Regression Discontinuity

EC 607, Set 9

Edward Rubin
Spring 2020

Prologue

Schedule

Last time

- Introduction to selection-on-unobservables designs
- Instrumental variables (IV) and two-stage least squares (2SLS)

Today

Regression discontinuity [†]

Upcoming

- Problem set; office hours Friday?
- Project

† These notes largely follow notes from Michael Anderson, Imbens and Lemieux (2008), and notes from Teppei Yamamoto.

Regression discontinuity

Regression discontinuity

Setup

We're still in the game of estimating the effect of a potentially endogenous treatment $\mathbf{D}_i$ on an outcome $\mathbf{Y}_i$.

Regression discontinuity (**RD**) offers a particularly clear/clean research design based upon an arbitrary threshold (the *discontinuity*).

That said, most RDs boil down to an implementation of IV.

In addition, while RD is all the rage in modern applied econometrics, Thistlewaite and Campbell wrote about it back in 1960.

Our framework

Back to our potential-outcome framework.

We want to know the effect of $\mathbf{D}_i$ on $\mathbf{Y}_i$.

$$\mathbf{Y}_i = \mathbf{D}_i \mathbf{Y}_{1i} + (1 - \mathbf{D}_i)\mathbf{Y}_{0i}$$

New: Suppose $\mathbf{D}_i$ is determined† by whether some variable $\mathbf{X}_i$ crosses a threshold c (the discontinuity).

The variable $\mathbf{X}_i$ need not be randomly assigned—we will assume it is not (*i.e.*, $\mathbf{X}_i$ correlates with $\mathbf{Y}_{0i}$ and $\mathbf{Y}_{1i}$).

We will assume that $\mathbf{Y}_{0i}$ and $\mathbf{Y}_{1i}$ vary smoothly in $\mathbf{X}_i$.

† At least in part.

Regression discontinuity

Examples

We often apply regression-discontinuity designs in setting with some arbitrary threshold embeded within some bureaucratic decision.

- An elector candidate wins if her vote share exceeds her competitors.
- Election runoffs are triggered if "winner" is below 50%.
- Antidiscrimination laws only apply to firms with >15 employees.
- Prisoners are eligible for early parole if some score exceeds a threshold.
- An individual is eligible for Medicare if her age is at least 65.
- You get a ticket if your speed exceeds the speed limit.
- Fifteen-percent discount at Sizzler if your age exceeds 60.
- Counties with $PM_{2.5}$ > 35 μg/m^3 are *out of attainment*.

In some cases, "treatment" is definite once we exceed the threshold.

Regression discontinuity

Sharp *vs.* fuzzy

We distinguish RDs by how strong/definitive of the threshold is.

In **sharp RDs**, individuals move from control to treatment when their $\mathbf{X}_i$ passes our threshold c, *i.e.*, D_i switches from 0 to 1 when $\mathbf{X}_i$ moves across c.

E.g., a politician wins an election when the difference between her vote share and her competitor's vote share exceeds zero.

In **fuzzy RDs**, the *probability of treatment* $\mathbf{Pr}(\mathrm{D}_i = 1)$ discontinuously jumps at the threshold c, but it does not move from 0 to 1.

E.g., crossing some GRE threshold discontinuously increases your chances of getting into some grad schools (but doesn't guarantee admittance).

Sharp RDs

Sharp RDs

Setup

With **sharp regression discontinuity**, the probability of treatment changes from 0 to 1 as $\mathbf{X}_i$ moves across threshold $\boldsymbol{c}$.

Thus, treatment status totally depends upon whether $\mathbf{X}_i \geq \boldsymbol{c}$, *i.e.*,

$$\mathbf{D}_i = \mathbb{I}\{\mathrm{X}_i \geq \boldsymbol{c}\}$$

To estimate the causal effect of $\mathbf{D}_i$ on $\mathbf{Y}_i$, we compare the mean of $\mathbf{Y}_i$ just *above* the threshold to the mean of $\mathbf{Y}_i$ just *below* the threshold.

Sharp RDs

More formally

We can write the comparison of means at the threshold as

$$\lim_{x \downarrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x] - \lim_{x \uparrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x]$$

$$= \lim_{x \downarrow c} E[\mathrm{Y}_{1i} \mid \mathrm{X}_i = x] - \lim_{x \uparrow c} E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = x]$$

$$= \tau_{\mathrm{SRD}}$$

Assumption $E[\mathrm{Y}_{1i} \mid \mathrm{X}_i = x]$ and $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = x]$ are continuous in x.

$$\implies \tau_{\mathrm{SRD}} = E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \mid \mathrm{X}_i = c]$$

I.e., Because we don't observe Y_{0i} for treated individuals, we extrapolate $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = c - \varepsilon]$ to $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = x + \varepsilon]$ for small ε.

Sharp RDs

Estimation

Thus, we estimate

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x] - \lim_{x \uparrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x]$$

as the diffrence between two regression functions estimated "near" c.

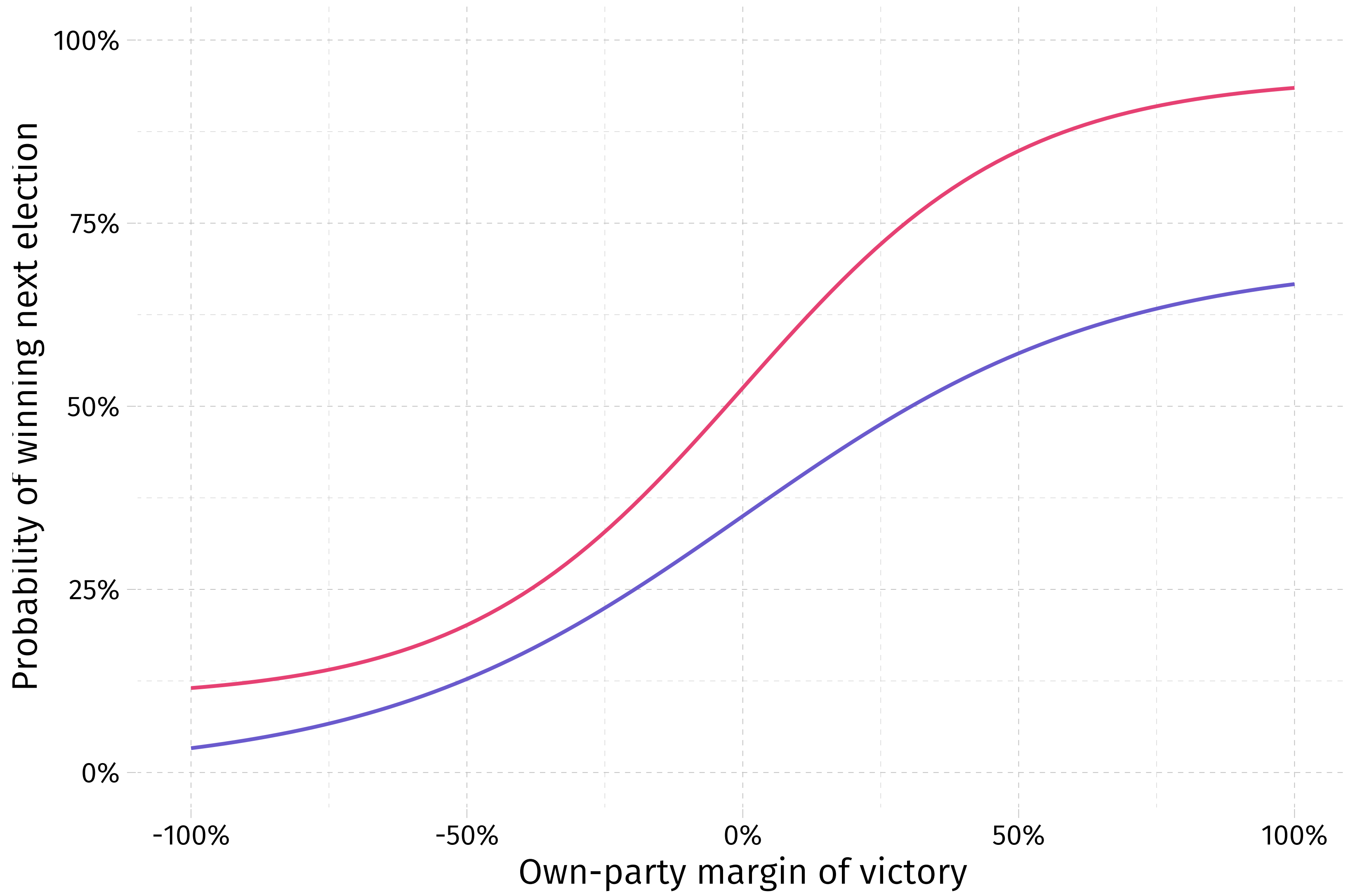
We must stay "near" to c to minimize the bias from extrapolating $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = c - \varepsilon]$ to $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = c + \varepsilon]$ (and assuming continuity).

Ex. Is there effect of a political party winning an election on that party's likelihood of winning the following election?

