RDD in practice

PSCI 2301: Quantitative Political Science II

Prof. Brenton Kenkel

brenton.kenkel@gmail.com

Vanderbilt University

March 19, 2025

Recap

Last time: Estimating treatment effects by regression discontinuity

- Another method for observational data w/ unobserved confounders
- Key assumptions
    1. Treatment assignment determined by sharp cutoff in continuous "running variable"
    2. No major dissimilarities in other background characteristics just below/above the cutoff
- Linear relationship between running and outcome: linear RDD
- Nonlinear relationship: polynomial RDD, or local linear RDD w/in bandwidth

Today's agenda

Working with RDD in practice, using Hall's data

1. Initial visual inspections
2. Estimating RDD models
    - Linear RDD via `lm()`
    - Polynomial RDD via `lm()`
    - Automatic bandwidth selection via `rdrobust()`
3. Assessing balance

Effects of ideological extremism in House races

Research design

What is the effect of ideological extremism on a candidate's election results?

Population: US House races, 1980–2010

- Only those with a competitive primary
- ...and a discernible ideological difference b/w primary candidates

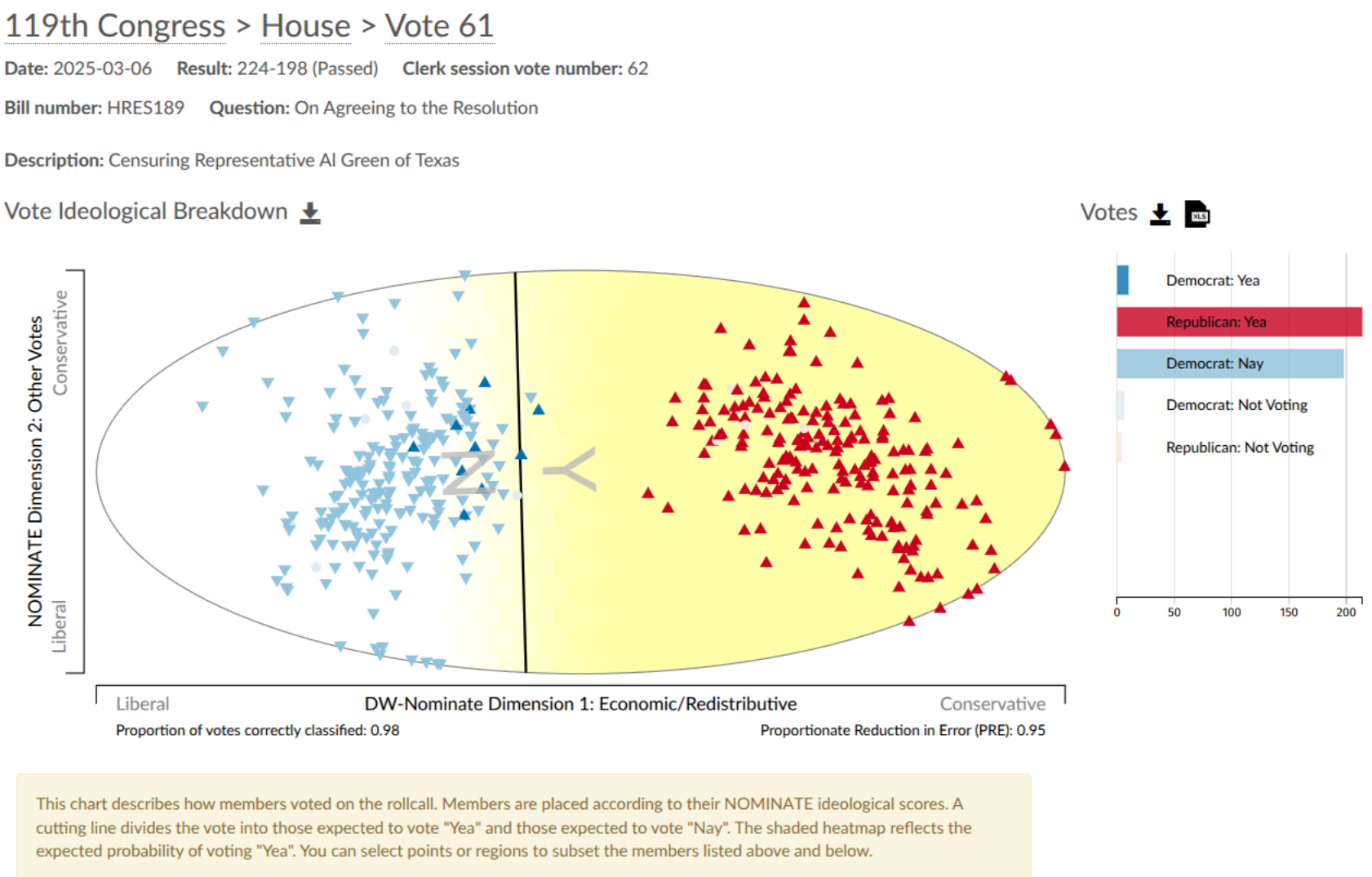
Outcome: Vote share in general election

Treatment group: Ideological extremists

Comparison group: Ideological moderates

Measuring primary candidate ideology

Typically measure legislator ideology by how they voted on bills



Measuring primary candidate ideology

Problem: Don't observe bill votes for candidates who weren't elected

Hall's solution: Measure ideology by donation patterns

- Lots of data: All donations of $200+ must be publicly reported
- Key assumption: Donors favor ideologically close candidates

