# Regression Discontinuity (RD) Designs

Sebastián Vallejo Vera
University of Western Ontario

# Reference

This lecture borrows from:

Angrist, J.D. and J.S. Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect.* Princeton University Press. (Chapter 4)

# The Idea behind RD designs

Sometimes nature or institutions/rules provide situations where individuals, cities, states and provinces or any other units are "as-if-randomly" assigned to different states of the world or conditions.

For example:

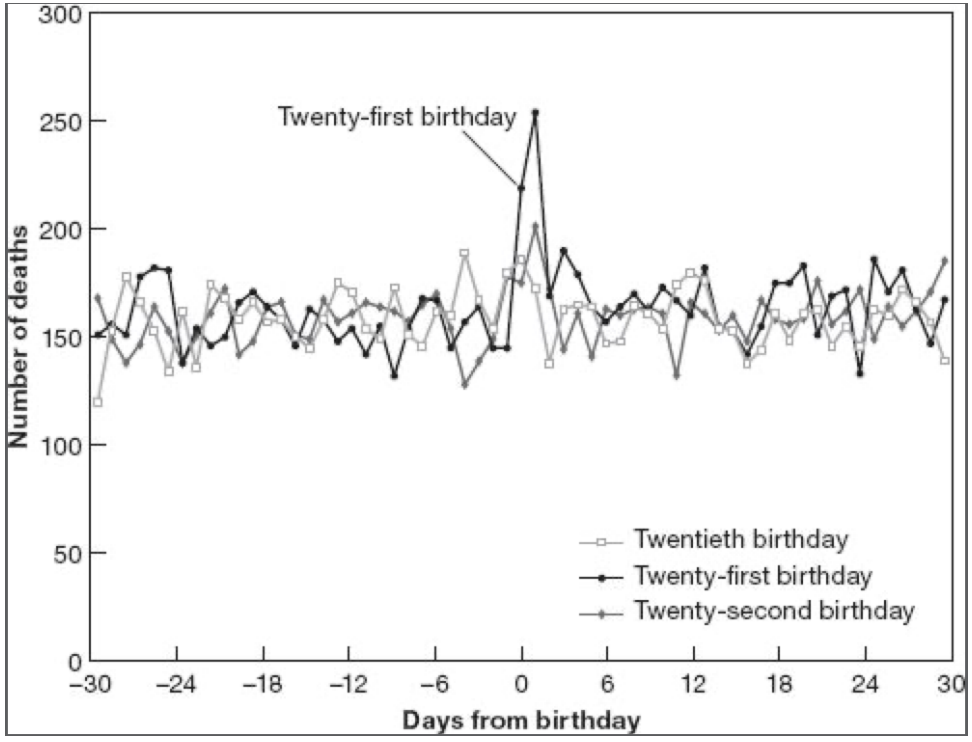
- voters in Brazil gain the right to vote at 16 but voting becomes compulsory only when they turn 18;
- students are admitted into programmes if their GPA or test score reaches a specific threshold;
- electoral candidates can win or lose elections by a handful of votes;
- people can benefit from cash transfer programs if they fall below certain income.

In all of these examples, we can examine how such abrupt and arbitrarily changes affect behaviour.

# The Idea Behind RD designs

The US has a minimum legal drinking age (MLDA) of 21. Some argue that such a late age for drinking encourages binge drinking and causes deaths.

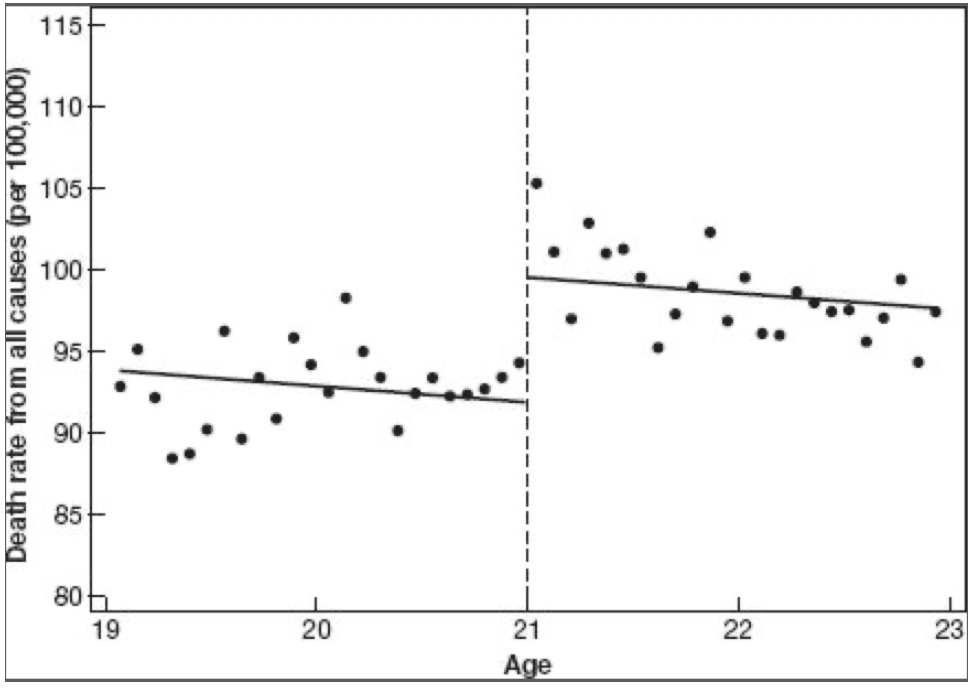
**Figure 1:** Birthdays and deaths



# The Idea behind RD designs

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

**Figure 2:** A sharp RD estimate of MLDA mortality effects



# Sharp RD: definitions and notation

The example of the MLDA and deaths serves to illustrates the causal effect of legal access to alcohol on death rates.

The treatment variable can be written as $D_a$, 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.

Formally,

$$D_a = \begin{cases} 1 & \text{if } a \geq 21 \\ 0 & \text{if } a < 21. \end{cases} \tag{1}$$

This representation illustrates 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.

