Chapter 4

Regression Discontinuity Designs

YOUNG CAINE: Master, may we speak further on the forces of destiny?

MASTER PO: Speak.

CAINE: As we stand with two roads before us, how shall we know whether the left road or the right road will lead us to our destiny?

MASTER PO: You spoke of chance, Grasshopper. As if such a thing were certain to exist. In the matter you speak of, destiny, there is no such thing as chance.

Kung Fu, Season 3, Episode 62

Our Path

Human behavior is constrained by rules. The State of California limits elementary school class size to 32 students; 33 is one too many. The Social Security Administration won’t pay you a penny in retirement benefits until you’ve reached age 62. Potential armed forces recruits with test scores in the lower deciles are ineligible for American military service. Although many of these rules seem arbitrary, with little grounding in science or experience, we say: bring ’em on! For rules that constrain the role of chance in human affairs often generate interesting experiments. Masters of ’metrics exploit these experiments with a tool called the *regression discontinuity* (RD) design. RD doesn’t work for all causal questions, but it works for many. And when it does, the results have almost the same causal force as those from a randomized trial.

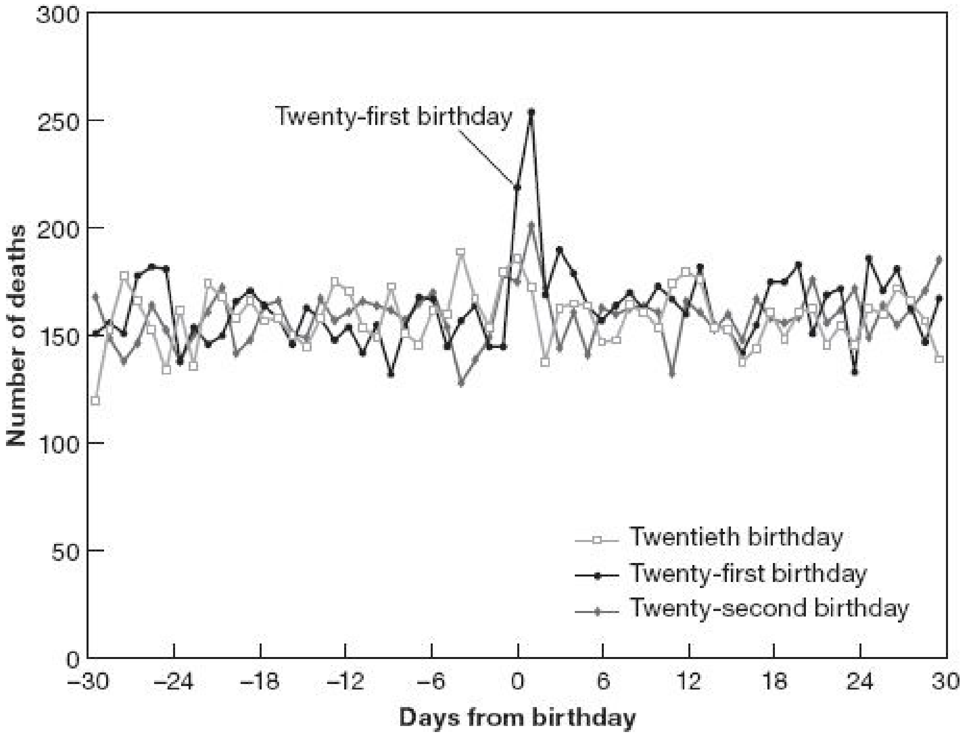
4.1 Birthdays and Funerals

> KATY: Is this really what you're gonna do for the rest of your life?
>
> BOON: What do you mean?
>
> KATY: I mean hanging around with a bunch of animals getting drunk every weekend.
>
> BOON: No! After I graduate, I'm gonna get drunk every night.
>
> *Animal House,* 1978 … of course

Your twenty-first birthday is an important milestone. American over-21s can drink legally, "at last," some would say. Of course, those under age drink as well. As we learn from the exploits of Boon and his fraternity brothers, not all underage drinking is in moderation. In an effort to address the social and public health problems associated with underage drinking, a group of American college presidents have lobbied states to return the minimum legal drinking age (MLDA) to the Vietnamera threshold of 18. The theory behind this effort (known as the Amethyst Initiative) is that legal drinking at age 18 discourages binge drinking and promotes a culture of mature alcohol consumption. This contrasts with the traditional view that the age-21 MLDA, while a blunt and imperfect tool, reduces youth access to alcohol, thereby preventing some harm.