Rough outline of how this works:

1. Measure ideology of incumbents the usual way
2. Measure donor ideology via which incumbents they contribute to
3. Measure candidate ideology as weighted average of donor ideology

Accessing Hall's data

Can download manually from https://www.andrewbenjaminhall.com/

...or can copy-paste this:

```
library("archive")
library("haven")
url_hall <- "https://www.dropbox.com/s/1o0lrqemdlyh7ha/Hall_Extremist_Primaries_APSR_Replication.zip?dl=
con <- archive_read(url_hall, file = "primary_analysis.dta")
df_hall <- read_dta(con)
```

Hall's data

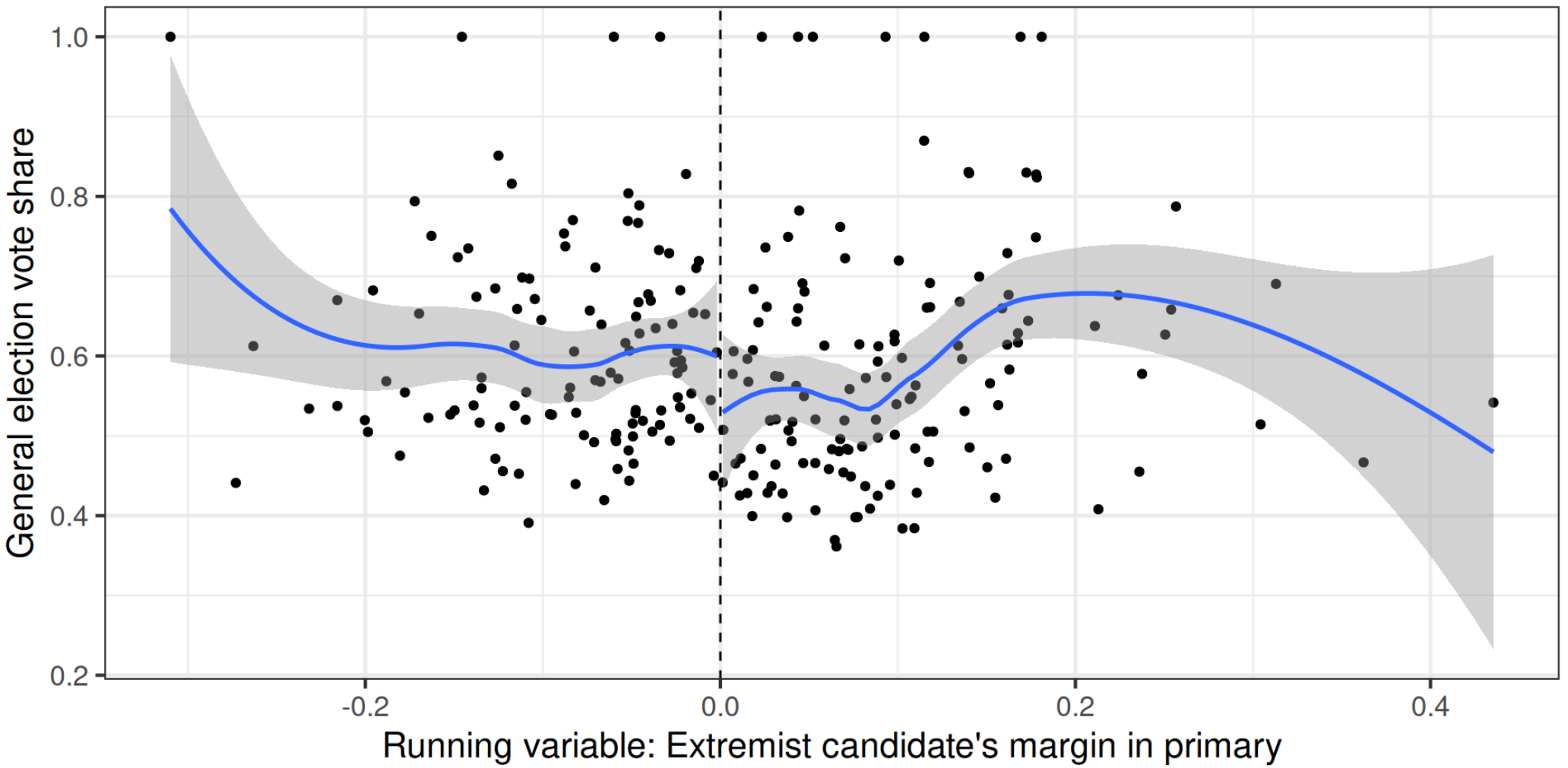
```
# Only use races w/ above-median ideological difference
# (as in Hall's main analysis)
df_hall <- df_hall |>
  filter(absdist > median(absdist)) |>
  relocate(state, dist, year, dem, dv, rv, treat)

df_hall
```

```
# A tibble: 252 × 70
  state  dist  year   dem    dv      rv treat redist1 redist2 vote_P0 cand_dwnom0 prim_total0
  <chr> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>       <dbl>       <dbl>
1 AL        2  1992     1 0.492 -0.0711     0     103     108   0.429      -0.312       70849
2 AL        3  1996     1 0.484  0.110      1     103     108   0.610      -0.172       71650
3 AL        5  1990     1 0.671 -0.105      0      98     103   0.395      -0.166       46017
4 AL        6  1980     0 0.520  0.0535     1      93      98   0.554       0.255        8942
5 AL        7  1982     1 1     -0.146      0      98     103   0.354      -0.317       28650
# i 247 more rows
# i 58 more variables: vote_P1 <dbl>, cand_dwnom1 <dbl>, fully_open_general <dbl>,
#   this_primary_open <dbl>, this_primary_open_other_inc <dbl>, this_primary_inc_other_open <dbl>,
#   both_primaries_inc <dbl>, tot_amountQ <dbl>, tot_amountY <dbl>, absdist <dbl>, rv_treat <dbl>,
#   margin <dbl>, rv2 <dbl>, rv3 <dbl>, rv4 <dbl>, prim_total_winner <dbl>, prim_pac_share <dbl>,
#   prim_share <dbl>, winner_score <dbl>, inc_winner <dbl>, party_share <dbl>, group_share <dbl>,
#   pnv <dbl>, lag_pnv <dbl>, dv_win <dbl>, lag_dv_party <dbl>, vote_G_comb <dbl>, ...
```