# Sharp RD: definitions and notation

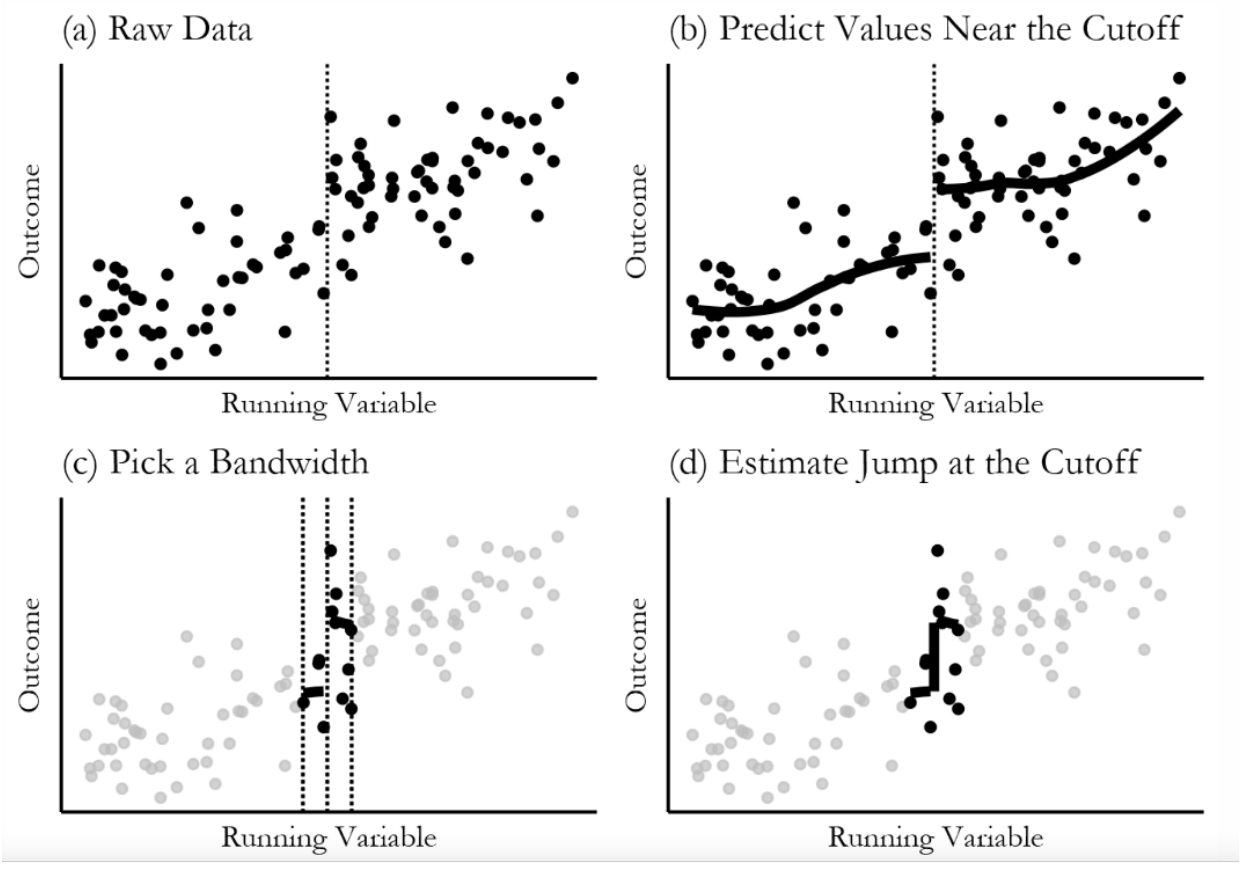
The variable that determines treatment, age ($a$) in this case, is called the *running variable*.

In *sharp* RD designs, treatment switches cleanly off or on as the running variable passes a cutoff.

In *fuzzy* RD designs, it is the probability or intensity of treatment that jumps at a cutoff. In other words, some treated units take the treatment at the cutoff but others do not. In fuzzy RD designs, some untreated units may also take the treatment. (Fuzzy RD designs are not covered here).

# How to RDD

**Figure 3:** Estimating causals effects in an RDD



# Estimation of causal effects with RD designs

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

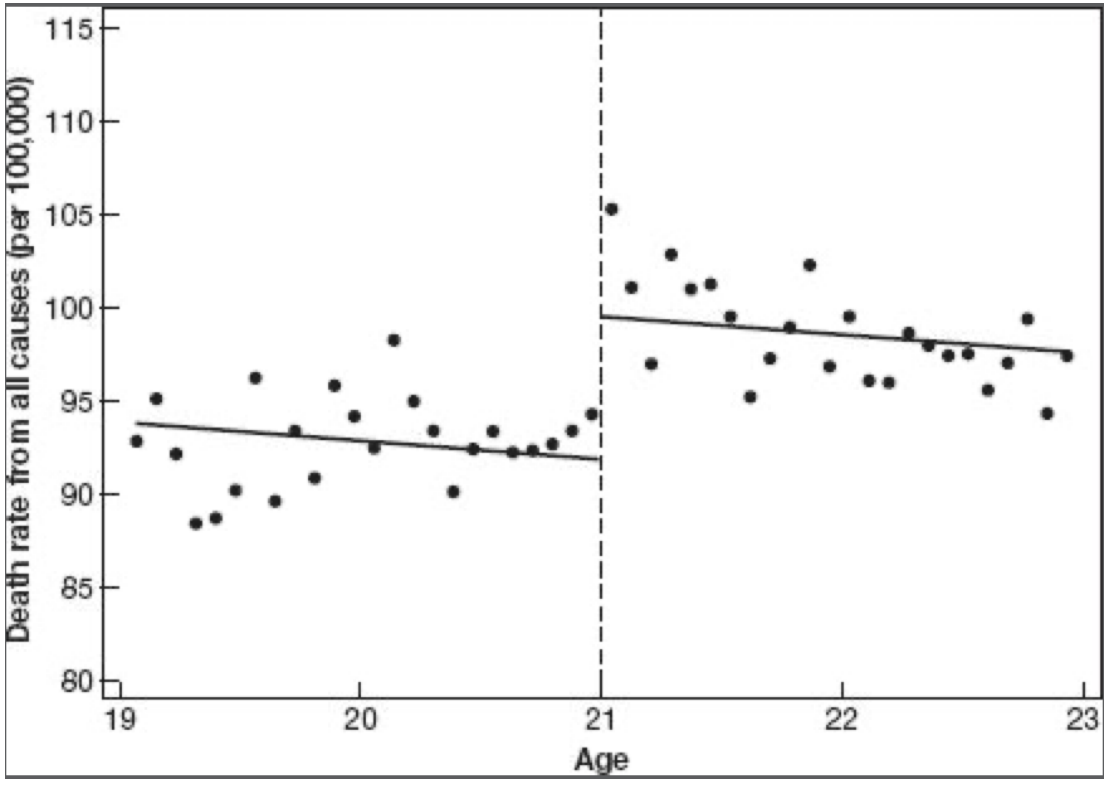
$$\bar{M}_a = \beta_0 + \beta_1 a + \beta_2 D_a + \beta_3 a * D_a + \epsilon \tag{2}$$

where $\bar{M}_a$ is the death rate in a month $a$ (month is defined as a 30-day interval counting from the twenty-first birthday). The equation includes the treatment dummy, $D_a$. The cutoff ($a = 0$) in this case is the twenty-first birthday and the running variable is age, $a$.