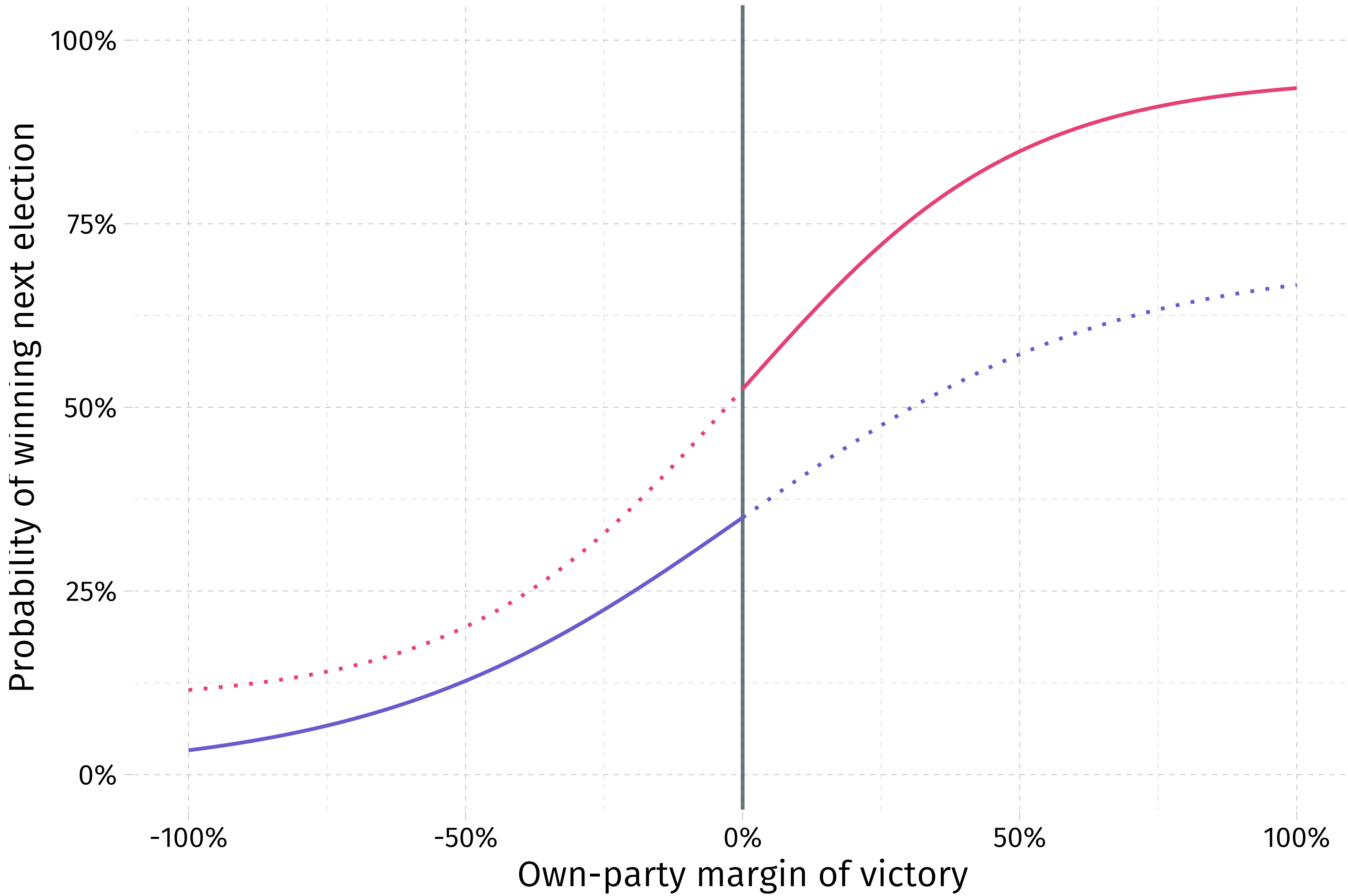
Is there a benefit of incumbency (at the party level)?†

† Lee (2008) addresses this question via RD. Caughey and Sekhon (2011) discuss RD in this setting.

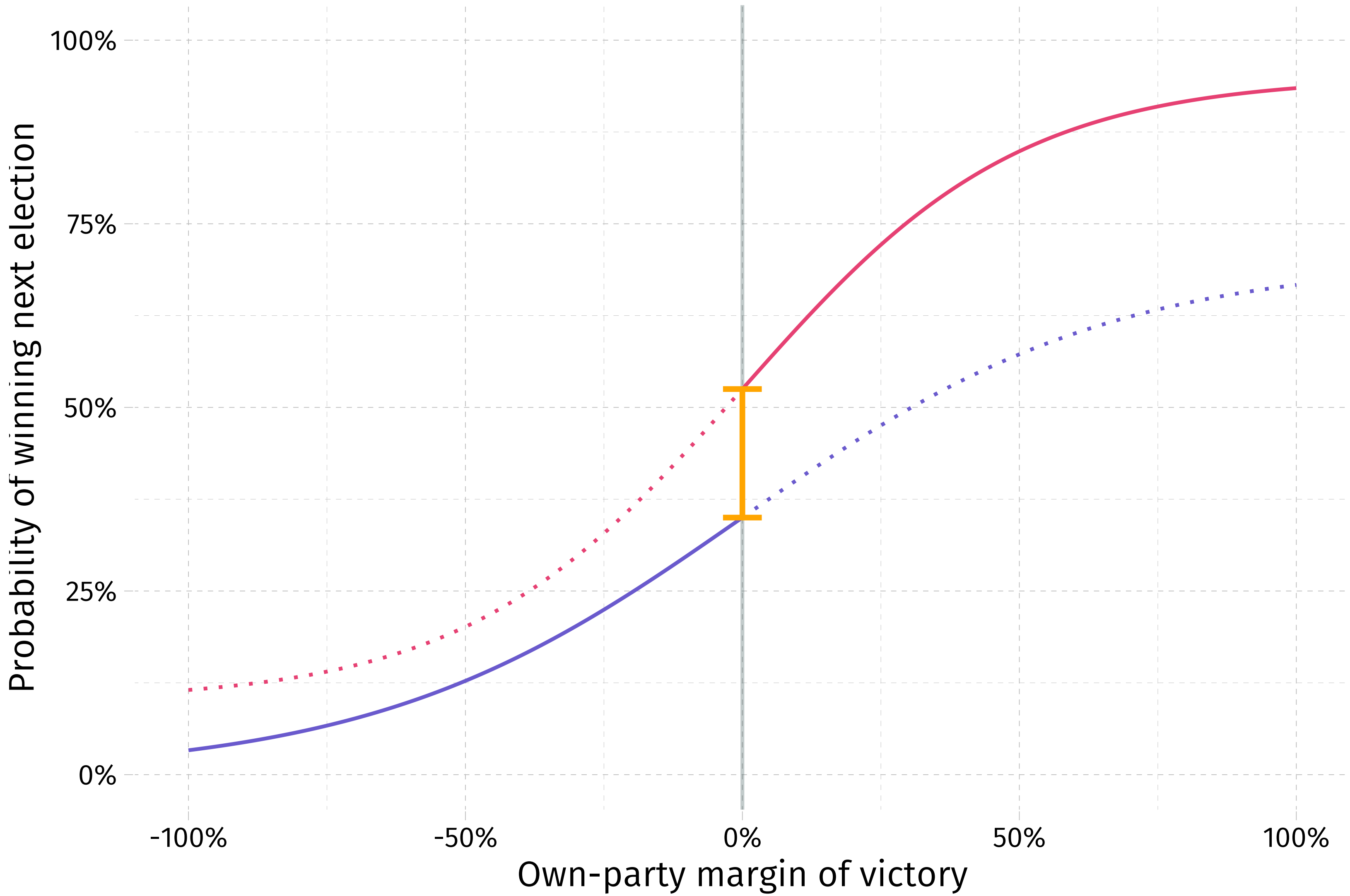
Let's start with $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i]$ and $E[\mathrm{Y}_{1i} \mid \mathrm{X}_i]$.



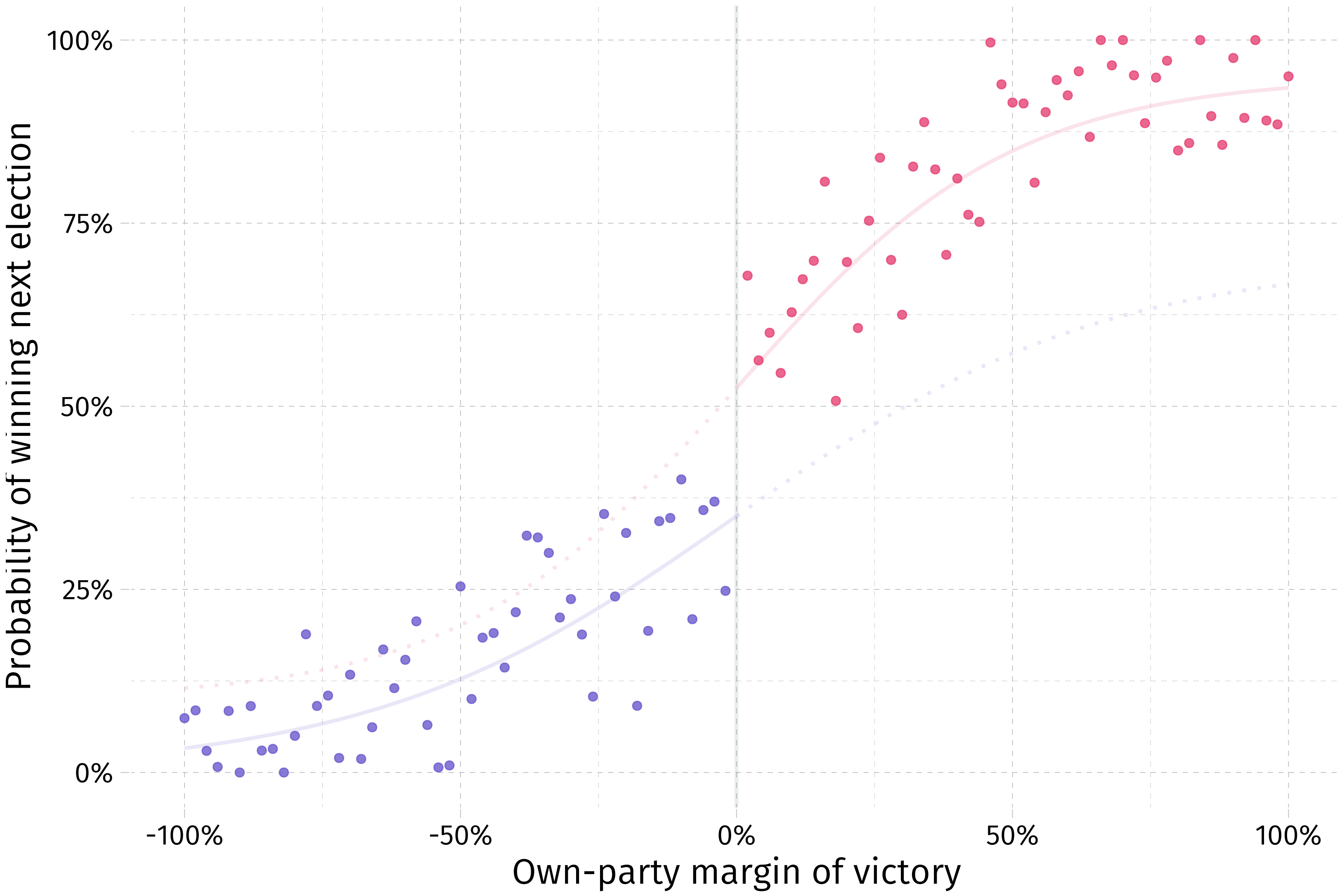
You only win an election if your **margin of victory exceeds zero**.