Fortunately, the history of the MLDA generates two natural experiments that can be used for a sober assessment of alcohol policy. We discuss the first experiment in this chapter and the second in the next.[1] The first MLDA experiment emerges from the fact that a small change in age (measured in months or even days) generates a big change in legal access. The difference a day makes can be seen in Figure 4.1, which plots the relationship between birthdays and funerals. This figure shows the number of deaths among Americans aged 20–22 between 1997 and 2003. Deaths here are plotted by day, relative to birthdays, which are labeled as day 0. For example, someone who was born on September 18, 1990, and died on September 19, 2012, is counted among deaths of 22-year-olds occurring on day 1.

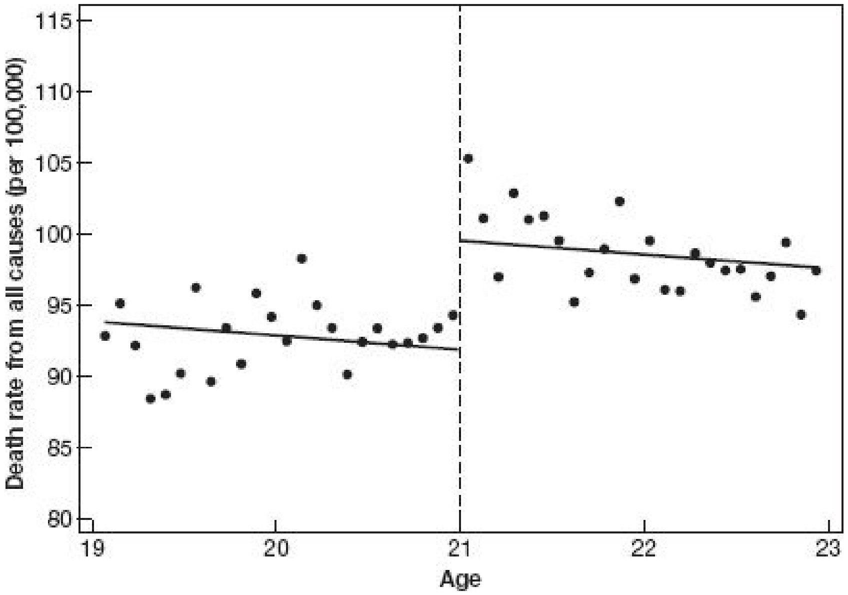
FIGURE 4.1
Birthdays and funerals



Mortality risk shoots up on and immediately following a twenty-first birthday, a fact visible in the pronounced spike in daily deaths on these days. This spike adds about 100 deaths to a baseline level of about 150 per day. The age-21 spike doesn't seem to be a generic party-hardy birthday effect. If this spike reflects birthday partying alone, we should expect to see deaths shoot up after the twentieth and twenty-second birthdays as well, but that doesn't happen. There's something special about the twenty-first birthday. It remains to be seen, however, whether the age-21 effect can be attributed to the MLDA, and whether the elevated mortality risk seen in Figure 4.1 lasts long enough to be worth worrying about.

FIGURE 4.2

A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

Sharp RD

The story linking the MLDA with a sharp and sustained rise in death rates is told in Figure 4.2. This figure plots death rates (measured as deaths per 100,000 persons per year) by month of age (defined as 30-day intervals), centered around the twenty-first birthday. The *X*-axis extends 2 years in either direction, and each dot in the figure is the death rate in one monthly interval. Death rates fluctuate from month to month, but few rates to the left of the age-21 cutoff are above 95. At ages over 21, however, death rates shift up, and few of those to the right of the age-21 cutoff are below 95.

Happily, the odds a young person dies decrease with age, a fact that can be seen in the downward-sloping lines fit to the death rates plotted in Figure 4.2. But extrapolating the trend line drawn to the left of the cutoff, we might have expected an age-21 death rate of about 92, while the trend line to the right of 21 starts markedly higher, at around 99.

The jump in trend lines at age 21 illustrates the subject of this chapter, regression discontinuity designs (RD designs for short). RD is based on the seemingly paradoxical idea that rigid rules—which at first appear to reduce or even eliminate the scope for randomness—create valuable experiments.