# Estimation of causal effects with RD designs

$$\bar{M}_a = \beta_0 + \beta_1 a + \beta_2 D_a + \beta_3 a * D_a + \epsilon \tag{3}$$

**Figure 4:** A sharp RD estimate of MLDA mortality effects



# Estimation of causal effects with RD designs

Fitted values from equation (2) produce the lines in Figure 4. The negative slope, captured by $\beta_3$, reflects smoothly declining death rates among young people as they mature. The parameter $\beta_2$ captures the jump in deaths at age 21.

The regression equation (2) generates an estimate of $\beta_2$ equal to 7.7. When cast against average death rates of around 95 per 100,000 persons per year, this estimate indicates a substantial increases in risk at the MLDA cutoff.

# Estimation of causal effects with RD designs

**Is this a credible estimate of the causal effect of the MLDA?**

**Should we not control for other things?**

We have *omitted variable bias* if the difference between the estimate of $\beta_2$ in this short regression (2) and the results any longer regression might produce depend on the *correlation* between variables added to the long regression and $D_a$.

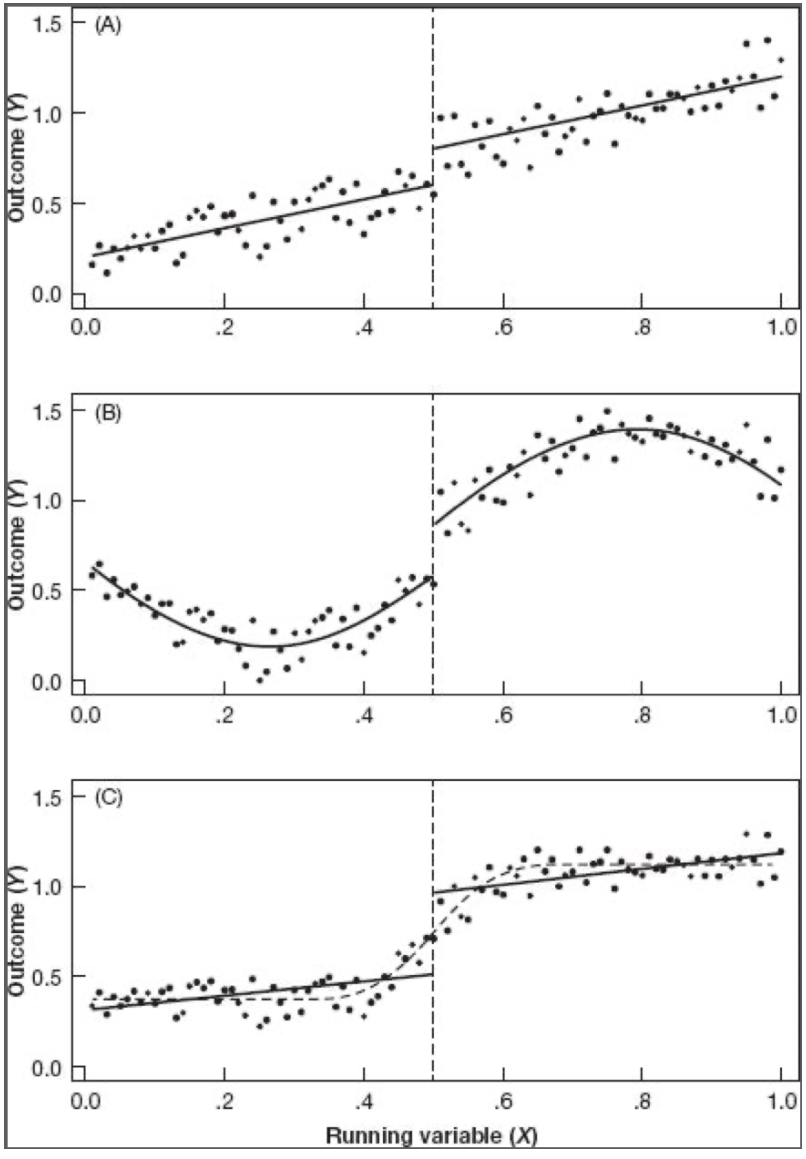
But we know from equation (1) that $D_a$ is determined *solely* by $a$.

Thus, assuming that the effect of $a$ on death rates is captured appropriately (here, as a linear function), we can be sure that no omitted variable bias afflicts this short regression.

In other words, 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.

# RD tools aren't guaranteed to produce reliable causal estimates

**Figure 5:** RD in action, three ways




In all three panels, the vertical dashed line indicates a hypothetical RD cutoff.

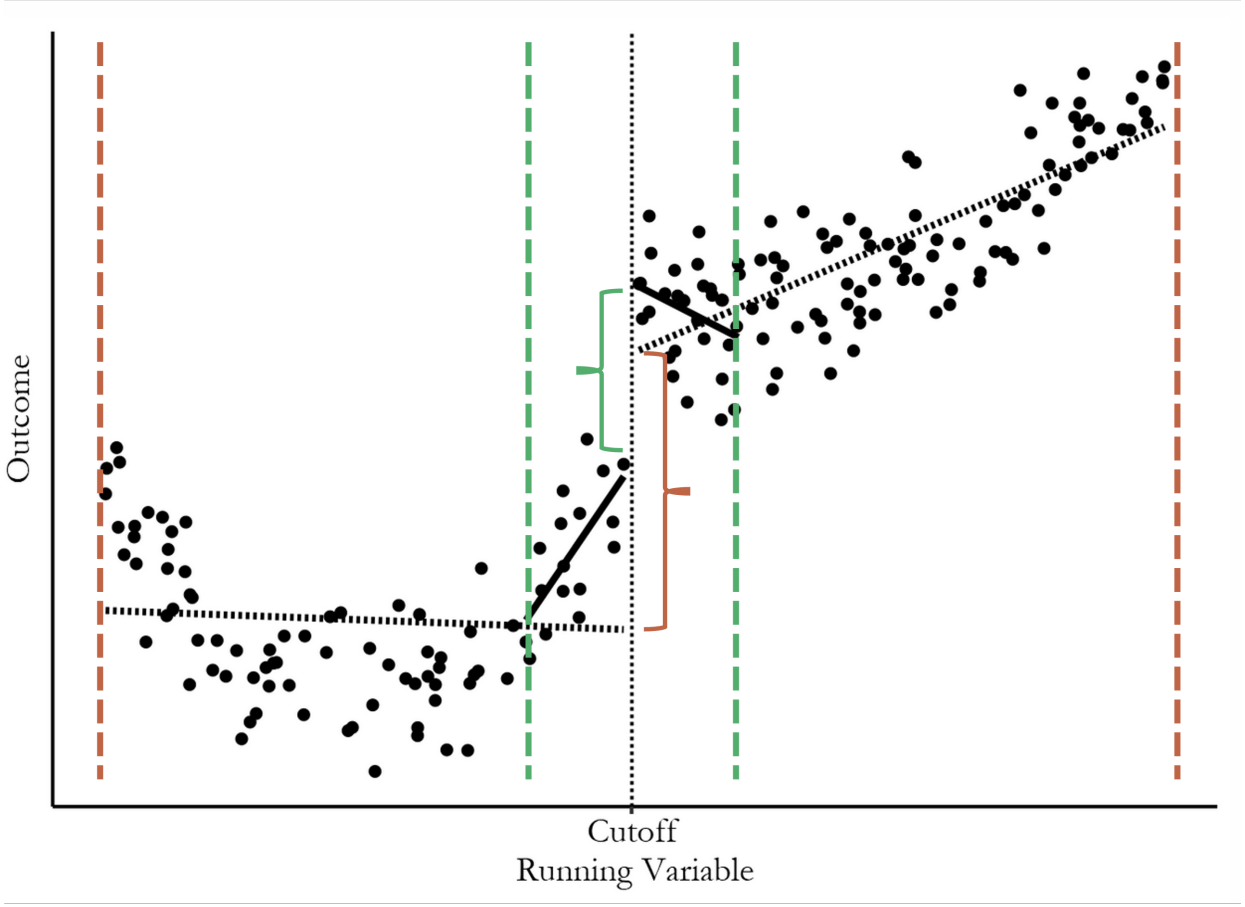
Panel A shows RD with a linear model and we observe a clear jump at $X = .5$.