$E[\mathrm{Y}_{1i} \mid \mathrm{X}_i] - E[\mathrm{Y}_{0i} \mid \mathrm{X}_i]$ **at the discontinuity** gives τ_{SRD}.



Real data are a bit trickier. We must estimate $E[\mathrm{Y}_{1i} \mid \mathrm{X}_i]$ and $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i]$.

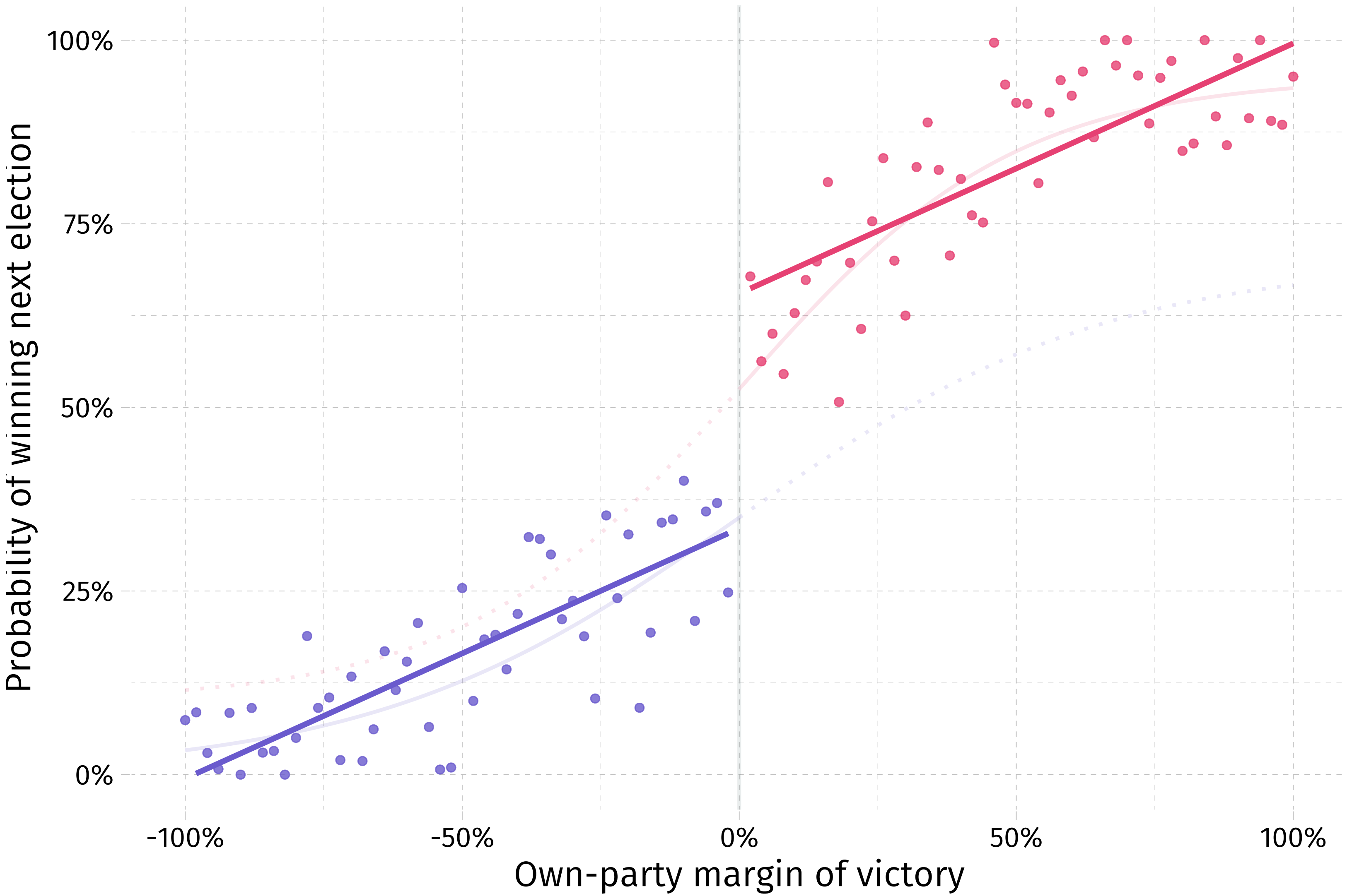


Questions

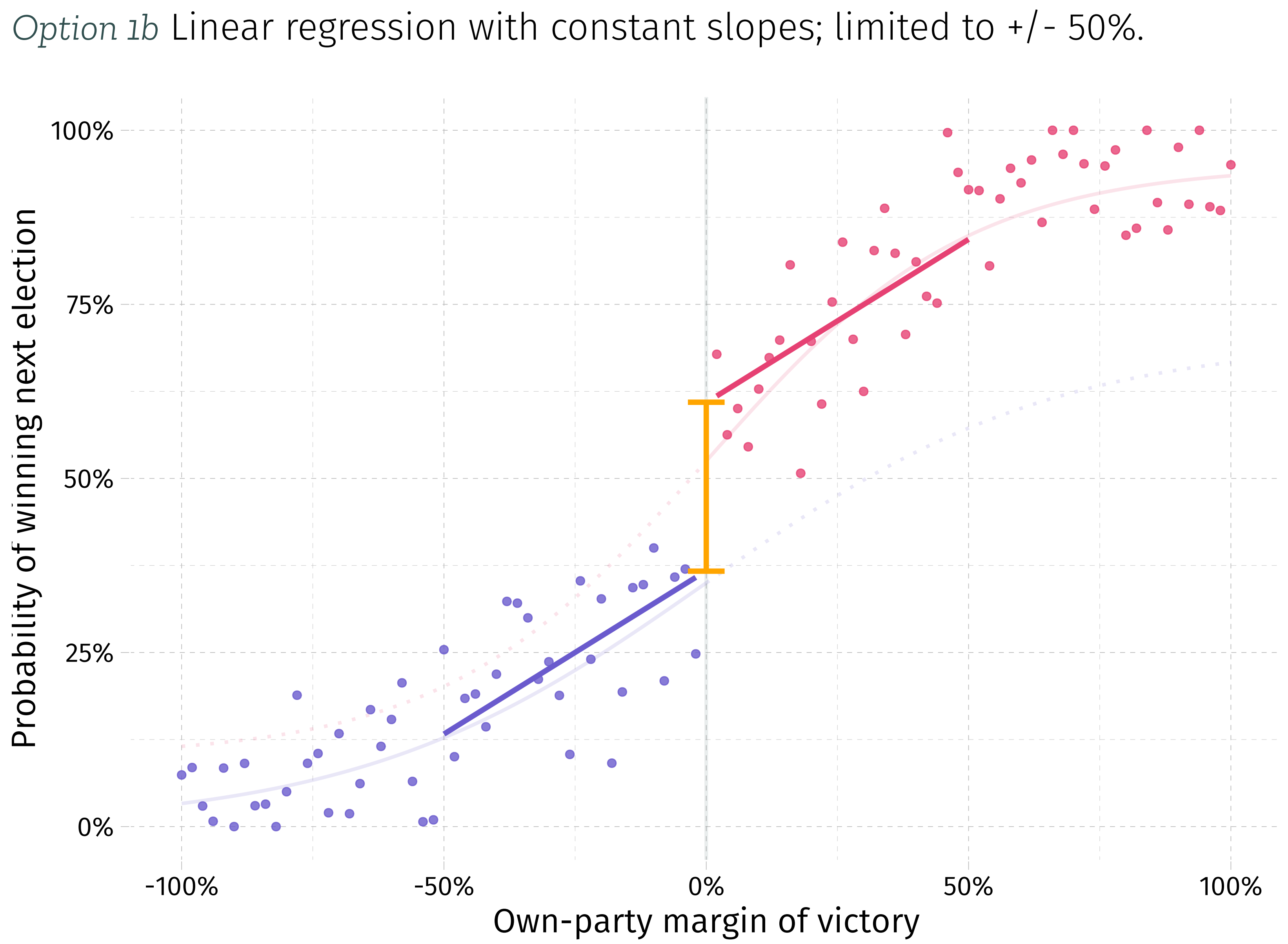
1. How should we estimate $E[\mathbf{Y}_{1i} \mid \mathbf{X}_i]$ and $E[\mathbf{Y}_{0i} \mid \mathbf{X}_i]$?