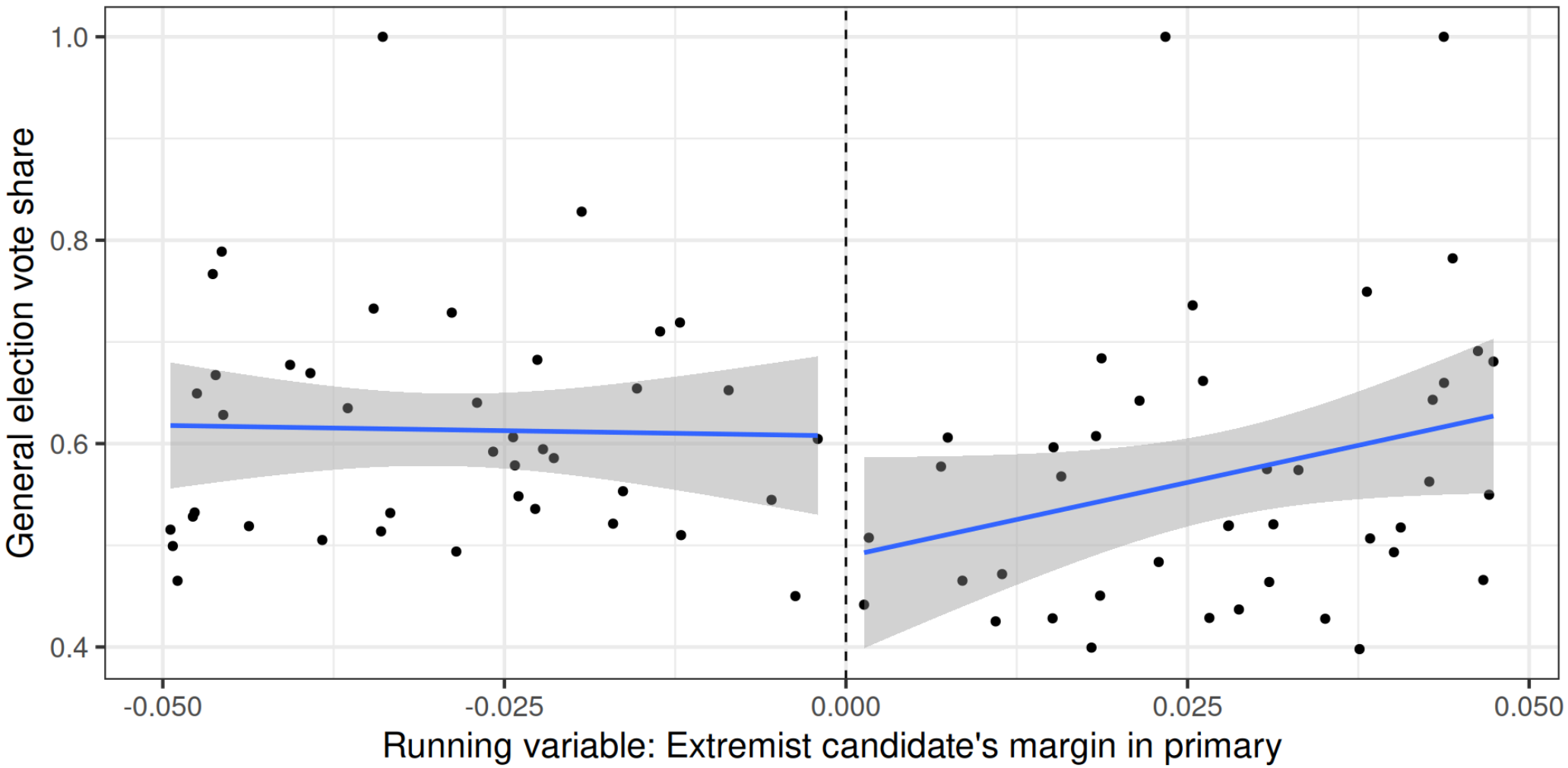
Looking at the data

Plot raw data to assess linearity + discontinuity



A closer look at the data

Restricting to when extremist won/lost by 5% or less



Polynomial RDD

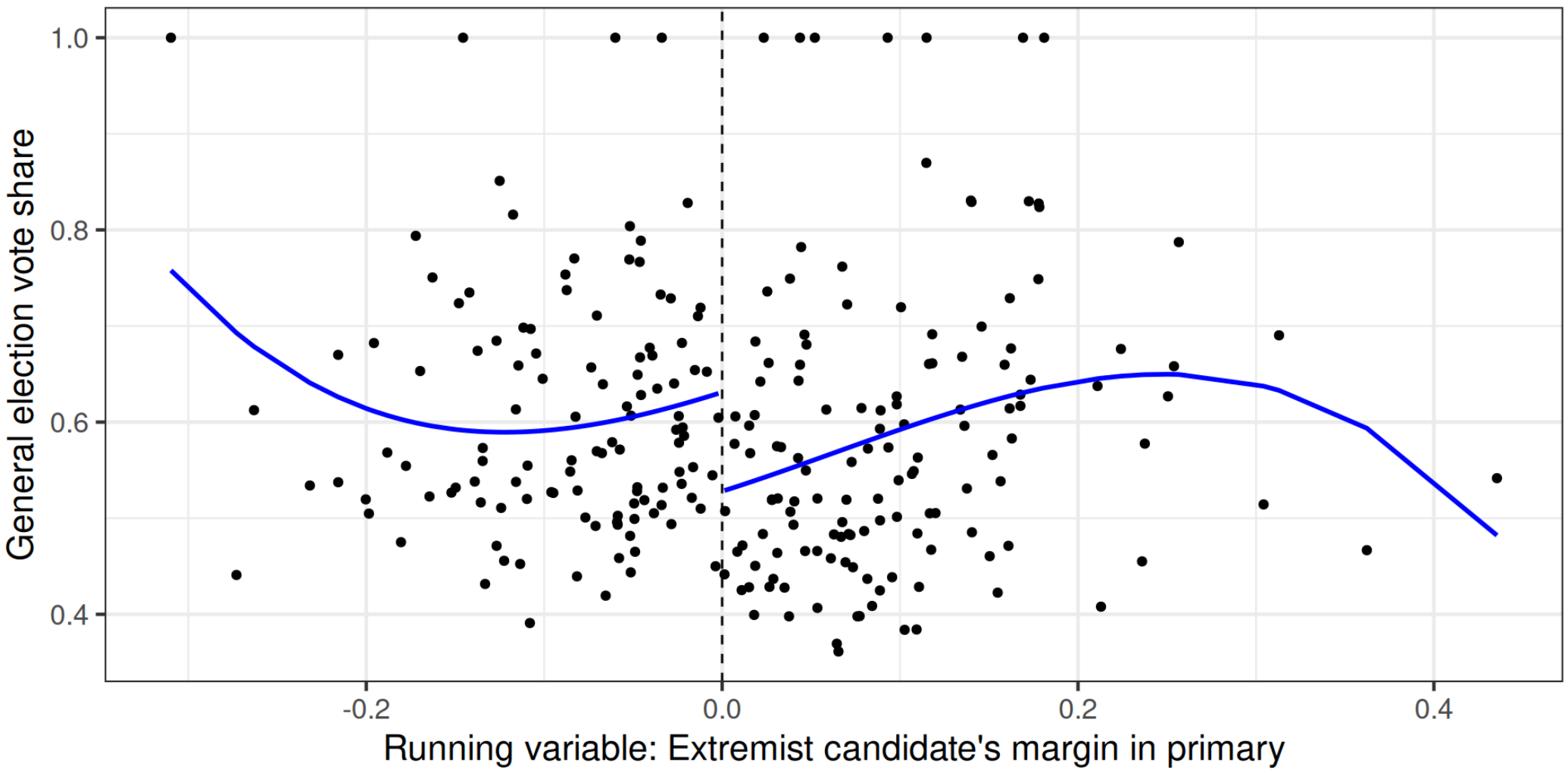
```
fit_cubic <- lm(
  dv ~ treat + rv + I(rv^2) + I(rv^3),
  data = df_hall
)
tidy(fit_cubic)
```

```
# A tibble: 5 × 5
  term        estimate std.error statistic
  <chr>          <dbl>     <dbl>     <dbl>
1 (Intercept)    0.631    0.0198     31.8
2 treat         -0.103    0.0346     -2.98
3 rv             0.585    0.201       2.91
4 I(rv^2)        1.22     0.501       2.43
5 I(rv^3)       -6.44     2.22       -2.90
# i 1 more variable: p.value <dbl>
```