Panel B adds some curvature but we still observe a clear jump at $X = .5$.

Panel C shows nonlinearity mistaken for a discontinuity. There is no real jump at $X = .5$.

# RD specifics

**Figure 6:** Problems!



# RD specifics

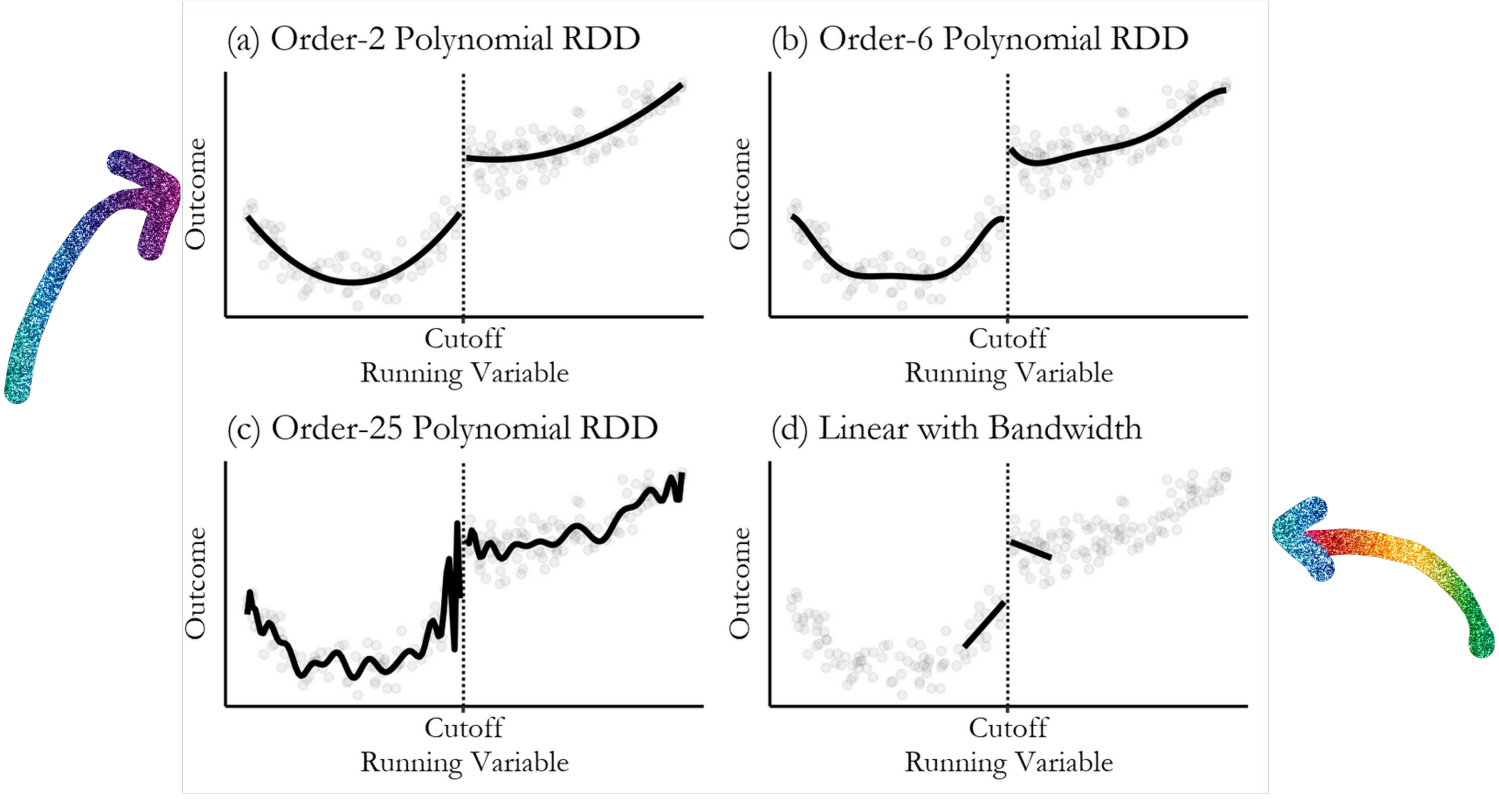
Two strategies reduce the likelihood of RD mistakes, though neither provides perfect insurance.

The first strategy models nonlinearities directly. 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.

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.