2. How much data should we use—*i.e.*, what is the right **bandwidth** size?

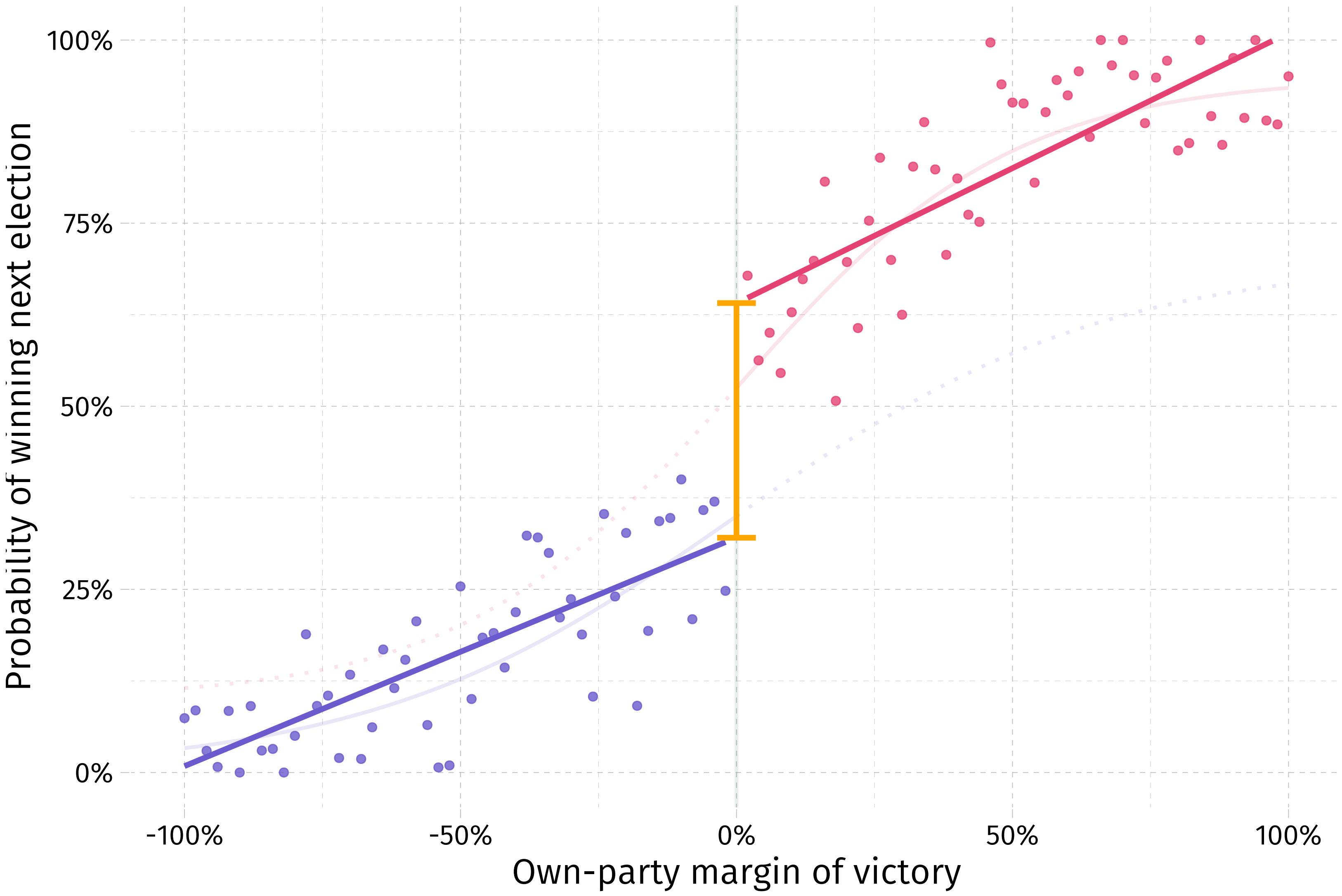
Option 1a Linear regression with constant slopes (and all data)



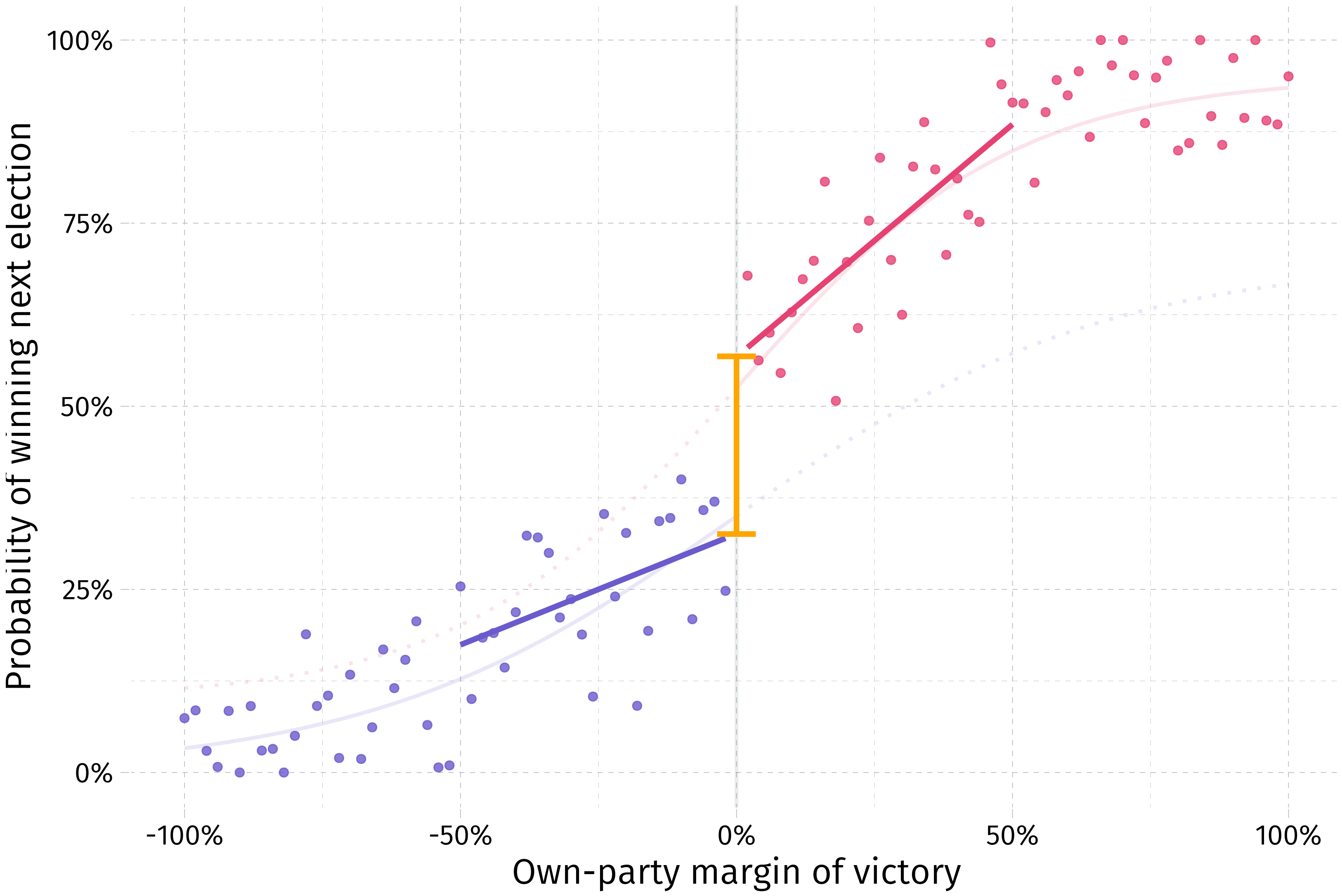
Option 1a Linear regression with constant slopes (and all data)


Option 1b Linear regression with constant slopes; limited to +/- 50%.
100%
75%
50%
25%
0%
Probability of winning next election
-100%
-50%
0%
50%
100%
Own-party margin of victory

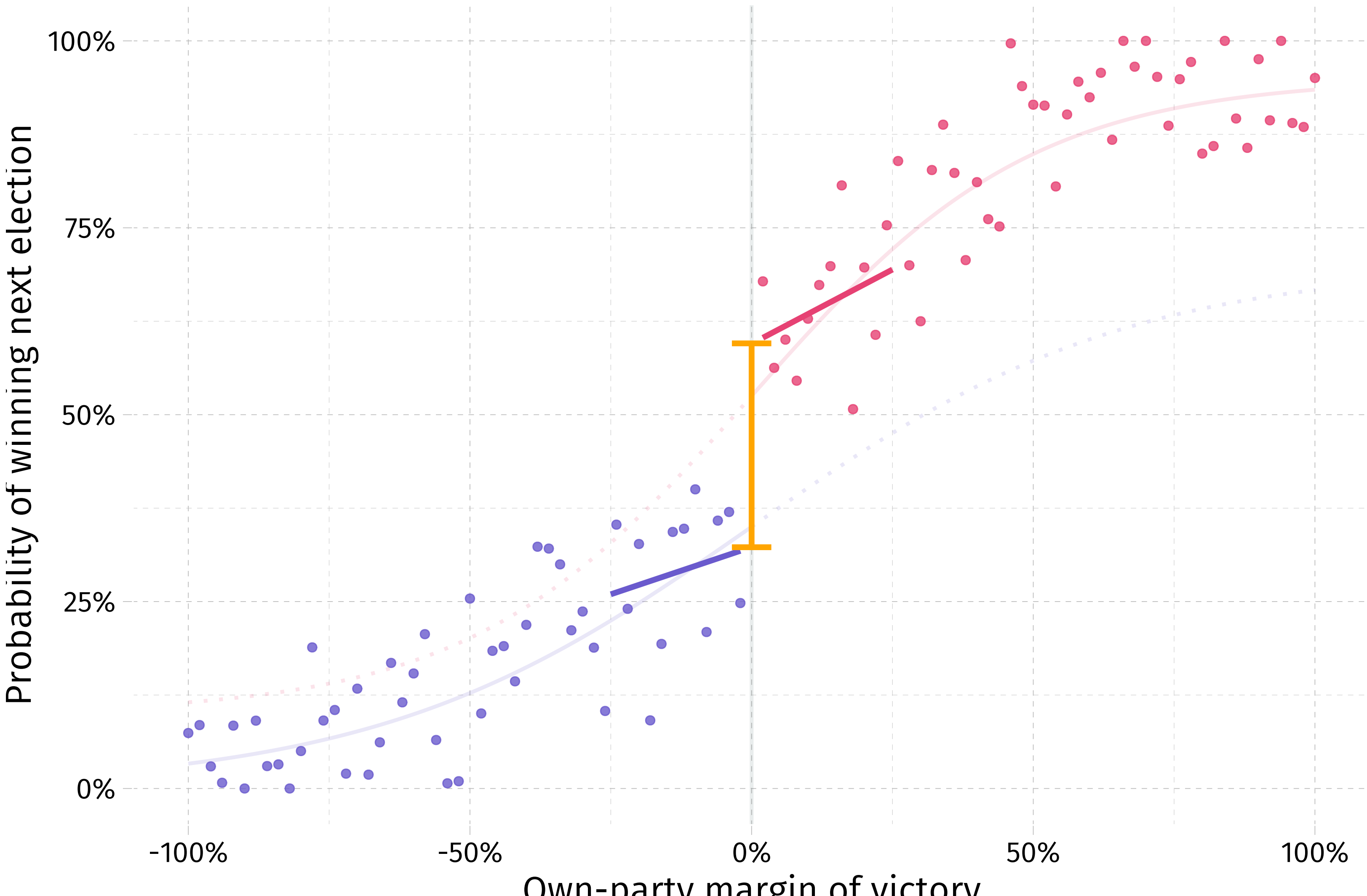
Option 2a Linear regression with differing slopes (and all data)



Option 2b Linear regression with differing slopes; limited to +/- 50%.