The causal question addressed by Figure 4.2 is the effect of legal access to alcohol on death rates. The treatment variable in this case can be written D_a, where $D_a = 1$ indicates legal drinking and is 0 otherwise. D_a is a function of age, a: the MLDA transforms 21-year-olds from underage minors to legal alcohol consumers. We capture this transformation in mathematical notation by writing

$$D_a = \begin{cases} 1 & \text{if } a \geq 21 \\ 0 & \text{if } a < 21. \end{cases} \quad (4.1)$$

This representation highlights two signal features of RD designs:

- Treatment status is a deterministic function of a, so that once we know a, we know D_a.
- Treatment status is a discontinuous function of a, because no matter how close a gets to the cutoff, D_a remains unchanged until the cutoff is reached.

The variable that determines treatment, age in this case, is called the *running variable.* Running variables play a central role in the RD story. In *sharp* RD designs, treatment switches cleanly off or on as the running variable passes a cutoff. The MLDA is a sharp function of age, so an investigation of MLDA effects on mortality is a sharp RD study. The second half of the chapter discusses a second RD scenario, known as *fuzzy RD*, in which the probability or intensity of treatment jumps at a cutoff.

Mortality clearly changes with the running variable, a, for reasons unrelated to the MLDA. Death rates from disease-related causes like cancer (known to epidemiologists as internal causes) are low but increasing for those in their late teens and early 20s, while deaths from external causes, primarily car accidents, homicides, and suicides, fall. To separate this trend variation from any possible MLDA effects, an RD analysis controls for smooth variation in death rates generated by a. RD gets its name from the practice of using regression models to implement this control.

A simple RD analysis of the MLDA estimates causal effects using a regression like

$$\bar{M}_a = \alpha + \rho D_a + \gamma a + e_a, \qquad (4.2)$$

where $\bar{M}_a$ is the death rate in month a (again, month is defined as a 30-day interval counting from the twenty-first birthday). Equation (4.2) includes the treatment dummy, D_a, as well as a linear control for age in months. Fitted values from equation (4.2) produce the lines drawn in Figure 4.2. The negative slope, captured by γ, reflects smoothly declining death rates among young people as they mature. The parameter ρ captures the jump in deaths at age 21. Regression (4.2) generates an estimate of ρ equal to 7.7. When cast against average death rates of around 95, this estimate indicates a substantial increase in risk at the MLDA cutoff.

Is this a credible estimate of the causal effect of the MLDA? Should we not control for other things? The OVB formula tells us that the difference between the estimate of ρ in this short regression and the results any longer regression might produce depend on the correlation between variables added to the long regression and D_a. But equation (4.1) tells us that D_a is determined solely by a. Assuming that the effect of a on death rates is captured by a linear function, we can be sure that no OVB afflicts this short regression.

The lack of OVB in equation (4.2) is the payoff to inside information: although treatment isn't randomly assigned, we know where it comes from. Specifically, treatment is determined by the running variable—an implication of the deterministic link noted above. The question of causality therefore turns on whether the relationship between the running variable and outcomes has indeed been nailed by a regression with a linear control for age.

Although RD uses regression methods to estimate causal effects, RD designs are best seen as a distinct tool that differs importantly from the regression methods discussed in Chapter 2. In Chapter 2, we compared treatment and control outcomes at particular values of the control variables, in the hope that treatment is as good as randomly assigned after conditioning on controls. Here, there is no value of the running variable at which we get to observe both treatment and control observations. Whoa, Grasshopper! Unlike the matching and regression strategies discussed in Chapter 2, which are based on treatment-control comparisons conditional on covariate values, the validity of RD turns on our willingness to extrapolate across values of the running variable, at least for values in the neighborhood of the cutoff at which treatment switches on.