# RD specifics

**Figure 7:** Polynomials and RDD



# RD specifics

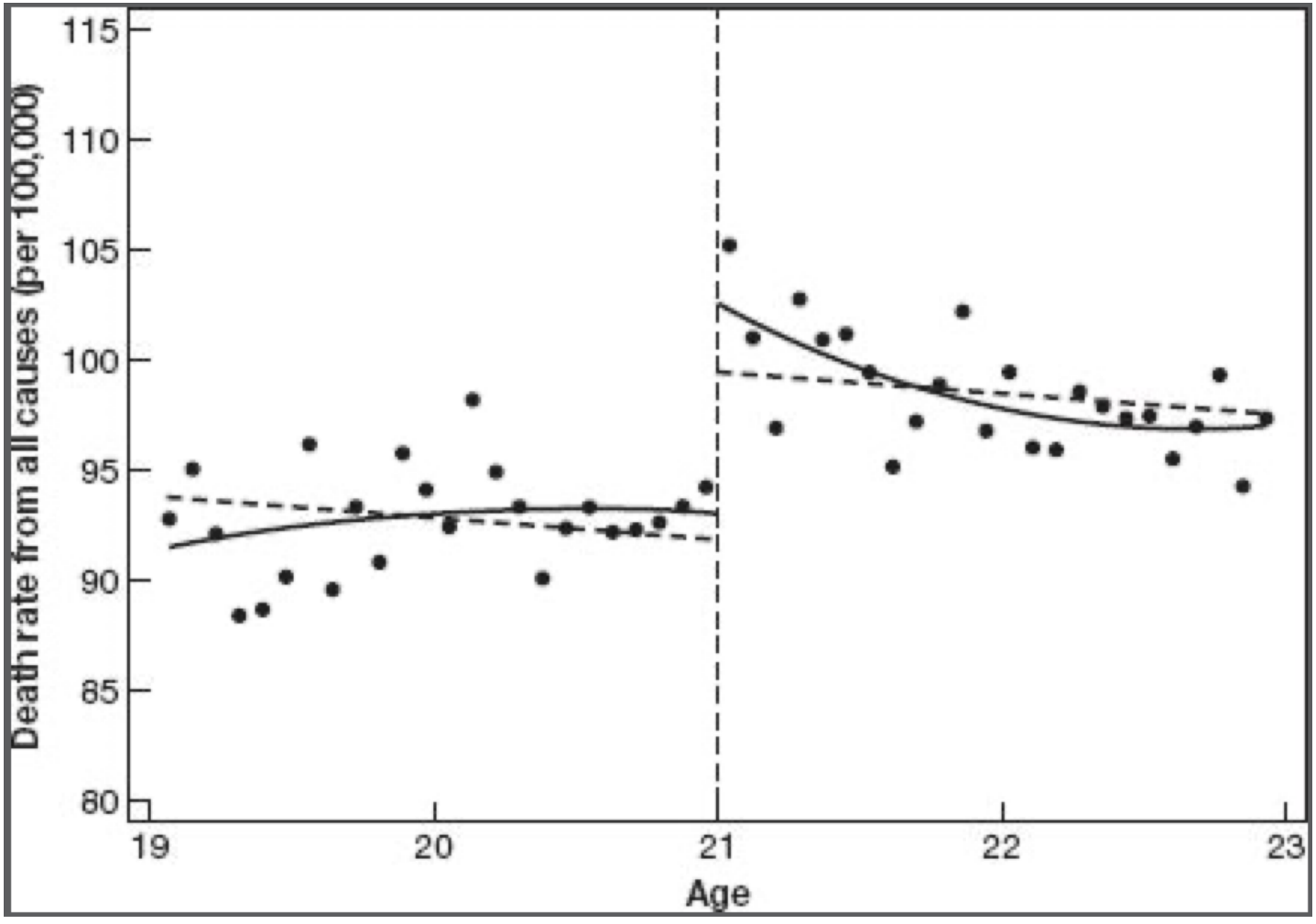
A subtle implication is that estimates away from the cutoff constitute a bold extrapolation of the effect and, thus, should be taken cautiously.

In other words, we are very confident that those *just under the cutoff* provide a good counterfactual comparison for those *just above it*.

We are less confident for those *away* from the cutoff, however, as the counterfactual comparison is not nearly as good. Presumably, other characteristics come into play, reducing our ability to draw a causal inference about the treatment.

# RD specifics

**Figure 8:** Quadratic control on *both* sides of the cutoff (as in equation (5))



# RD specifics

One may wonder, **which model is best?**

The simple model or the more fancy model with curvature (and potentially different trends) on both sides of the cutoff.

There are no general rules and no substitute for a thoughtful look at the data. That said, when the results are *not* highly sensitive to the details of our modeling choices, we know we are onto something real.

# RD specifics

The second strategy one can employ to reduce the likelihood of RD mistakes is to focus solely on observations near the cutoff.

This 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. This approach, however, trades precision (due to the lower number of observations) for lower bias.

# RD specifics

**But how shall we pick the bandwidth?**

On the 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.

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.

# RD specifics

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.

In practice, bandwidth choice—like the choice of polynomial in parametric models—requires a judgment call.

The goal is not so much to find the one perfect bandwidth as to show that the findings are *not* too sensitive to the choice of any particular bandwidth.

# Threats to RD designs

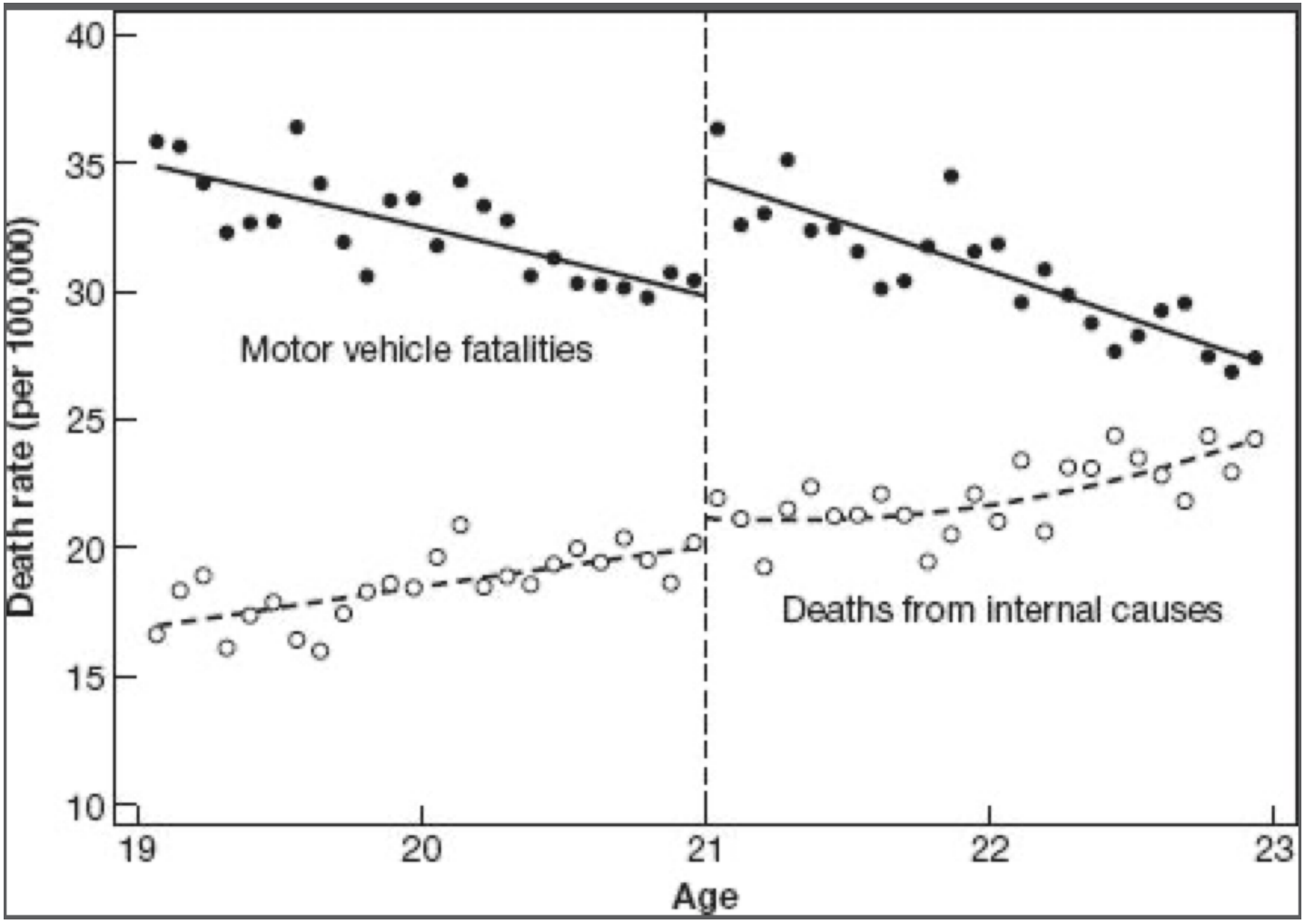
One important threat to RD designs is the possibility that the treatment at the cutoff is confounded with some other factor.

