one year, it's likely to be higher than average in the next. Because panel data sets combine repeated observations for individual states (in our MLDA example) or regions (in our Mississippi experiment), such data are often serially correlated. When the dependent variable in a regression is serially correlated, the residuals from any regression model explaining this variable are often serially correlated as well. A combination of serially correlated residuals and serially correlated regressors changes the formula required to calculate standard errors.

If we ignore serial correlation and use the simple standard error formula, equation (2.15), the resulting statistical conclusions are likely to be misleading. The penalty for ignoring serial correlation is that you exaggerate the precision of regression estimates. This is because the sampling theory for regression inference laid out in the appendix to Chapter 1 presumes the data at hand come from random samples. Serial correlation is a deviation from randomness, with the important consequence that each new observation in a serially correlated time series contains less information than would be the case if the sample were random.

FIGURE 5.7

John Snow's DD recipe

TABLE XII.

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	
St. Olave	157	161	
St. John, Horsleydown	192	148	
St. James, Bermondsey	249	362	
St. Mary Magdalen	259	244	
Leather Market	226	237	
Rotherhithe*	352	282	Southwark & Vauxhall Company only.
Wandsworth	97	59	
Battersea	111	171	
Putney	8	9	
Camberwell	235	240	
Peckham	92	174	
Christchurch, Southwark	256	113	
Kent Road	267	174	
Borough Road	312	270	
London Road	257	93	
Trinity, Newington	318	210	
St. Peter, Walworth	446	388	Lambeth Company, and Southwark and Vauxhall Compy.
St. Mary, Newington	143	92	
Waterloo Road (1st)	193	58	
Waterloo Road (2nd)	243	117	
Lambeth Church (1st)	215	49	
Lambeth Church (2nd)	544	193	
Kennington (1st)	187	303	
Kennington (2nd)	153	142	
Brixton	81	48	
Clapham	114	165	
St. George, Camberwell	176	132	
Norwood	2	10	
Streatham	154	15	Lambeth Company only.
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

* A small part of Rotherhithe is now supplied by the Kent Water Company.

Just as the robust standard errors discussed in the appendix to Chapter 1 correct for heteroskedasticity, there's a modified standard error formula that answers the serial correlation challenge. The appropriate formula in this case is known as a *clustered standard error*. The formula for clustered standard errors is more complicated than the formula for robust standard errors given in equation (2.16); we won't ask you to learn it for the test. The important thing is that clustering (an option in

most regression software) allows for correlated data within researcher-defined clusters. In contrast with the assumption that all data are randomly sampled, the formula for clustered standard errors requires only that clusters be sampled randomly, with no random sampling assumption invoked for what's inside them.

In the MLDA example discussed in this chapter, states are clusters. Often, it's individual people who appear in our samples repeatedly. Participants in the RAND HIE contributed up to five annual observations on their health-care use in the sample used to construct Table 1.4, and children appear in two separate grades in the sample used to estimate the peer effects model, equation (4.9). In these examples, we adjust for the fact that repeated outcomes for the same person tend to be correlated by clustering on individual.

In the Mississippi experiment, clusters are Federal Reserve Districts. There are only two of these, an important caution. Serial correlation might not be a problem in the Mississippi experiment, but if it is, we'll need more data before we can say anything conclusive about the effects of liquidity on bank survival. Once you start clustering, the formal theory behind statistical inference presumes you have many clusters instead of (or in addition to) many individual observations within clusters. In practice, "many" might be only a few dozen, as with American states. That's probably OK, but a pair or a handful of clusters may not be enough.[12]

Clustered standard errors are appropriate for a wide variety of settings, not only for panel data. In principle, clustering solves any sort of dependence problem in your data (though you might not like the large standard errors that result). For example, data from achievement tests taken by schoolchildren are likely to be correlated within classrooms if children in the same classes share a teacher and have similar family backgrounds. When reporting estimates of the effects of educational interventions like peer effects in equation (4.6) or the effects of private universities in Chapter 2, masters cluster their standard errors on class, school, or university.

[1] Carmen Reinhart and Kenneth Rogoff, *This Time Is Different: Eight Centuries of Financial Folly,* Princeton University Press, 2009; and Milton Friedman and Anna Schwartz, *A Monetary History of the United States, 1867–1960,* Princeton University Press, 1963.

[2] From Chapter IV.4 in Walter Bagehot, *Lombard Street: A Description of the Money Market*, Henry S. King and Co., 1873.

[3] Gary Richardson and William Troost, "Monetary Intervention Mitigated Banking Panics during the Great Depression: Quasi-Experimental Evidence from a Federal Reserve District Border, 1929–1933," *Journal of Political Economy*, vol. 117, no. 6, December 2009, pages 1031–1073. Numbers in this section are our tabulations from the Richardson and Troost data.

[4] In fact, as we explain in the chapter appendix, it's hard to gauge the precision of a DD estimate constructed from only two cross-sectional units and two periods.

[5] Milton Friedman and Rose D. Friedman, *Two Lucky People: Memoirs,* University of Chicago Press, 1998, page 233.

[6] Carpenter and Dobkin, "The Minimum Legal Drinking Age," *Journal of Economic Perspectives,* 2011, analyzed the MLDA in a DD framework.

[7] We include one less time effect than there are years in our data. Time effects measure temporal changes relative to a starting point, usually the first year in the sample.

[8] Here's another way to see how the notation works. Consider an observation for s = NY. Then we have

$$\sum_{k=\text{Alaska}}^{\text{Wyoming}} \beta_k STATE_{ks} = \beta_{\text{NY}},$$

so the sum of all possible state dummies picks up the New York state effect, β_{NY}, when observations are from New York. All the other dummies in the sum are zero. Likewise, if t = 1980, then we have

$$\sum_{j=1971}^{1983} \gamma_j YEAR_{jt} = \gamma_{1980},$$

so the sum picks up the 1980 year effect when observations are from 1980.

[9] Regression residuals, defined in the appendix to Chapter 2, are the differences between the fitted values generated by the model we're estimating and the dependent variable in this model.

[10] One requirement is that the underlying CEF be linear. The appendix to Chapter 2 notes, however, that many regression models are only linear approximations to the CEF.

[11] John Snow, *On the Mode of Communication of Cholera,* John Churchill, second edition, 1855.

[12] For a more detailed discussion of this point, see our book, *Mostly Harmless Econometrics,* Princeton University Press, 2009. In an analysis of hundreds of counties on either side of Federal Reserve District borders, Andrew Jalil adds clusters to the Mississippi experiment. See "Monetary Intervention Really Did Mitigate Banking Panics during the Great Depression: Evidence along the Atlanta Federal Reserve District Border," *Journal of Economic History,* vol. 74, no. 1, March 2014, pages 259–273.