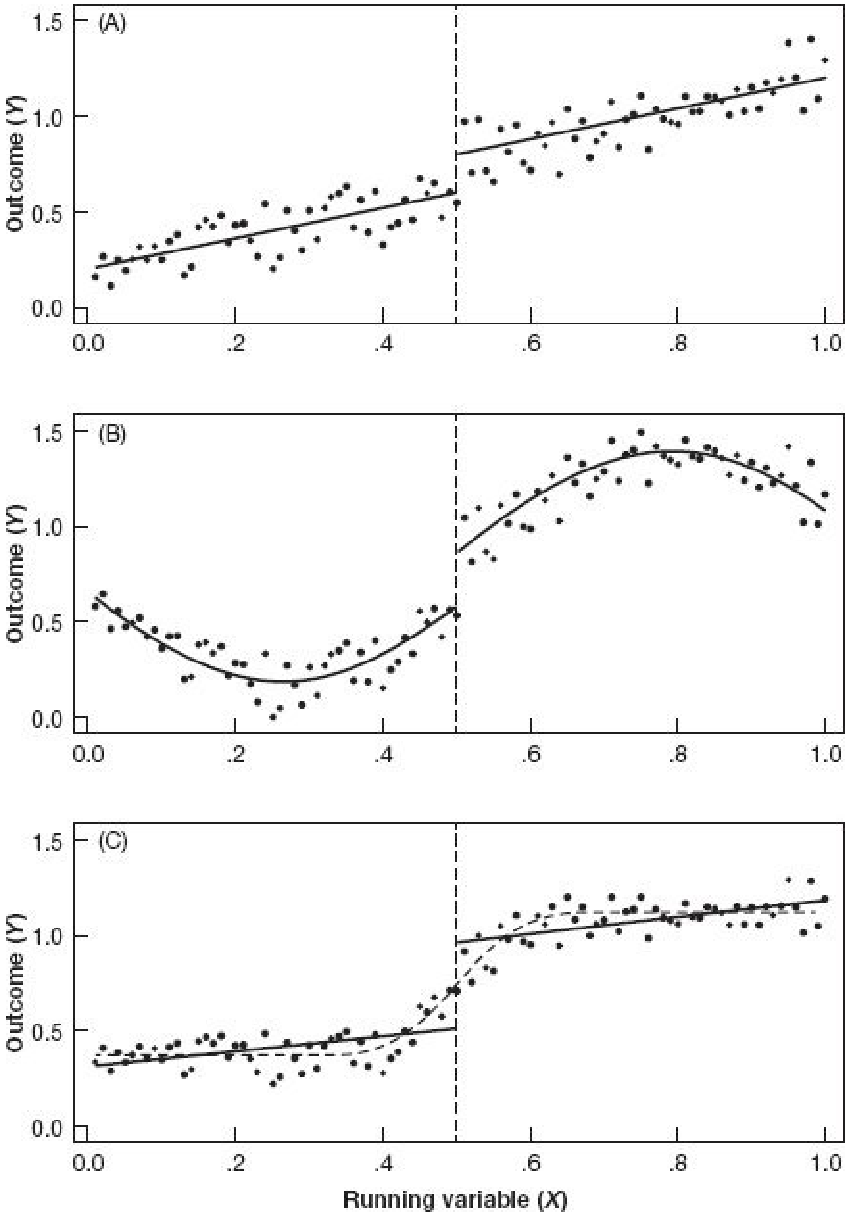
The local nature of such neighborly comparisons is apparent in Figure 4.2. The jump in trend lines at the MLDA cutoff implicitly compares death rates for people on either side of—but close to—a twenty-first birthday. In other words, the notional experiment here involves changes in access to alcohol for young people, in a world where alcohol is freely available to adults. The results from this experiment, though relevant for contemporary discussions of alcohol policy, need not tell us much about the consequences of more dramatic policy changes, such as Prohibition.

RD Specifics

RD tools aren't guaranteed to produce reliable causal estimates. Figure 4.3 shows why not. In panel A, the relationship between the running variable (X) and the outcome (Y) is linear, with a clear jump in $E[Y|X]$ at the cutoff value of one-half. Panel B looks similar, except that the relationship between average Y and X is nonlinear. Still, the jump at $X = .5$ is plain to see. Panel C of Figure 4.3 highlights the challenge RD designers face. Here, the figure exhibits a baroque nonlinear trend, with sharp turns to the left and right of the cutoff, but no discontinuity.

Estimates constructed using a linear model like equation (4.2) mistake this nonlinearity for a discontinuity.

FIGURE 4.3
RD in action, three ways



Notes: Panel A shows RD with a linear model for $E[Y_i|X_i]$; panel B adds some curvature. Panel C shows nonlinearity mistaken for a discontinuity. The vertical dashed line indicates a hypothetical RD cutoff.

Two strategies reduce the likelihood of RD mistakes, though neither provides perfect insurance. The first models nonlinearities directly, while the second focuses solely on observations near the cutoff. We start with the nonlinear modeling strategy, briefly taking up the second approach at the end of this section.

Nonlinearities in an RD framework are typically modeled using polynomial functions of the running variable. Ideally, the results that emerge from this approach are insensitive to the degree of nonlinearity the model allows. Sometimes, however, as in the case of panel C of Figure 4.3, they are not. The question of how much nonlinearity is enough requires a judgment call. A risk here is that you'll pick the model that produces the results that seem most appealing, perhaps favoring those that conform most closely to your prejudices. RD practitioners therefore owe their readers a report on how their RD estimates change as the details of the regression model used to construct them change.