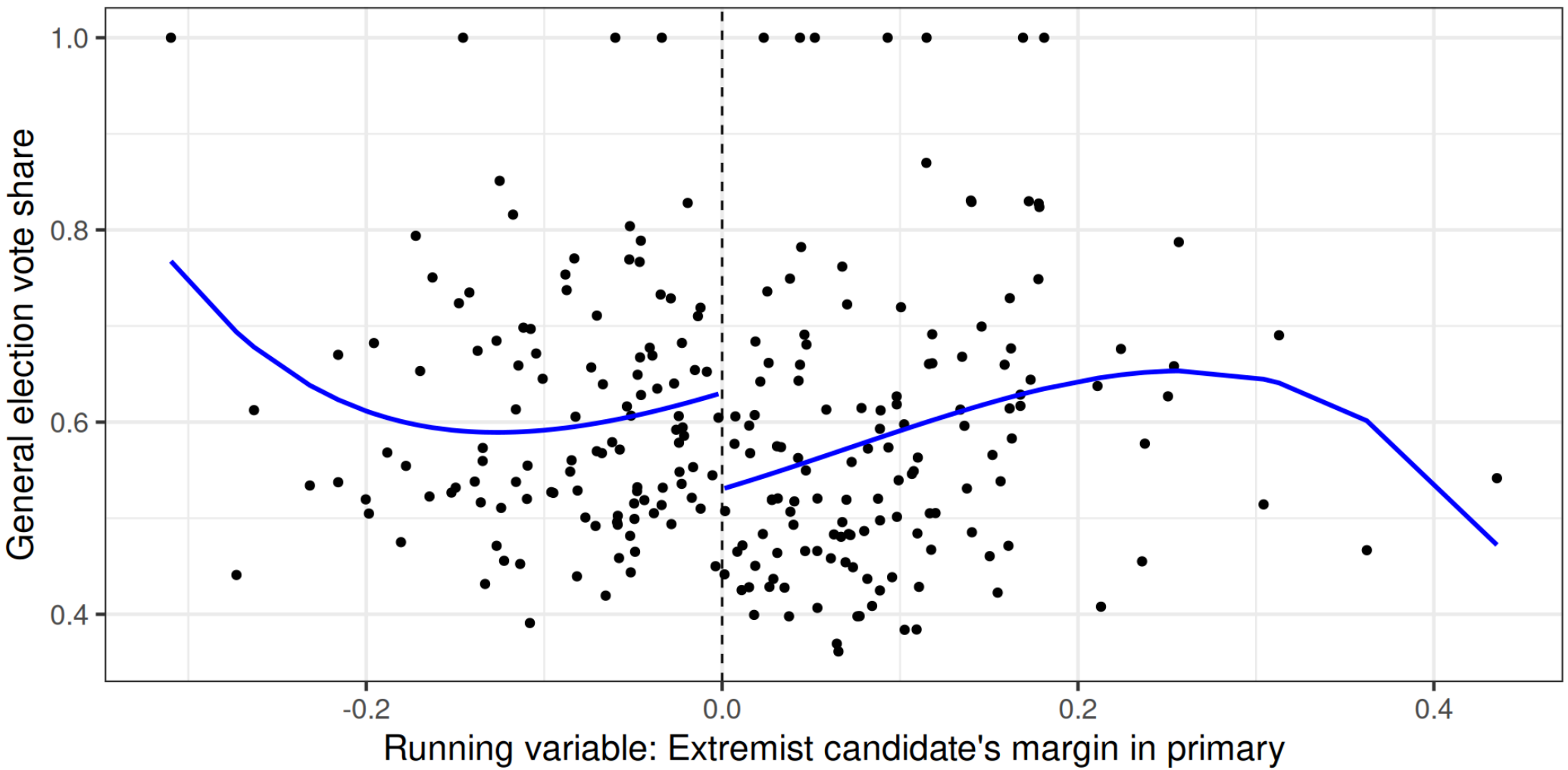
```
fit_quintic <- lm(
  dv ~ treat + rv + I(rv^2) + I(rv^3) +
    I(rv^4) + I(rv^5),
  data = df_hall
)
tidy(fit_quintic)
```

```
# A tibble: 7 × 5
  term        estimate std.error statistic
  <chr>          <dbl>     <dbl>     <dbl>
1 (Intercept)    0.630    0.0233    27.1
2 treat         -0.100    0.0402    -2.49
3 rv             0.550    0.301      1.83
4 I(rv^2)        1.06     1.23       0.865
5 I(rv^3)       -5.18     7.31      -0.709
6 I(rv^4)        2.44    16.4        0.149
7 I(rv^5)      -10.2     52.7       -0.193
# i 1 more variable: p.value <dbl>
```

Checking fit of polynomial RDD: Cubic model



Checking fit of polynomial RDD: Quintic model



Local linear RDD with manual bandwidth choice

```
fit_local_linear <- lm(
  dv ~ treat * rv,
  data = df_hall,
  subset = margin < 0.05
)
tidy(fit_local_linear)
```

```
# A tibble: 4 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.608    0.0454    13.4   5.04e-22
2 treat         -0.119    0.0635    -1.87  6.53e- 2
3 rv            -0.207    1.39      -0.149 8.82e- 1
4 treat:rv       3.12     2.01       1.56  1.24e- 1
```