In other words, is the treatment effect observed at the cutoff solely attributed to our treatment of interest or possibly to some other factor? Can the cutoff be manipulated?

Thinking about the MLDA example, how convincing is the argument that the jump in the death rate observed at 21 is indeed due to drinking? Can the drinking age be manipulated? Data on death rates by cause of death help us make the case.

# Threats to RD designs

**Figure 9:** RD estimates of MLDA effects on mortality by cause of death



# Additional robustness checks for RD designs

One common practice in regression discontinuity designs is to check for covariate balance to conclude that the randomization process was "successful." The test constitutes in evaluating whether the units on each side of the cutoff share common attributes beyond being on either side of the cutoff. The expectation is that units slightly to the left and the right of the cutoff are no different.

It is also conventional to conduct a series of placebo tests for jumps at points other than the cutoff (discontinuity) of interest. See Figure 1 among twentieth and twenty-second birthdays.

Finally, it is worth noting one important limitation of RD designs: they allow to estimate causal effects from observational data but the estimates are local, that is, they only speak to those around the cutoff. The analysis of the MLDA, for example, does not tell us anything about drinking and deaths related to drinking among people much older than 21.

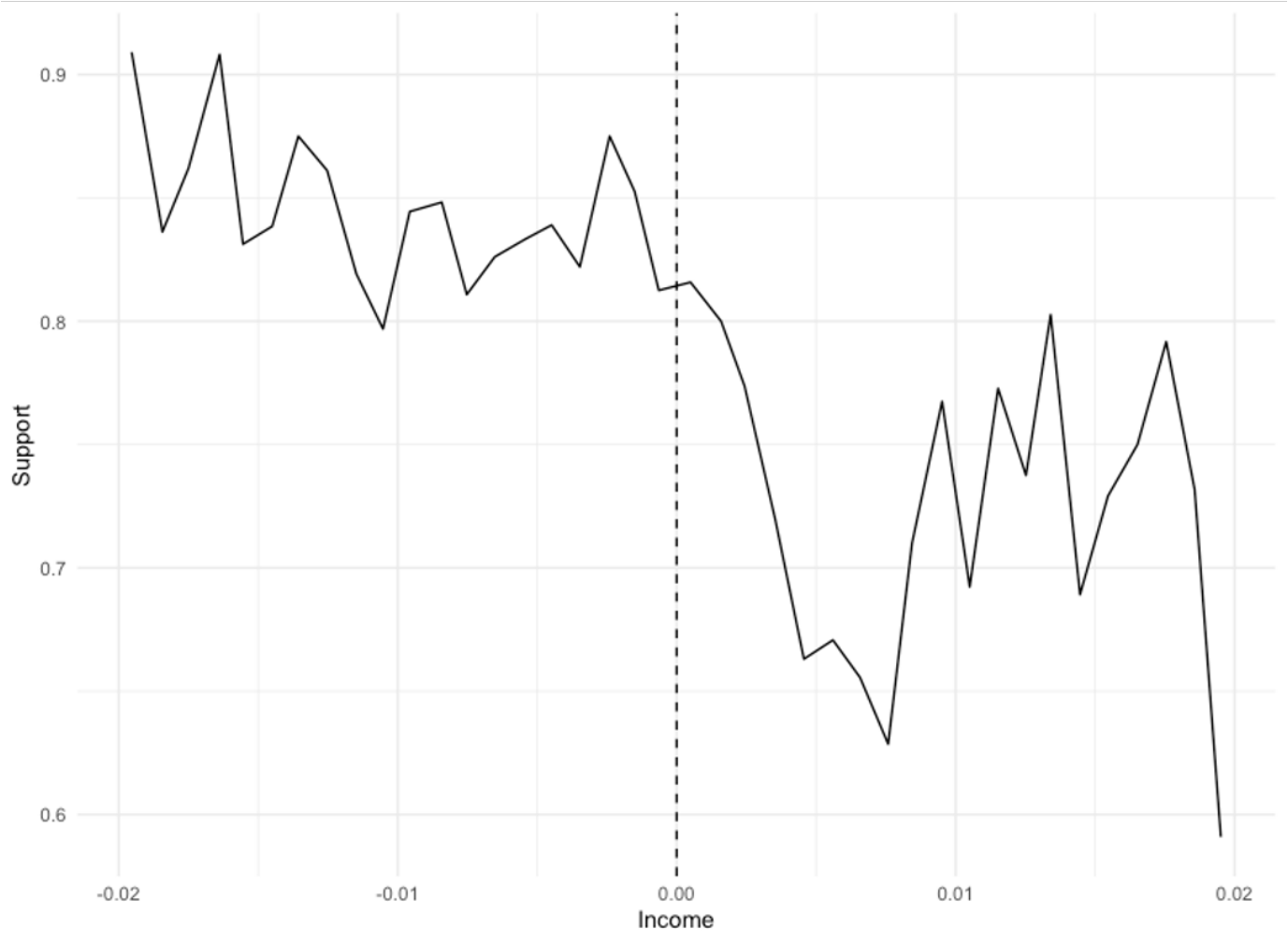
# Example: PANES

In 2007, the Uruguay government implemented a cash transfer program for low-income households (Plan de Atención Nacional a la Emergencia Social, PANES). To access the program, the government used a series of variable to predict income and determine eligibility. Eligibility was determined by earning below a certain income threshold.