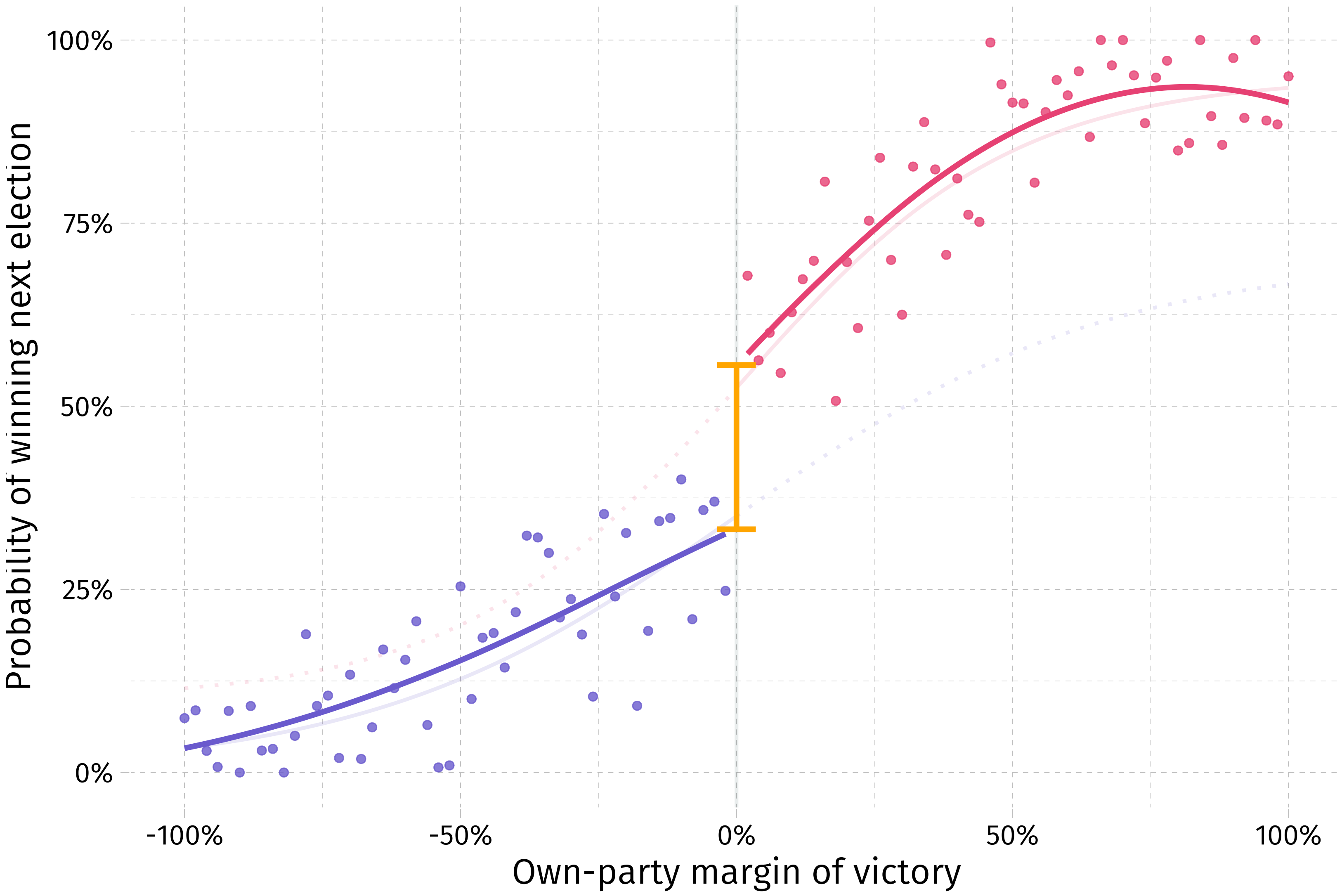
Option 2c Linear regression with differing slopes; limited to +/- 25%.



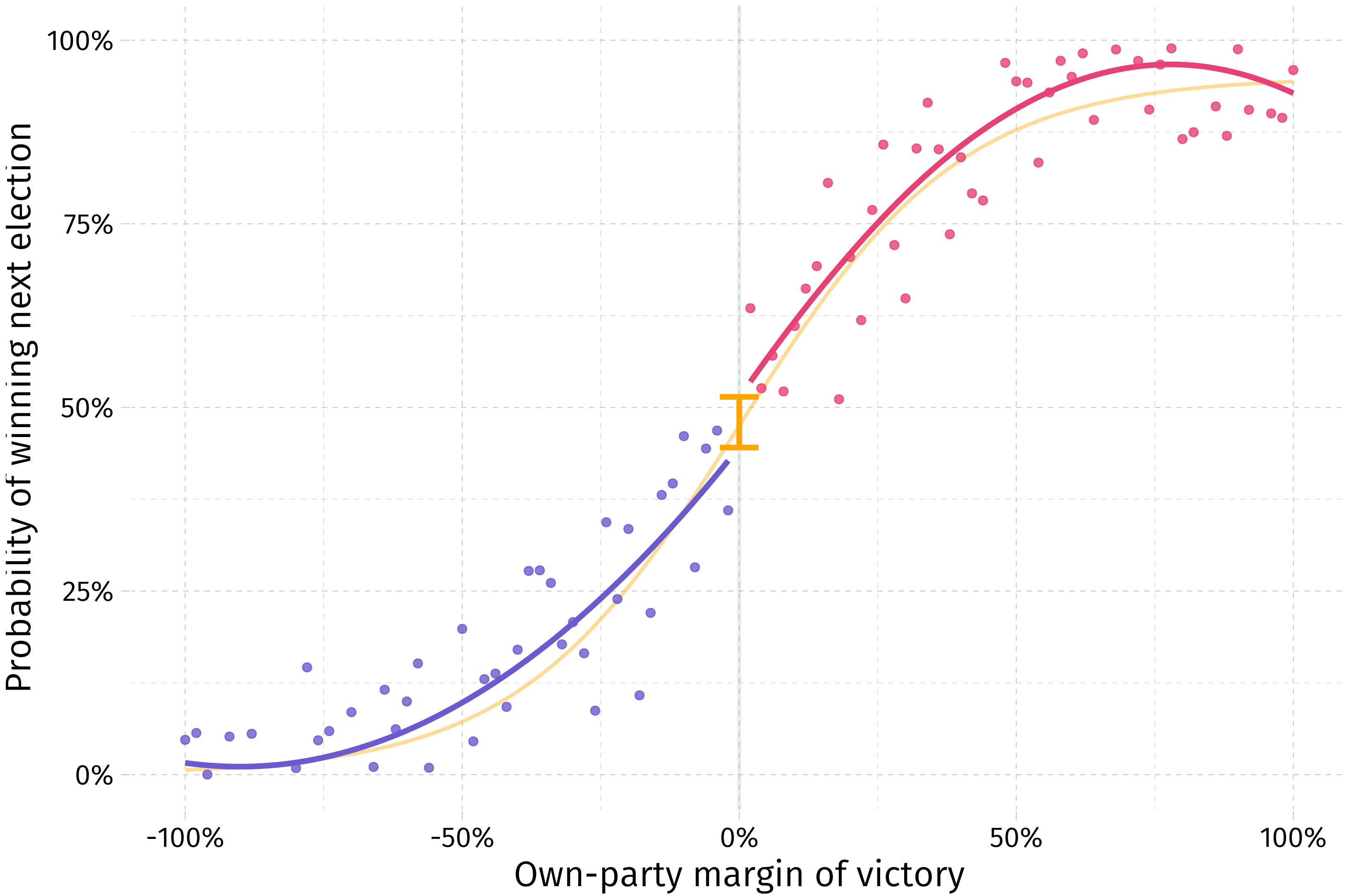
Option 3 Differing quadratic regressions (limited to +/- 50%).

Option 4a Differing local (LOESS) regressions (limited to +/- 50%).