Figure 4.2 suggests the possibility of mild curvature in the relationship between $\bar{M}_a$ and a, at least for the points to the right of the cutoff. A simple extension that captures this curvature uses quadratic instead of linear control for the running variable. The RD model with quadratic running variable control becomes

$$\bar{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a,$$

where $\gamma_1 a + \gamma_2 a^2$ is a quadratic function of age, and the γs are parameters to be estimated.

A related modification allows for different running variable coefficients to the left and right of the cutoff. This modification generates models that interact a with D_a. To make the model with interactions easier to interpret, we center the running variable by subtracting the cutoff, a_0. Replacing a by $a - a_0$ (here, $a_0 = 21$), and adding an *interaction term*, $(a - a_0)D_a$, the RD model becomes

$$\bar{M}_a = \alpha + \rho D_a + \gamma (a - a_0) + \delta [(a - a_0) D_a] + e_a. \quad (4.3)$$

Centering the running variable ensures that ρ in equation (4.3) is still the jump in average outcomes at the cutoff (as can be seen by setting $a = a_0$ in the equation).

Why should the trend relationship between age and death rates change at the cutoff? Data to the left of the cutoff reflect the relationship between age and death rates for a sample whose drinking behavior is restricted by the MLDA. In this sample, we might expect steadily declining death rates as young people mature and take fewer risks. After age 21, however, unrestricted access to alcohol might change this process, perhaps slowing a declining trend. On the other hand, if the college presidents who back the Amethyst Initiative are right, responsible legal drinking accelerates the development of mature behavior. The direction of such a change in slopes is merely a hypothesis—the main point is that equation (4.3) allows for slope changes either way.

A subtle implication of the model with interaction terms is that away from the a_0 cutoff, the MLDA treatment effect is given by $\rho + \delta(a - a_0)$. This can be seen by subtracting the regression line fit to observations where D_a is switched off from the line fit to observations where D_a is switched on:

$$\begin{aligned}&[\alpha + \rho + (\gamma + \delta)(a - a_0)] - [\alpha + \gamma(a - a_0)]\\&\quad = \rho + \delta(a - a_0).\end{aligned}$$

Estimates away from the cutoff constitute a bold extrapolation, however, and should be consumed with a slice of lime and a shaker of salt. There is no data on counterfactual death rates in a world where drinking at ages substantially older than 21 is forbidden. Likewise, far to the left of the cutoff, it's hard to say what death rates would be in a world where drinking at very young ages is allowed. By contrast, it seems reasonable to say that those just under 21 provide a good counterfactual comparison for those just over 21. This leads us to see estimates of the parameter ρ (the causal effect right at the cutoff) as most reliable, even when the model used for estimation implicitly tells us more than that.

Nonlinear trends and changes in slope at the cutoff can also be combined in a model that looks like

$$\begin{aligned}\bar{M}_a = {} & \alpha + \rho D_a + \gamma_1(a - a_0) + \gamma_2(a - a_0)^2 \\ & + \delta_1[(a - a_0)D_a] + \delta_2\left[(a - a_0)^2 D_a\right] + e_a.\end{aligned} \tag{4.4}$$

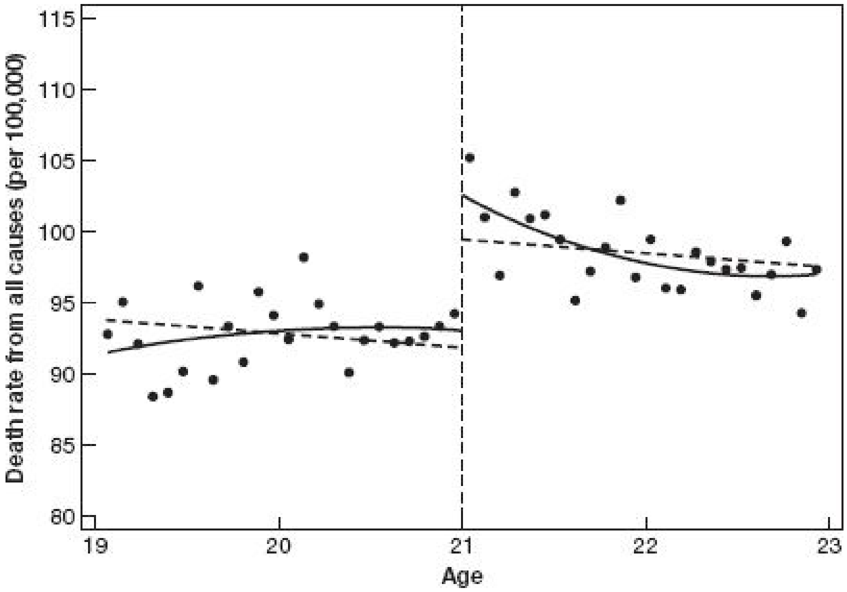
In this setup, both the linear and quadratic terms change as we cross the cutoff. As before, the jump in death rates at the MLDA cutoff is captured by the MLDA treatment effect, ρ. The treatment effect away from the cutoff is now $\rho + \delta_1(a - a_0) + \delta_2(a - a_0)^2$, though again the causal interpretation of this quantity is more speculative than the causal interpretation of ρ itself.