Manacorda et al. (2011) take advantage of this discontinuity to see the effect cash transfer programs have on political support to the government party. The researchers used (predicted) income as the running variable and treatment was assigned if the family was below the cutoff. Close to 14% of the population received the transfers.

# Example: PANES

**Figure 10:** Relation between income and government support.



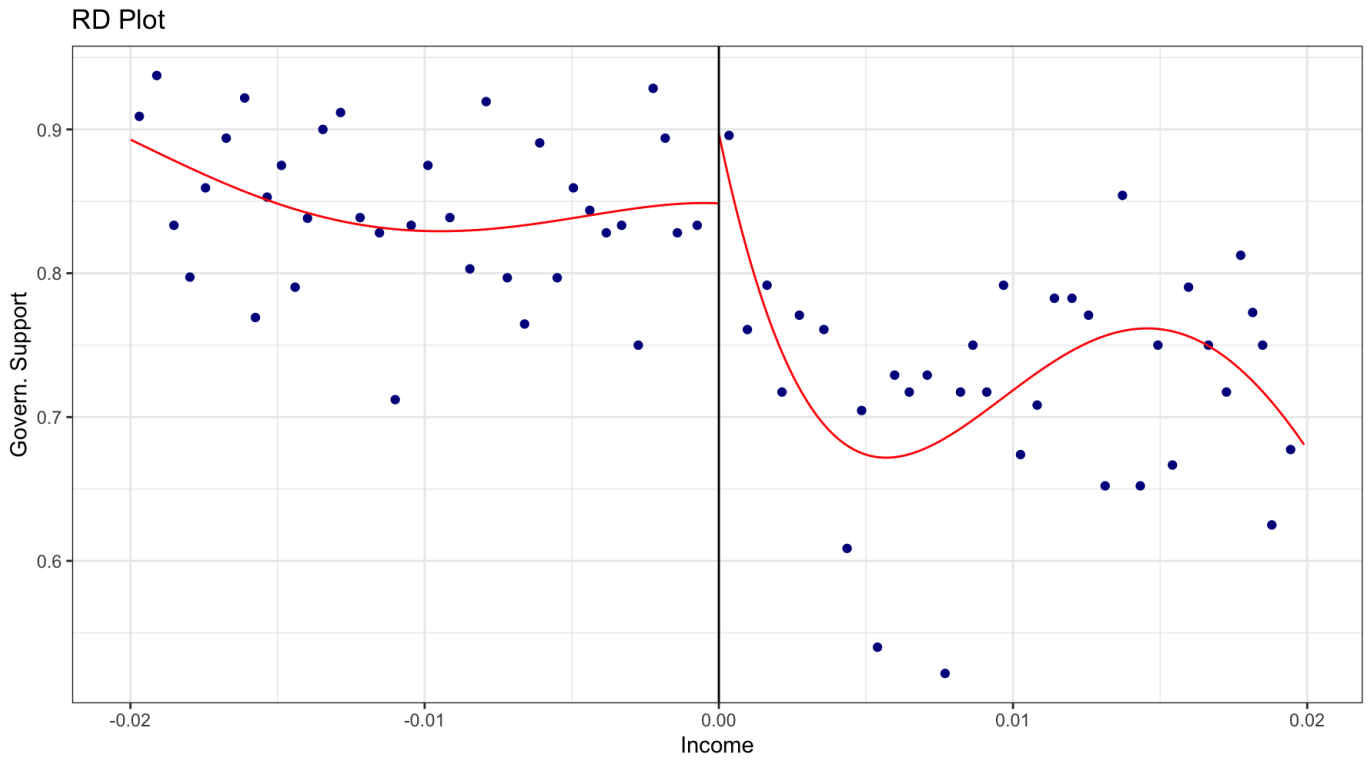
# Example: PANES

**Figure 11:** RDD (Wrong).

```
> m1 <- rdrobust(gt$Support, gt$Income_Centered, c = 0,
+               kernel = "triangular",bwselect = 'certwo')
[1] "Mass points detected in the running variable."
> summary(m1)
Call: rdrobust

Number of Obs.                 1948
BW type                       certwo
Kernel                     Triangular
VCE method                         NN

Number of Obs.                 1127         821
Eff. Number of Obs.             158         121
Order est. (p)                    1           1
Order bias  (q)                   2           2
BW est. (h)                   0.003       0.003
BW bias (b)                   0.009       0.008
rho (h/b)                     0.353       0.389
Unique Obs.                     841         639

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     0.157     0.091     1.728     0.084    [-0.021 , 0.336]
        Robust         -         -     1.744     0.081    [-0.021 , 0.360]
=============================================================================
```