Considerations in bandwidth choice

We face a **bias-variance** tradeoff in selecting a bandwidth

Narrow bandwidth

- Low bias: Only comparing very similar observations
- High variance: Throwing away lots of data

Wide bandwidth

- High(er) bias: Comparing observations with more baseline diffs
- Low(er) variance: Using greater fraction of the data

Statisticians have developed methods to balance these two considerations

Choosing bandwidth automatically

Using the **rdrobust** package:

```
# install.packages("rdrobust") if not installed already
library("rdrobust")
```

`rdrobust()` function automatically estimates bandwidth + corrects SEs

```
fit_rdr <- rdrobust(y = df_hall$dv, x = df_hall$rv, c = 0)
```

In here we'll just use the defaults, but there are **lots** of options to check before using `rdrobust()` for production/publication work

Choosing bandwidth automatically

Sadly our friend `tidy()` doesn't work for `rdrobust()` output

```
summary(fit_rdr)
```

```
Sharp RD estimates using local polynomial regression.

Number of Obs.                  252
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                  116         136
Eff. Number of Obs.              54          49
Order est. (p)                    1           1
Order bias  (q)                   2           2
BW est. (h)                   0.064       0.064
BW bias (b)                   0.111       0.111
rho (h/b)                     0.577       0.577
Unique Obs.                     116         136
```

Checking balance

Key RDD assumption: Obs essentially similar on either side of threshold

Want to verify there's **no** discontinuity for observed confounders

Confounding variables in Hall's analysis:

- `winner_female`: Was the winning candidate female?
- `inc_winner`: Was the winning candidate an incumbent?
- `qual`: Did the winning candidate have prior political experience?
- `winner_share`: How big was the winning candidate's fundraising (dis)advantage?

Balance check regressions

```r
# Extract optimal bandwidth from dv analysis
bandwidth <- fit_rdr$bws[1]
df_hall_subset <- filter(df_hall, abs(rv) <= bandwidth)

# Local linear RDD for each confounding variable
fit_fem <- lm(winner_female ~ treat * rv, data = df_hall_subset)
fit_inc <- lm(inc_winner ~ treat * rv, data = df_hall_subset)
fit_exp <- lm(qual ~ treat * rv, data = df_hall_subset)
fit_don <- lm(winner_share ~ treat * rv, data = df_hall_subset)
```

Balance check regressions

For once we want **high** p-values

```
tidy(fit_fem) |> filter(term == "treat")
```

```
# A tibble: 1 × 5
  term  estimate std.error statistic p.value
  <chr>    <dbl>     <dbl>     <dbl>   <dbl>
1 treat  -0.0761     0.148    -0.513   0.609
```

```
tidy(fit_exp) |> filter(term == "treat")
```

```
# A tibble: 1 × 5
  term  estimate std.error statistic p.value
  <chr>    <dbl>     <dbl>     <dbl>   <dbl>
1 treat 0.000709     0.224   0.00316   0.997
```

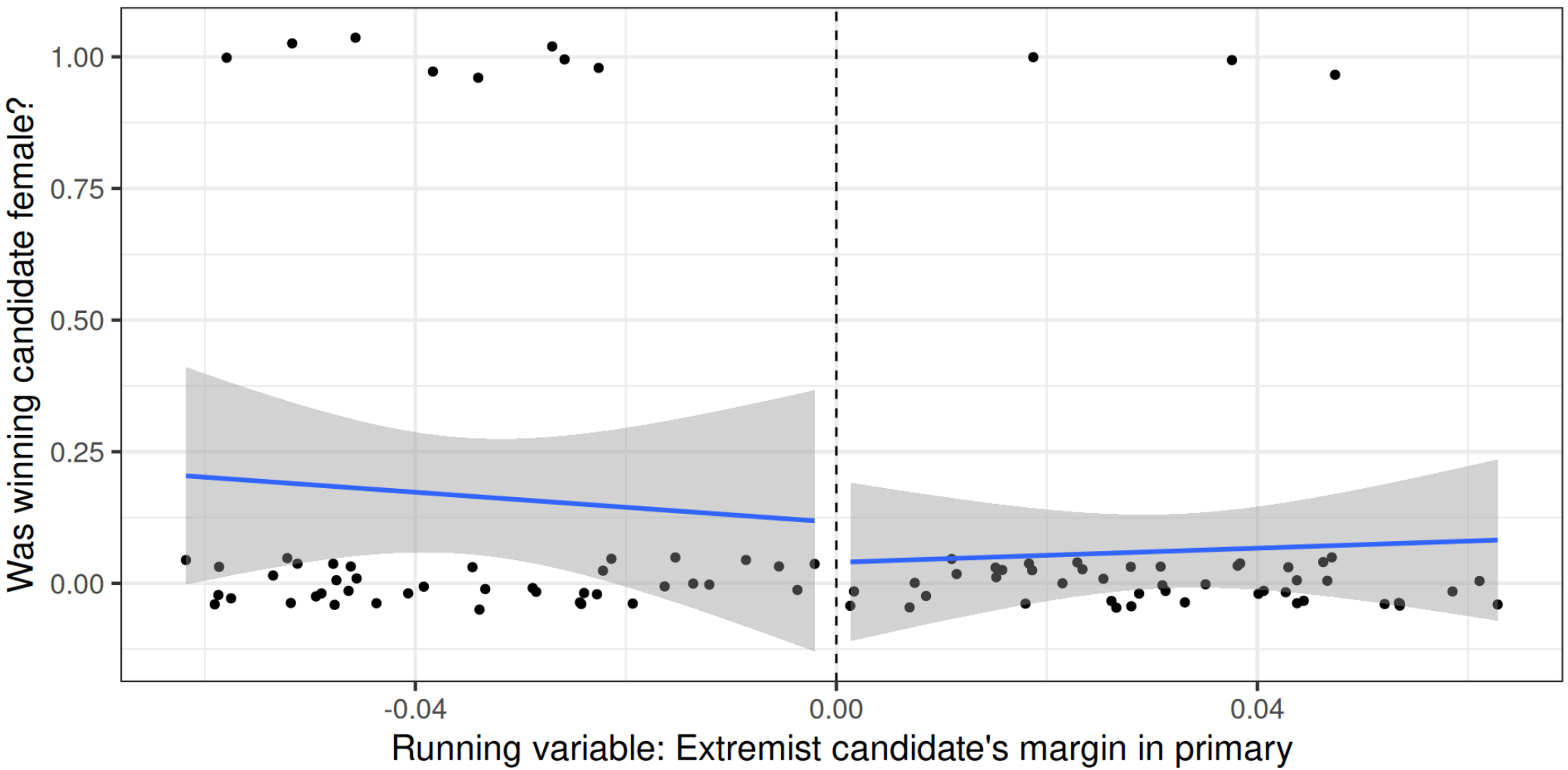
```
tidy(fit_inc) |> filter(term == "treat")
```

```
# A tibble: 1 × 5
  term  estimate std.error statistic p.value
  <chr>    <dbl>     <dbl>     <dbl>   <dbl>
1 treat   0.0397     0.174     0.228   0.820
```