Figure 4.4 shows that the estimated trend function generated by equation (4.4) has some curvature, mildly concave to the left of age 21 and markedly convex thereafter. This model generates a larger estimate of the MLDA effect at the cutoff than does a linear model, equal to about 9.5 deaths per 100,000. Figure 4.4 also shows the linear trend line generated by equation (4.2). The more elaborate model seems to give a better fit than the simple model: Death rates jump sharply at age 21, but then recover somewhat in the first few months after a twenty-first birthday. This echoes the spike in daily death rates on or around the twenty-first birthday seen in Figure 4.1. Unlike Boon and his fraternity brothers, many newly legalized drinkers seem eventually to tire of getting trashed every night. Specification (4.4) captures this jump—and decline—nicely, though at the cost of some technical fanciness.

Which model is better, fancy or simple? There are no general rules here, and no substitute for a thoughtful look at the data. We're especially fortunate when the results are not highly sensitive to the details of our modeling choices, as appears true in Figure 4.4. The simple RD model seems flexible enough to capture effects right at the cutoff, in this case around a twenty-first birthday. The fancier version fits the spike in death rates near twenty-first birthdays, while also capturing the subsequent partial recovery in death rates.

Effects at the cutoff need not be the most important. Suppose we raise the drinking age to 22. In a world where excess alcohol deaths are due entirely to MLDA birthday parties, such a change might extend some lives by a year but otherwise have little effect. The sustained increase in death rates apparent in Figure 4.4 is therefore important, since this suggests restricted alcohol access has lasting benefits. We commented above that evidence for effects away from the cutoff is more speculative than the evidence found in a jump near the cutoff. On the other hand, when the trend relationship between running variable and outcomes is approximately linear, limited extrapolation seems justified. The jump in death rates at the cutoff shows that drinking behavior responds to alcohol access in a manner that is reflected in death rates, an important point of principle, while the MLDA treatment effect extrapolated as far out as age 23 still looks substantial and seems believable, on the order of 5 extra deaths per 100,000. This pattern highlights the value of "visual RD," that is, careful assessment of plots like Figure 4.4.

FIGURE 4.4
Quadratic control in an RD design



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

How convincing is the argument that the jump in Figure 4.4 is indeed due to drinking? Data on death rates by cause of death help us make the case. Although alcohol is poisonous, few people die from alcohol poisoning alone, and deaths from alcohol-related diseases occur only at older ages. But alcohol is closely tied to motor vehicle accidents (MVA), the number-one killer of young people. If drunk driving is the primary alcohol-related cause of deaths, we should see a large jump in motor vehicle fatalities alongside little change in death rates due to internal causes. Like the balancing tests reported for the RAND HIE experiment in Table 1.3 and for the KIPP offer instrument in panel A of Table 3.1, zero effects on outcomes that should be unchanged by treatment raise our confidence in the causal effects we are after.