# Example: PANES

**Figure 12:** RDD (Wrong).



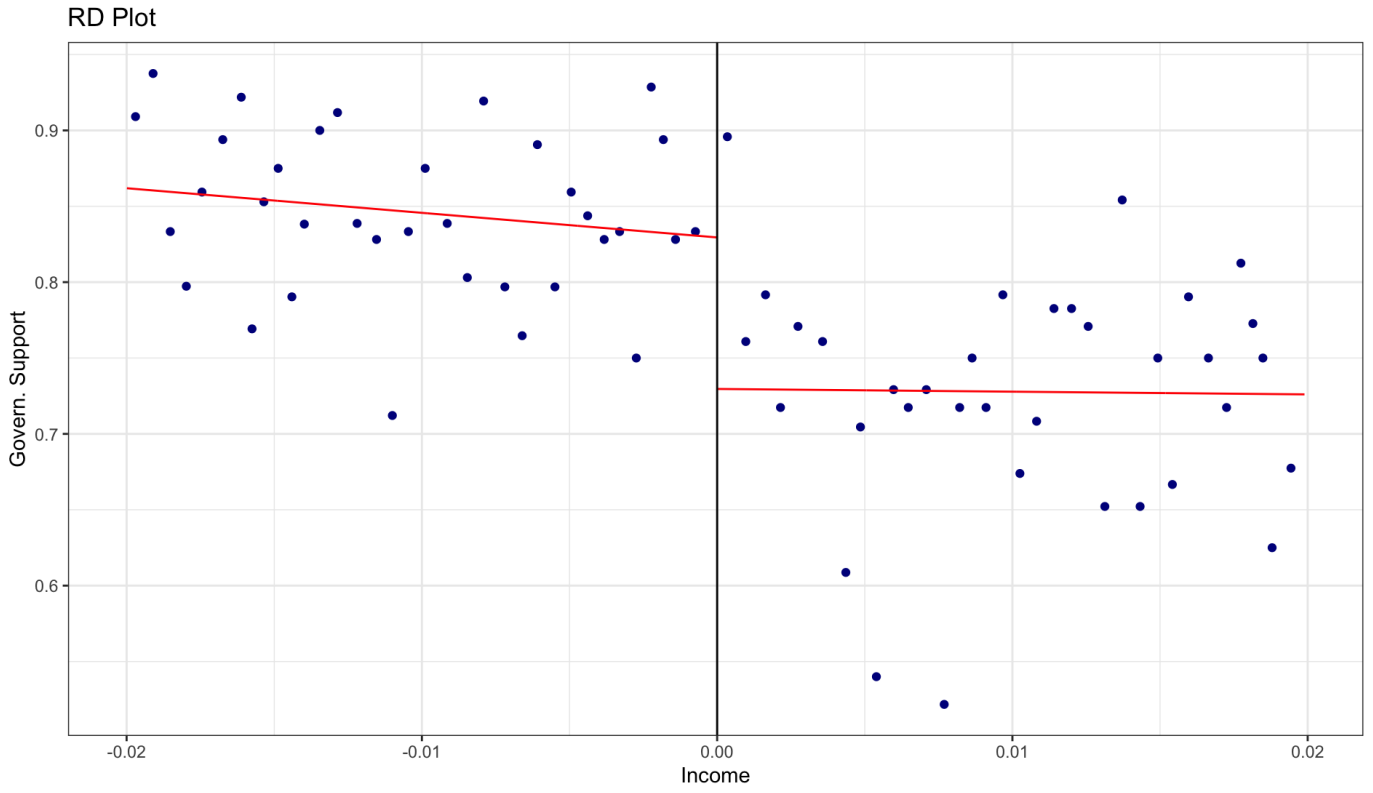
# Example: PANES

**Figure 13:** RDD (Correct).

```
> m2 <- rdrobust(gt$Support, gt$Income_Centered, c = 0,
+                kernel = "triangular",h=1, p=1)
[1] "Mass points detected in the running variable."
> summary(m2)
Call: rdrobust

Number of Obs.                 1948
BW type                      Manual
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 1127          821
Eff. Number of Obs.            1127          821
Order est. (p)                    1            1
Order bias  (q)                   2            2
BW est. (h)                   1.000        1.000
BW bias (b)                   1.000        1.000
rho (h/b)                     1.000        1.000
Unique Obs.                     841          639

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional    -0.100     0.030    -3.367     0.001    [-0.158 , -0.042]
        Robust         -         -    -2.148     0.032    [-0.177 , -0.008]
=============================================================================
```

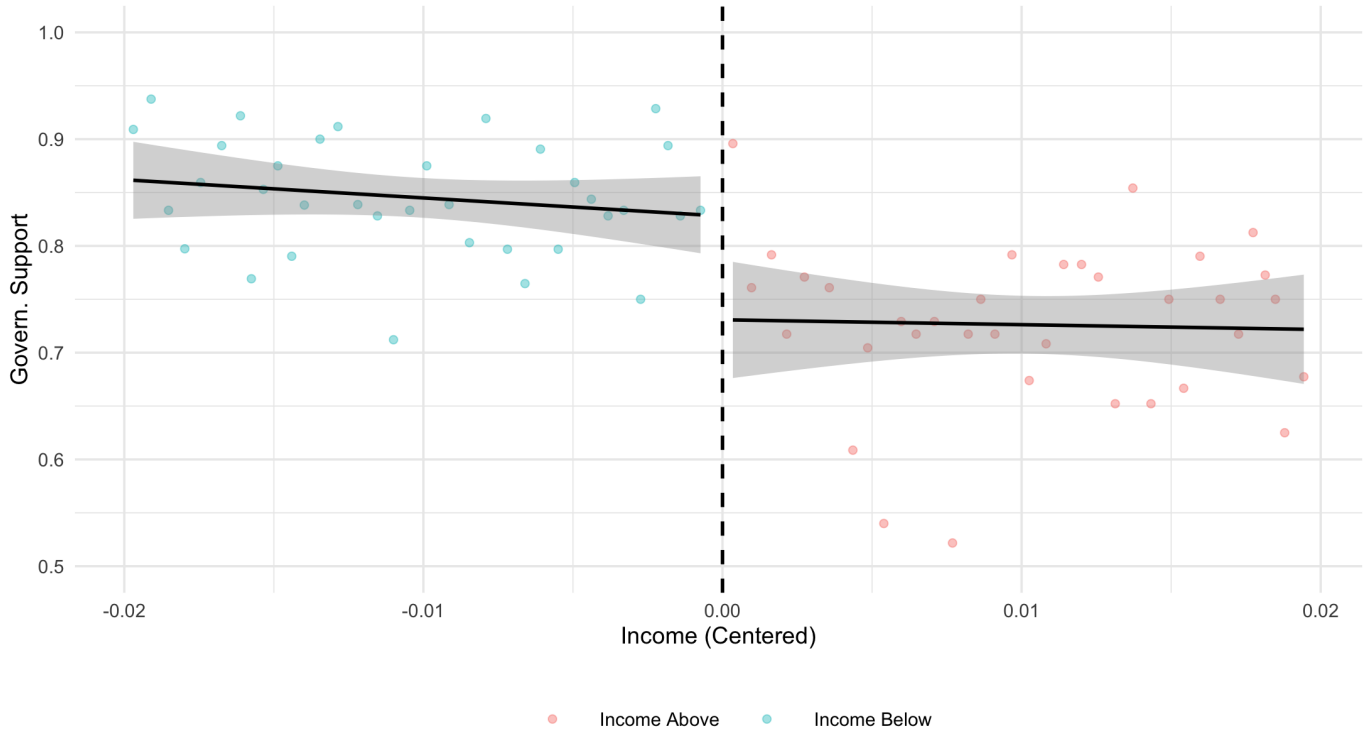
# Example: PANES

**Figure 14:** RDD (Correct).



# Example: PANES

**Figure 15:** RDD (Prettified).



# Resources

There are multiple R packages that can be used to estimate RD models. Notable packages include rdrobust, rdlocrand and others by Calonico and co-authors.

Visit their github page at `https://rdpackages.github.io/`.