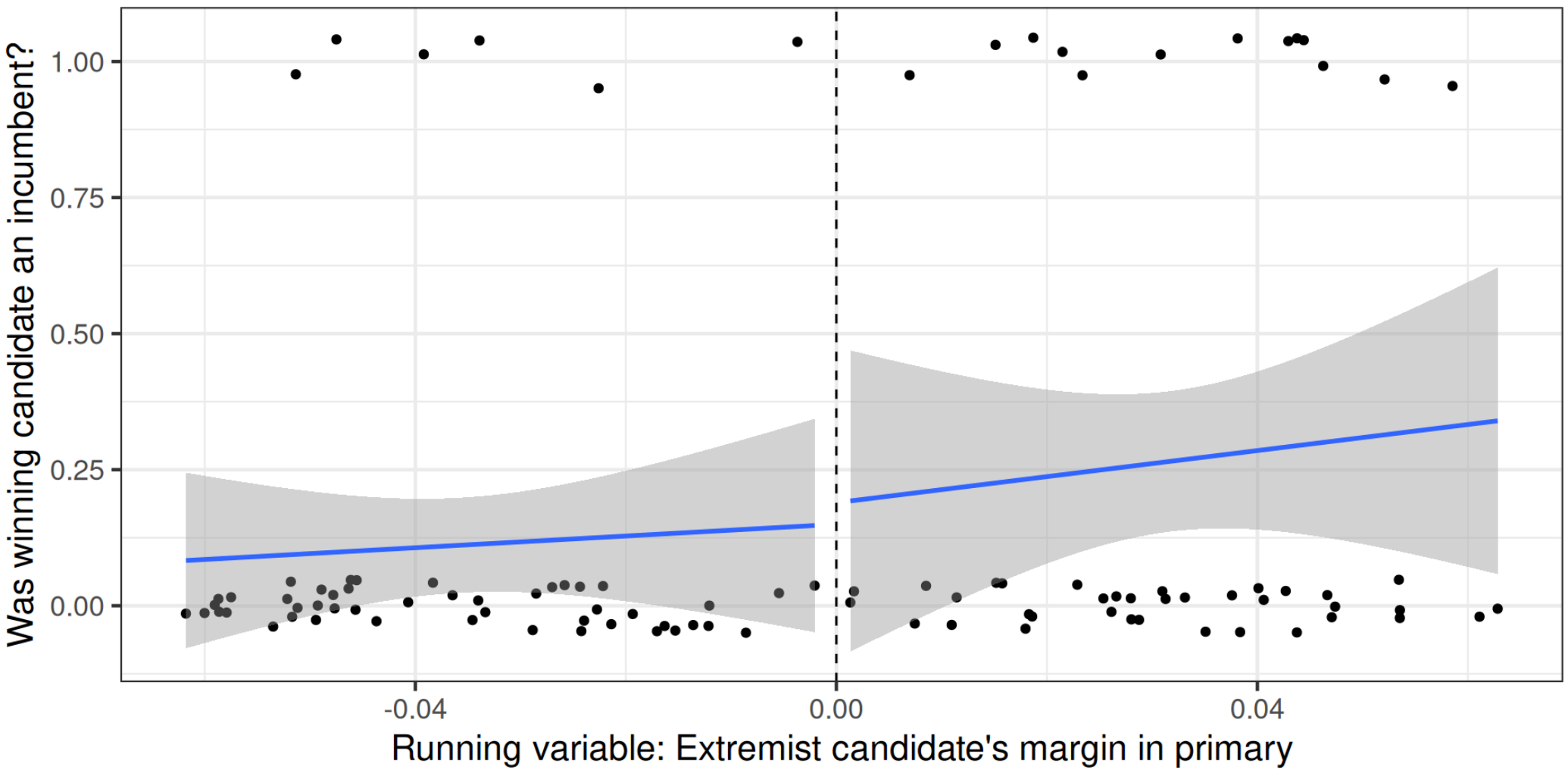
```
tidy(fit_don) |> filter(term == "treat")
```

```
# A tibble: 1 × 5
  term  estimate std.error statistic p.value
  <chr>    <dbl>     <dbl>     <dbl>   <dbl>
1 treat   0.0643     0.138     0.467   0.641
```