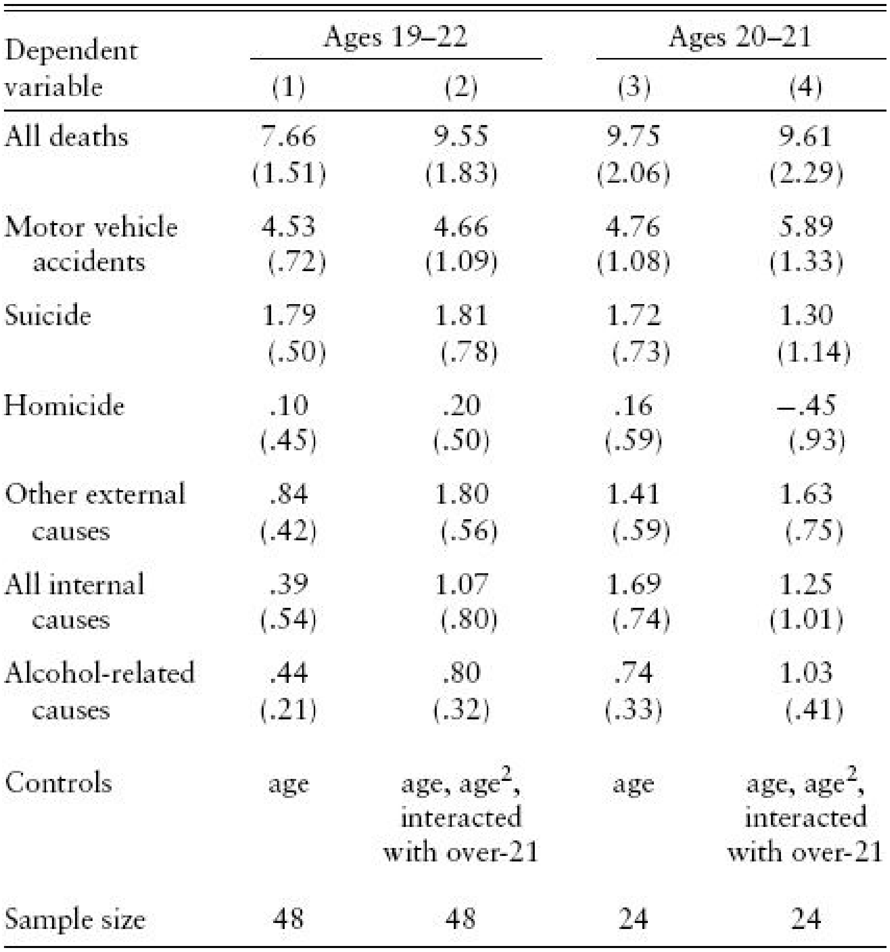
As a benchmark for results related to specific causes of death, the first row of Table 4.1 shows estimates for all deaths, constructed using both simple RD equation (4.2) and fancy RD equation (4.4). These are displayed in columns (1) and (2). The second row of Table 4.1 reveals strong effects of legal drinking on MVA fatalities, effects large enough to account for most of the excess deaths related to the MLDA. The estimates here are largely insensitive to whether the fancy or simple model is used to construct them. Other causes of death we might expect to see affected by drinking are suicide and other external causes, which include accidents other than car crashes. Indeed, estimated effects on suicide and deaths from other external causes (excluding homicide) also show small but statistically significant increases at the MLDA cutoff.

Importantly, the estimates reported in columns (1) and (2) for deaths from all internal causes (these include deaths from cancer and other diseases) are small and and not significantly different from zero. As the last row in the table shows, effects from direct alcohol poisoning also appear to be modest and of roughly the same magnitude as those from internal causes, though the estimated jump in deaths from alcohol poisoning is significantly different from zero. On balance, therefore, Table 4.1 supports the MLDA story, showing clear effects for causes most likely attributable to alcohol but little evidence of an increase due to internal causes.

Also in support of this conclusion, Figure 4.5 plots fitted values for MVA fatalities, constructed using the model that generates the estimates in column (2) of Table 4.1. The figure shows a clear break at the MLDA cutoff, with no evidence of potentially misleading nonlinear trends. At the same time, there isn't much of a jump in deaths due to internal causes, while the standard errors in Table 4.1 suggest that the small jump in internal deaths seen in the figure is likely due to chance.