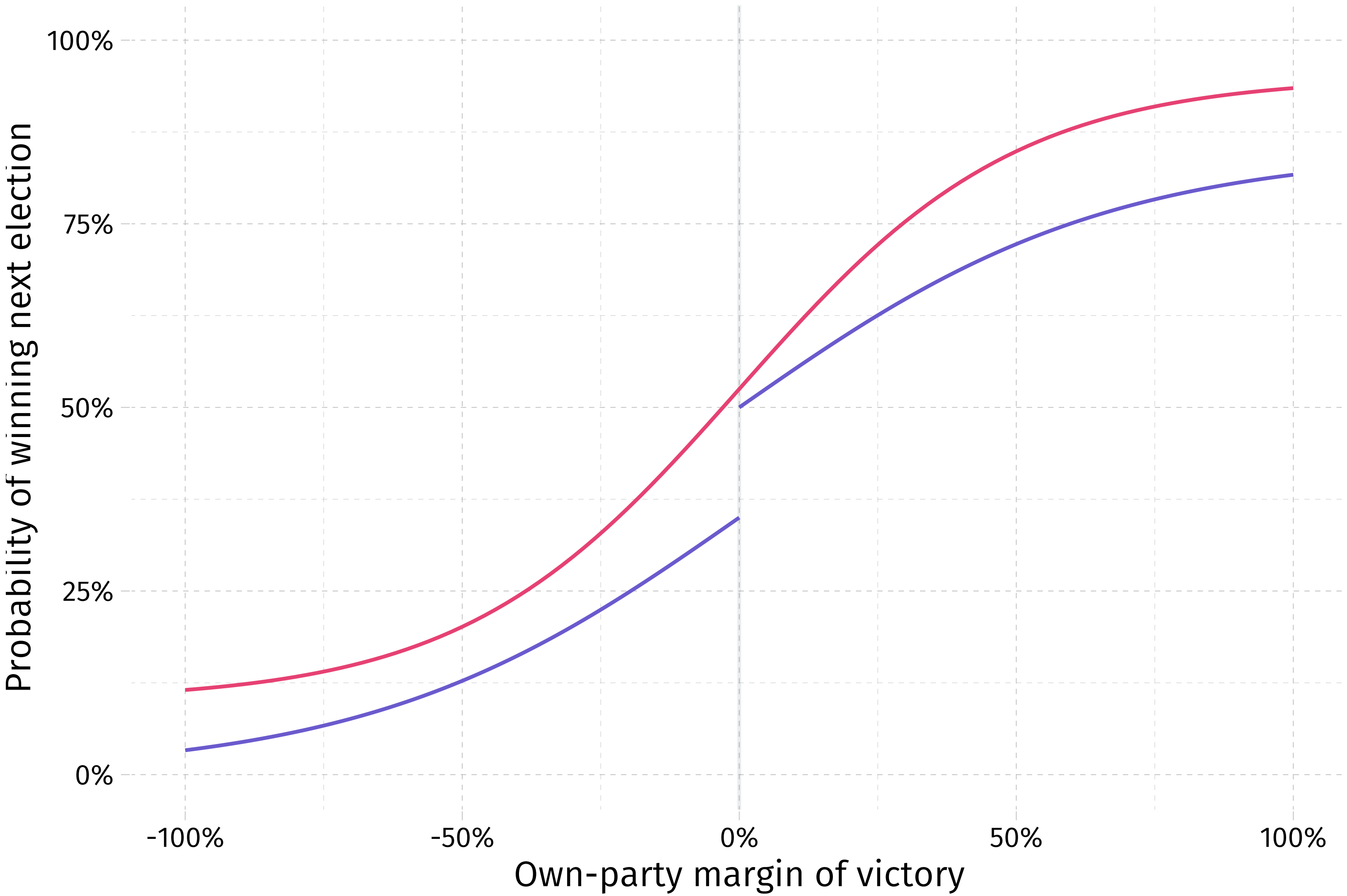
Option 4b Differing local (LOESS) regressions (all data).



Note Functional form can be very important.



The continuity of $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i = x]$ (in x) is also very important. No sorting.



In practice

Gelman and Imbens (2018) on functional form:

> We argue that controlling for global high-order polynomials in regression discontinuity analysis is a flawed approach with three major problems: it leads to noisy estimates, sensitivity to the degree of the polynomial, and poor coverage of confidence intervals. We recommend researchers instead use estimators based on local linear or quadratic polynomials or othersmooth functions.

See Imbens and Kalyanaraman (2012) for optimal bandwidth selection.

Sharp RDs

Estimation

1. **Trim data** to a reasonable window around the threshold c.

2. **Recode** X_i (the "forcing variable") as deviation from c, *i.e.*, $\widetilde{\text{X}}_i = \text{X}_i - c$

   - $\widetilde{\text{X}}_i = 0$ if $\text{X}_i = c$
   - $\widetilde{\text{X}}_i < 0$ if $\text{X}_i < c$ and thus $\text{D}_i = 0$
   - $\widetilde{\text{X}}_i > 0$ if $\text{X}_i > c$ and thus $\text{D}_i = 1$

3. Determine a model to **estimate** $E\left[\text{Y}_i \mid \widetilde{\text{X}}_i\right]$ for $\widetilde{\text{X}}_i$ above and below 0

   - Linear with common slopes for $E\left[\text{Y}_i \mid \widetilde{\text{X}}_i < 0\right]$ and $E\left[\text{Y}_i \mid \widetilde{\text{X}}_i > 0\right]$
   - Linear/quadratic/polynomial with differing slopes
   - LOESS, kernel regression, *etc.*

Estimation: Linear, common slope

Assumptions

1. $E[\mathrm{Y}_{0i}|\mathrm{X}_i = x]$ is linear in x, *i.e.*, $E[\mathrm{Y}_{0i} \mid \mathrm{X}_i] = \alpha + \beta \mathrm{X}_i$
2. Treatment effect does not depend upon τ, *i.e.*, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \mid \mathrm{X}_i] = \tau$

where (1) comes from linearity and (2) comes from common slopes.

$$\implies E[\mathrm{Y}_{1i} \mid \mathrm{X}_i] = \tau + E[\mathrm{Y}_{0i} \mid \mathrm{X}_i] = \tau + \alpha + \beta \mathrm{X}_i$$

Recall our definition of $\mathrm{Y}_i = \mathrm{D}_i \mathrm{Y}_{1i} + (1 - \mathrm{D}_i) \mathrm{Y}_{0i}$.

$$E[\mathrm{Y}_i \mid \mathrm{X}_i,\, \mathrm{D}_i] = \mathrm{D}_i\, E[\mathrm{Y}_{1i} \mid \mathrm{X}_i] + (1 - \mathrm{D}_i)\, E[\mathrm{Y}_{0i} \mid \mathrm{X}_i]$$
$$= \alpha + \tau \mathrm{D}_i + \beta \mathrm{X}_i = \alpha + \tau \mathrm{D}_i + \beta \left(\widetilde{\mathrm{X}}_i + c \right) = \widetilde{\alpha} + \tau \mathrm{D}_i + \beta \widetilde{\mathrm{X}}_i$$

which we can estimate by regressing Y_i on D_i and $\widetilde{\mathrm{X}}_i$.

Estimation: Linear, differing slopes

Assumption $E[\mathrm{Y}_{0i}|\mathrm{X}_i = x]$ and $E[\mathrm{Y}_{1i}|\mathrm{X}_i = x]$ are linear in x, *i.e.*,
$E[\mathrm{Y}_{0i} \mid \mathrm{X}_i] = \alpha_0 + \beta_0 \mathrm{X}_i$ and $E[\mathrm{Y}_{1i} \mid \mathrm{X}_i] = \alpha_1 + \beta_1 \mathrm{X}_i$

Now treatment effects can vary with X_i.

$$\implies E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} \mid \mathrm{X}_i] = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\,\mathrm{X}_i$$

$$\begin{aligned} E[\mathrm{Y}_i \mid \mathrm{X}_i,\, \mathrm{D}_i] &= \mathrm{D}_i\, E[\mathrm{Y}_{1i} \mid \mathrm{X}_i] + (1 - \mathrm{D}_i)\, E[\mathrm{Y}_{0i} \mid \mathrm{X}_i] \\ &= \alpha_0 + \beta_0 \mathrm{X}_i + (\alpha_1 - \alpha_0)\mathrm{D}_i + (\beta_1 - \beta_0)\mathrm{D}_i\mathrm{X}_i \\ &= \widetilde{\alpha} + \beta_0 \widetilde{\mathrm{X}}_i + \tau \mathrm{D}_i + \tilde{\beta}\mathrm{D}_i\widetilde{\mathrm{X}}_i \end{aligned}$$

τ is the LATE at $\widetilde{\mathrm{X}}_i = \mathbf{0}$ ($\mathrm{X}_i = c$). Estimate: Regress $\mathbf{Y}_i$ in $\widetilde{\mathbf{X}}_i$, $\mathbf{D}_i$, *and* $\mathbf{D}_i\widetilde{\mathbf{X}}_i$.[†]

† See Appendix for omitted steps.

Sharp RDs

Estimation: Additional

Fuzzy RDs

Fuzzy RDs

Setup

As with their sharp-RD relatives, **fuzzy RDs** take advantage of a discontinuous change in treatment assignment across some threshold c.

In a **fuzzy regression discontinuity**, the *probability* of treatment changes discontinuously as $\mathbf{X}_i$ crosses c, but it is no longer deterministic.

Formally,

$$0 < \lim_{x \downarrow c} \Pr(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x) - \lim_{x \uparrow c} \Pr(\mathrm{D}_i = 1 \mid \mathrm{X}_i = x) < 1$$

Ex., Exceeding a minimum GRE requirement for graduate school.

Fuzzy RDs

Threshold effects

We now have **two effects** of X_i crossing our threshold c.

1. The effect of X_i crossing c on our outcome

$$\lim_{x \downarrow c} E[\text{Y}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{Y}_i \mid \text{X}_i = x]$$

2. The effect of X_i crossing c on the probability of treatment

$$\lim_{x \downarrow c} E[\text{D}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{D}_i \mid \text{X}_i = x]$$

The treatment effect defined by a fuzzy RD is the ratio of (1) to (2)

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} E[\text{Y}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{Y}_i \mid \text{X}_i = x]}{\lim_{x \downarrow c} E[\text{D}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{D}_i \mid \text{X}_i = x]}$$

Fuzzy RDs

An old friend

This definition of the fuzzy-RD treatment effect

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} E[\text{Y}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{Y}_i \mid \text{X}_i = x]}{\lim_{x \downarrow c} E[\text{D}_i \mid \text{X}_i = x] - \lim_{x \uparrow c} E[\text{D}_i \mid \text{X}_i = x]}$$

should remind you of something—**IV**, where $\mathbf{Z}_i = \mathbb{I}\left\{\mathbf{X}_i \geq c\right\}$.

Accordingly, fuzzy RDs are going to have the **same requirements and interpretation as IV**.

More formally

Let $\mathbf{D}_i\left(x^*\right)$ denote the **potential treatment status** of i **with threshold** x^*.

Why write potential treatment status $\mathbf{D}_i$ a function of the threshold?

Changing the threshold (*e.g.*, voting age) generally makes more sense than changing $\mathbf{X}_i$ (*e.g.*, age).†

I.e., changing the threshold changes treatment statuses at the marginal.

Assumption $\mathbf{D}_i\left(x^*\right)$ is non-increasing in x^* at $x^* = c$.

This is our monotonicity assumption for fuzzy RDs. If we raise x^* from c to $c + \epsilon$, no one joins treatment—no defiers.

† This observation/motivation can help with inference.

Fuzzy RDs

Compliance

Our compliers in this setting are individuals such that

$$\lim_{x^* \downarrow \mathrm{X}_i} \mathrm{D}_i\left(x^*\right) = 0 \qquad \lim_{x^* \uparrow \mathrm{X}_i} \mathrm{D}_i\left(x^*\right) = 1$$

i.e., compliers are only treated when x^* (the threshold) is *below* their X_i.