TABLE 4.1

Sharp RD estimates of MLDA effects on mortality



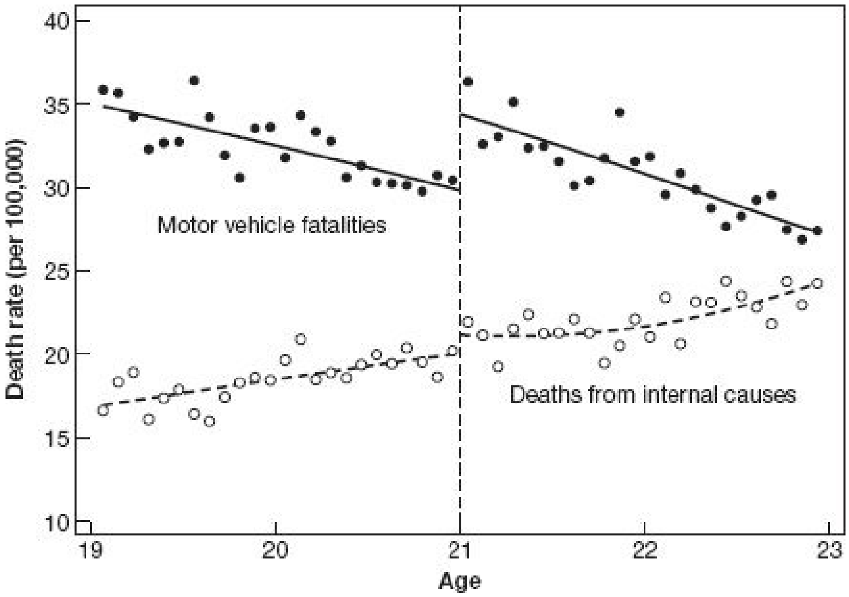
Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66 (1.51)	9.55 (1.83)	9.75 (2.06)	9.61 (2.29)
Motor vehicle accidents	4.53 (.72)	4.66 (1.09)	4.76 (1.08)	5.89 (1.33)
Suicide	1.79 (.50)	1.81 (.78)	1.72 (.73)	1.30 (1.14)
Homicide	.10 (.45)	.20 (.50)	.16 (.59)	−.45 (.93)
Other external causes	.84 (.42)	1.80 (.56)	1.41 (.59)	1.63 (.75)
All internal causes	.39 (.54)	1.07 (.80)	1.69 (.74)	1.25 (1.01)
Alcohol-related causes	.44 (.21)	.80 (.32)	.74 (.33)	1.03 (.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.

In addition to straightforward regression estimation, an approach that masters refer to as *parametric RD,* a second RD strategy exploits the fact that the problem of distinguishing jumps from nonlinear trends grows less vexing as we zero in on points close to the cutoff. For the small set of points close to the boundary, nonlinear trends need not concern us at all. This suggests an approach that compares averages in a narrow window just to the left and just to the right of the cutoff. A drawback here is that if the window is very narrow, there are few observations left, meaning the resulting estimates are likely to be too imprecise to be useful. Still, we should be able to trade the reduction in bias near the boundary against the increased variance suffered by throwing data away, generating some kind of optimal window size.

FIGURE 4.5

RD estimates of MLDA effects on mortality by cause of death



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

The econometric procedure that makes this trade-off is *nonparametric RD*. Nonparametric RD amounts to estimating equation (4.2) in a narrow window around the cutoff. That is, we estimate

$$\bar{M}_a = \alpha + \rho D_a + \gamma a + e_a;$$
$$\textit{in a sample such that } a_0 - b \leq a \leq a_0 + b. \quad (4.5)$$

The parameter b describes the width of the window and is called a *bandwidth*. The results in Table 4.1 can be seen as nonparametric RD with a bandwidth equal to 2 years of age for the estimates reported in columns (1) and (2) and a bandwidth half as large (that is, including only ages 20–21 instead of 19–22) for the estimates shown in columns (3) and (4). The choice of the simple model in equation (4.5) vs. the fancier equation (4.4) should matter little when both are estimated in narrower age windows around the cutoff. The results in Table 4.1 support this conjecture, though there is some wobbliness in the estimates across columns that we might reasonably attribute to sampling variance.[2]

Simple enough! But how shall we pick the bandwidth? On one hand, to obviate concerns about polynomial choice, we'd like to work with data close to the cutoff. On the other hand, less data means less precision. For starters, therefore, the bandwidth should vary as a function of the sample size. The more information available about outcomes in the neighborhood of an RD cutoff, the narrower we can set the bandwidth while still hoping to generate estimates precise enough to be useful. Theoretical econometricians have proposed sophisticated strategies for making such bias-variance trade-offs efficiently, though here too, the bandwidth selection algorithm is not completely data-dependent and requires researchers to choose certain parameters.[3] In practice, bandwidth choice—like the choice of polynomial in parametric models—requires a judgment call. The goal here is not so much to find the one perfect bandwidth as to show that the findings generated by any particular choice of bandwidth are not a fluke.

In this spirit, the studies upon which our investigation of the MLDA is based appear to have been written in RD heaven (perhaps a reward for their authors' temperance). The RD estimates generated by parametric models with alternative polynomial controls come out similar to one another and close to a corresponding set of nonparametric estimates. These nonparametric estimates are largely insensitive to the choice of bandwidth over a wide range.[4] This alignment of results suggests the findings generated by an RD analysis of the MLDA capture real causal effects. Some young people appear to pay the ultimate price for the privilege of downing a legal drink.

4.2 The Elite Illusion

KWAI CHANG CAINE: I seek not to know the answers, but to understand