Back to the fuzzy RD treatment effect

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x] - \lim_{x \uparrow c} E[\mathrm{Y}_i \mid \mathrm{X}_i = x]}{\lim_{x \downarrow c} E[\mathrm{D}_i \mid \mathrm{X}_i = x] - \lim_{x \uparrow c} E[\mathrm{D}_i \mid \mathrm{X}_i = x]}$$

$$= E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | i \text{ is a complier and } \mathrm{X}_i = c].$$

Thus, τ_{FRD} can be a *very local* LATE.

Graphical analysis

Graphical analysis

General

RD analyses hinge on their graphical analyses.

If the discontinuity is not graphically apparent, most people are not going to care about the results of a few tortured regressions.

You're arguing you know that treatment assignment changes across the threshold. If your reader/viewer cannot see it, they're likely not going to believe your regression tables.[†]

† This skepticism may be well founded. We know RDs are sensitive to functional form—and researchers have been known to *p-hack*.

Three figures

Most RD analyses will have some subset of three types of figures.

1. **Outcomes** by the running/forcing variable] $(\mathbf{X}_i)$
   Do we observe a treatment effect across the discontinuity?

2. **Covariates** by the running/forcing variable $(\mathbf{X}_i)$
   Are covariates smooth/balanced across the discontinuity?

3. **Density** of running/forcing variable $(\mathbf{X}_i)$
   Is there evidence of sorty into treatment (across the threshold)?

Outcomes by running variable

These figures tend to show the average value of the outcome $\mathbf{Y}_i$ at evenly spaced bins of the running variable $\mathbf{X}_i$.

We have two parameter choices

1. Binwidth (h)
2. Numbers of bins below and above threshold $(K_0,\, K_1)$

that yield $K = K_0 + K_1$ bins $(k = 1,\, \ldots,\, K)$

$$b_k = c - (K_0 - k + 1) \times h$$

We then calculate summaries for each bin.

Graphical analysis

Outcomes by running variable

The bin's **number of observations**, N_k

$$N_k = \sum_{i=1}^{N} \mathbb{I}\left\{b_k < \mathrm{X}_i \leq b_{k+1}\right\}$$

The **average treatment level** in the bin, $\overline{D}_k$ (for fuzzy RDs)

$$\overline{D}_k = \frac{1}{N_k} \sum_{i=1}^{N} \mathrm{D}_i \times \mathbb{I}\left\{b_k < \mathrm{X}_i \leq b_{k+1}\right\}$$

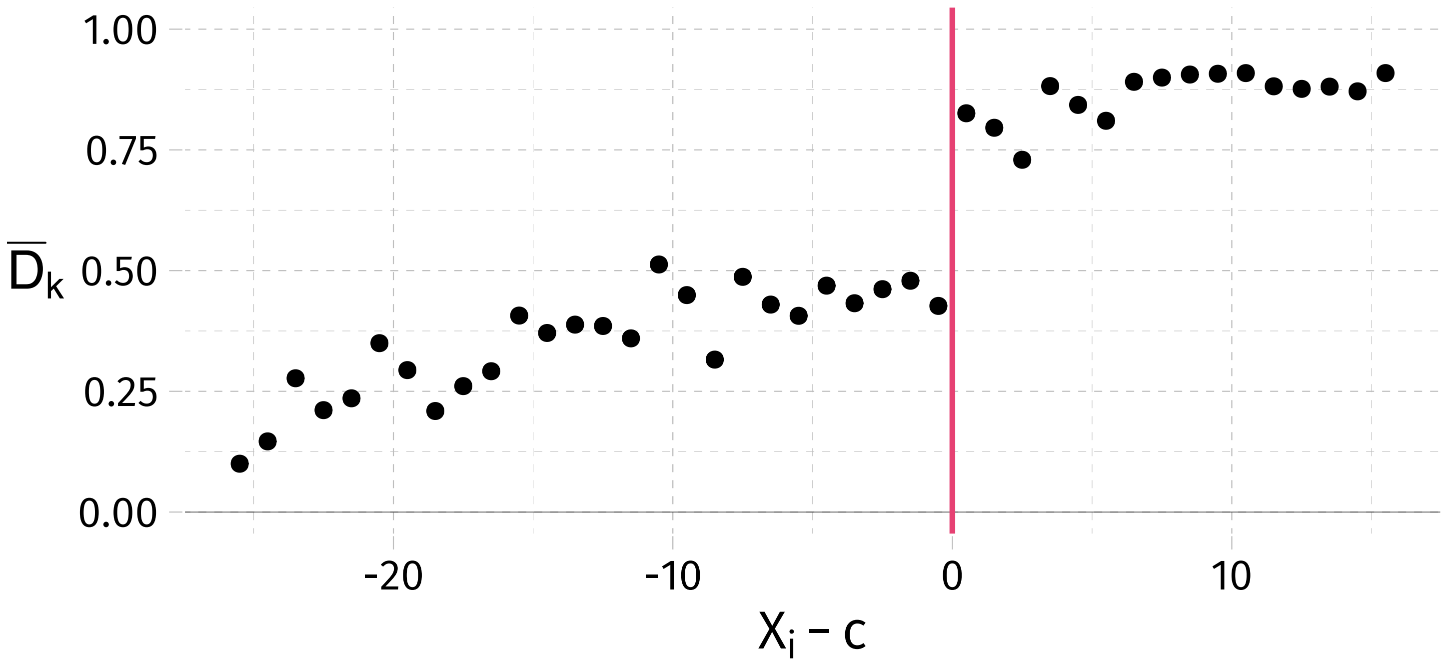
The **average outcome** in the bin, $\overline{Y}_k$

$$\overline{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} \mathrm{Y}_i \times \mathbb{I}\left\{b_k < \mathrm{X}_i \leq b_{k+1}\right\}$$

Outcomes by running variable

We then plot $\overline{\mathrm{D}}_k$ against the midpoint of each bin.

Q Does crossing c clearly affect $\mathbf{Pr}(\mathrm{D}_i = 1)$? (Fuzzy RD first stage)

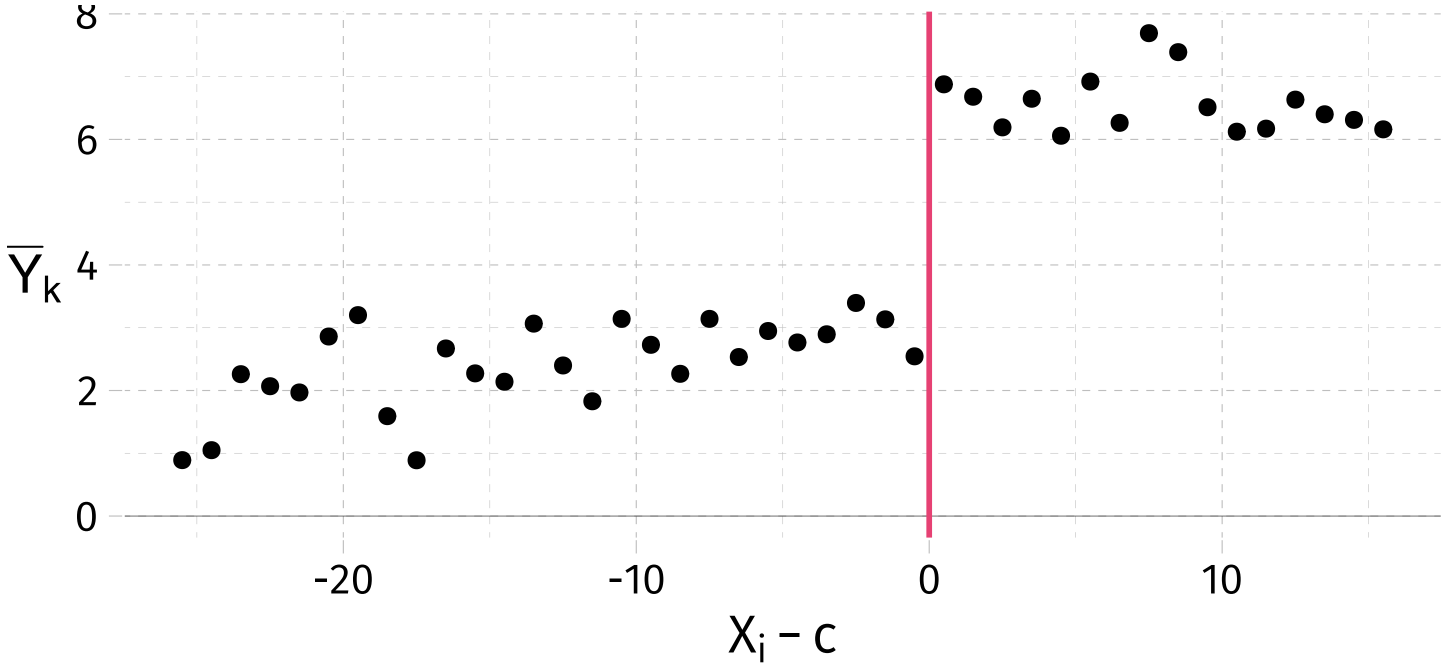


Graphical analysis

Outcomes by running variable

And then plot $\overline{\text{Y}}_k$ against the midpoint of each bin.

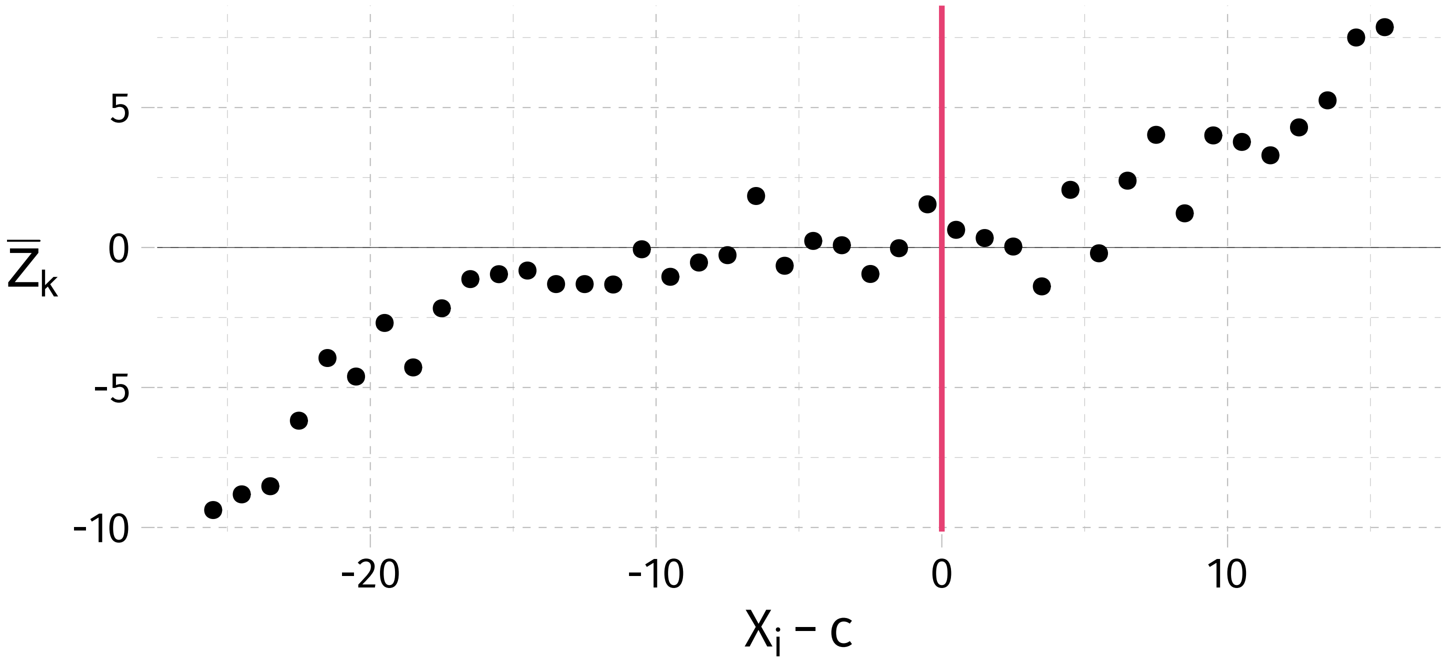
Q Does crossing c clearly affect our outcome Y_i? (Fuzzy RD reduced form)



Covariates by running variable

Now we apply the same approach to covariates (Z_i).

Q Are covariates **smooth** across c? If not, your RD may be invalid.



Graphical analysis

Density of running variable

Finally we looking for other violations of smoothness—particularly in form gaming the threshold.

In other words: Are individuals **bunching** just above or just below the threshold?

If so, folks just below the threshold don't give us the clean counterfactual that we want for the folks just above the threshold.

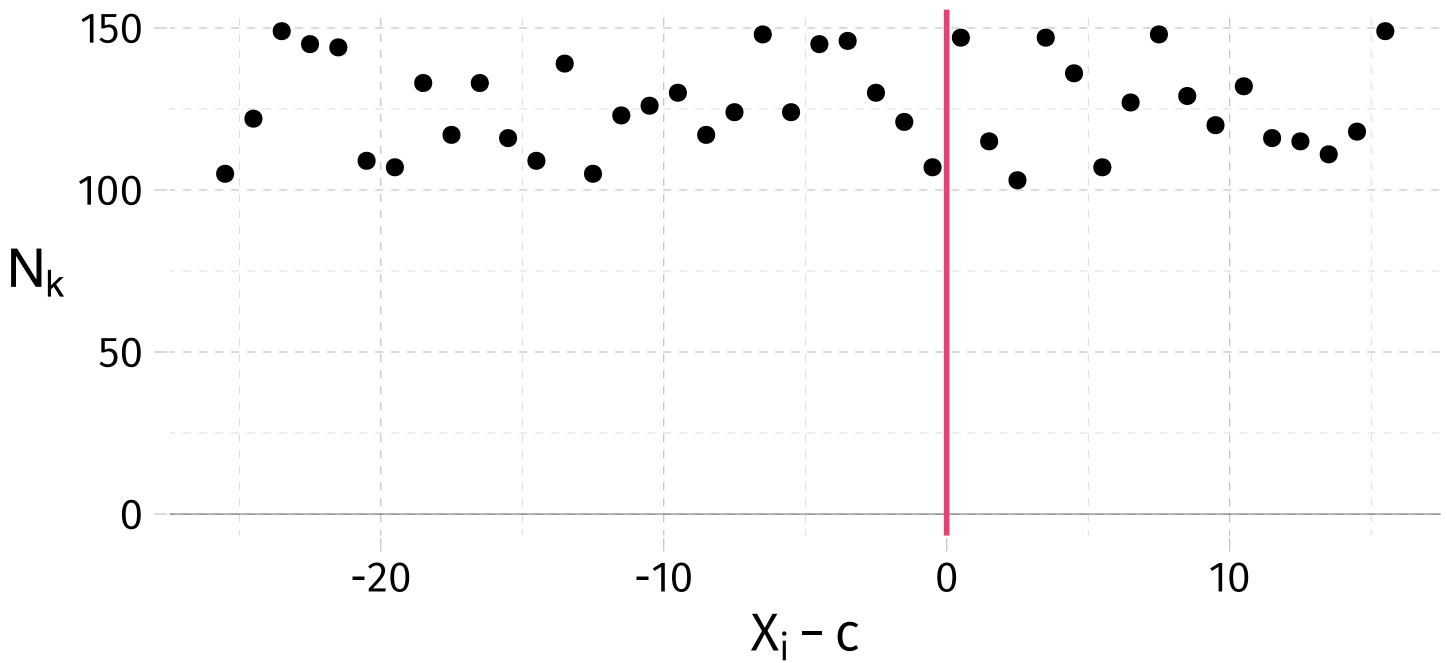
McCrary (2008) suggests testing the density of $\mathbf{X}_i$ at c.

Graphical analysis

Density of running variable

Effectively, we can plot N_k at the midpoint of each bin.

Q Is the distribution of X_i smooth across c?

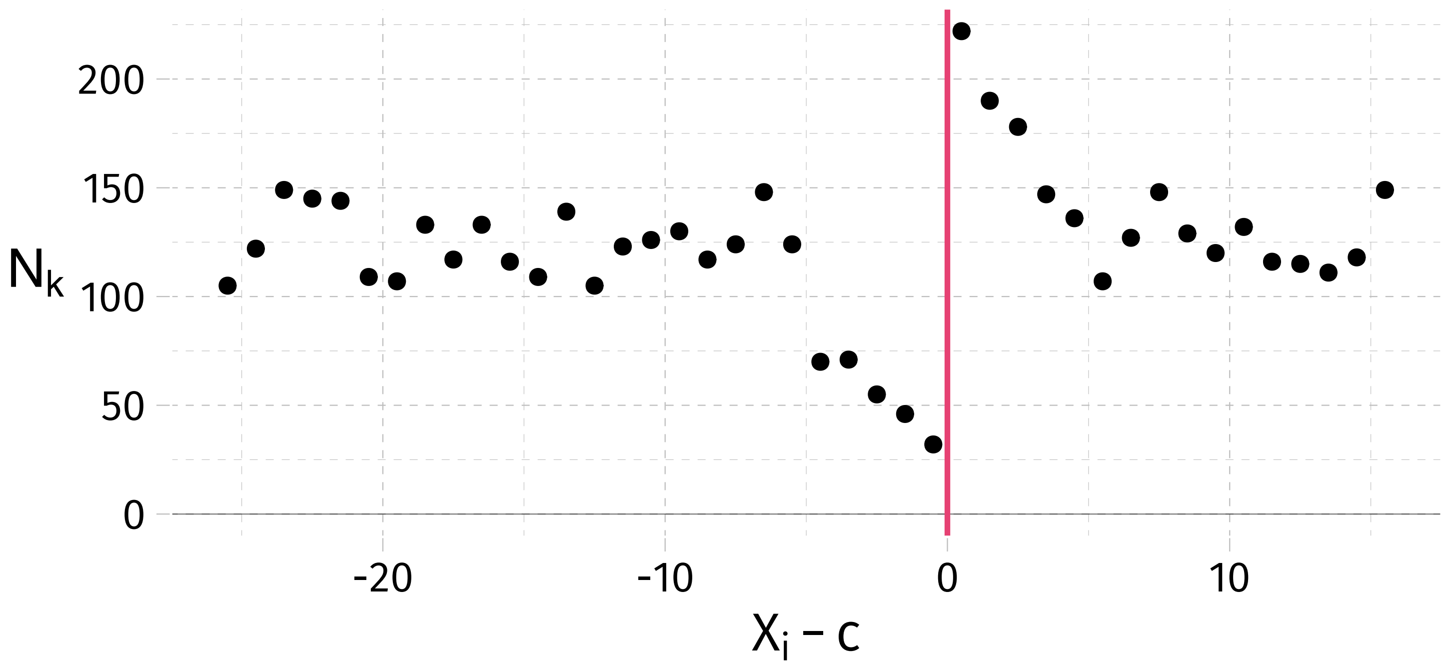


Graphical analysis

Density of running variable

Likely bunching (problem)

Q Is the distribution of X_i smooth across c?



Graphical analysis

Additional points

1. No bin should cross the threshold.
2. Are there discontinuities other than c? Should there be? Smoothness?

Again, if these graphs are not clear and convincing, it's going to be hard to make the case that you have a true/credible discontinuity.

Estimation: Linear, differing slopes

Definitions of terms that magically appear

- $\widetilde{\alpha} = \alpha_0 + \beta_0 \boldsymbol{c}$
- $\boldsymbol{\tau} = (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)\, \boldsymbol{c}$
- $\tilde{\beta} = (\beta_1 - \beta_0)$

Table of contents

Admin

1. Schedule

General RD

1. Setup
2. Framework
3. Examples
4. Sharp *vs.* fuzzy

Graphical analysis

1. General
2. Outcomes by $\mathbf{X}_i$
3. Covariates by $\mathbf{X}_i$
4. Density of $\mathbf{X}_i$

Sharp RDs

1. Setup
2. (Semi) Formally
3. Estimation
4. Examples
5. In practice
6. More estimation

Fuzzy RDs

1. Setup
2. As IV
3. Somewhat formally