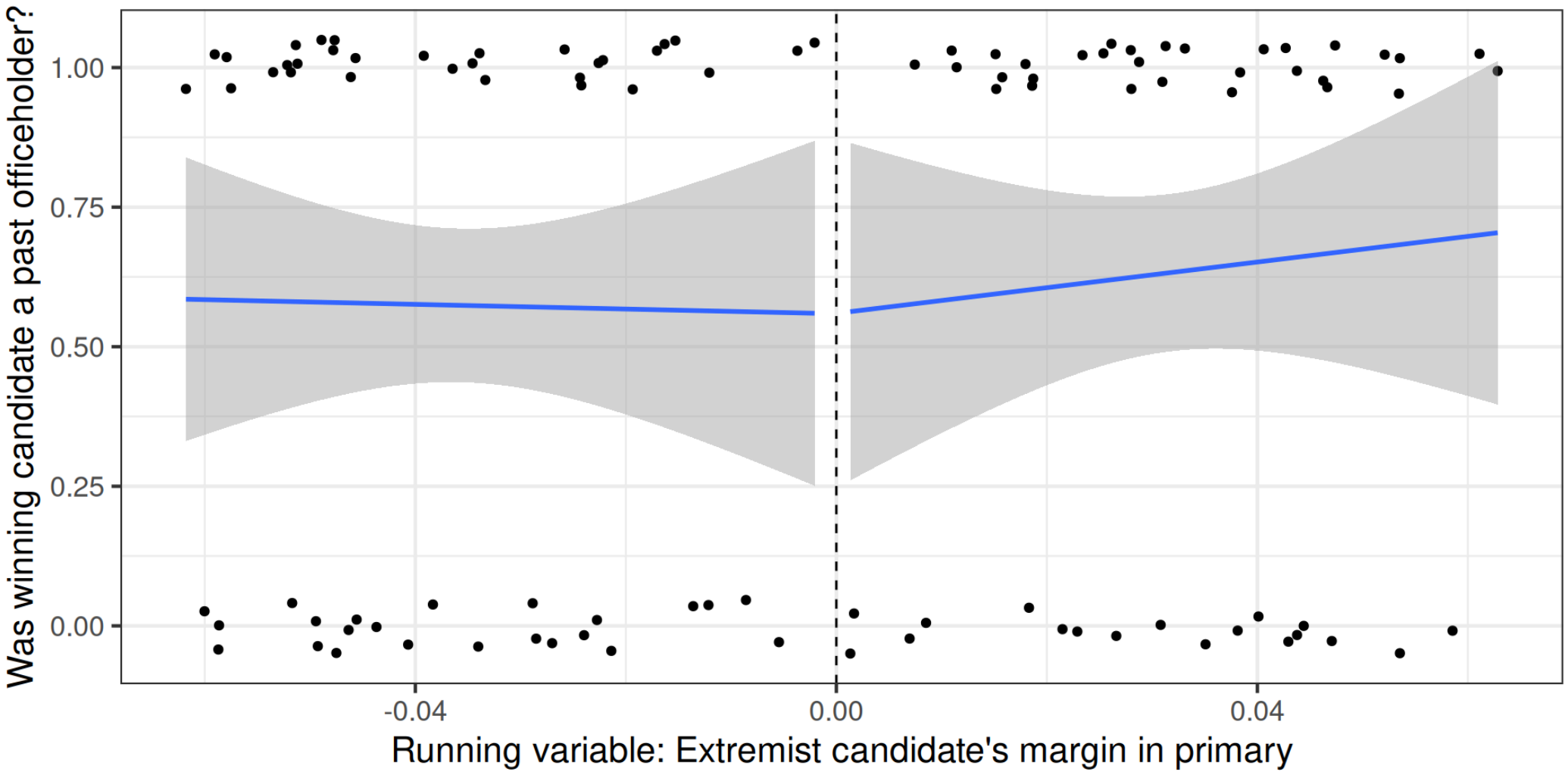
Visual balance check: Gender



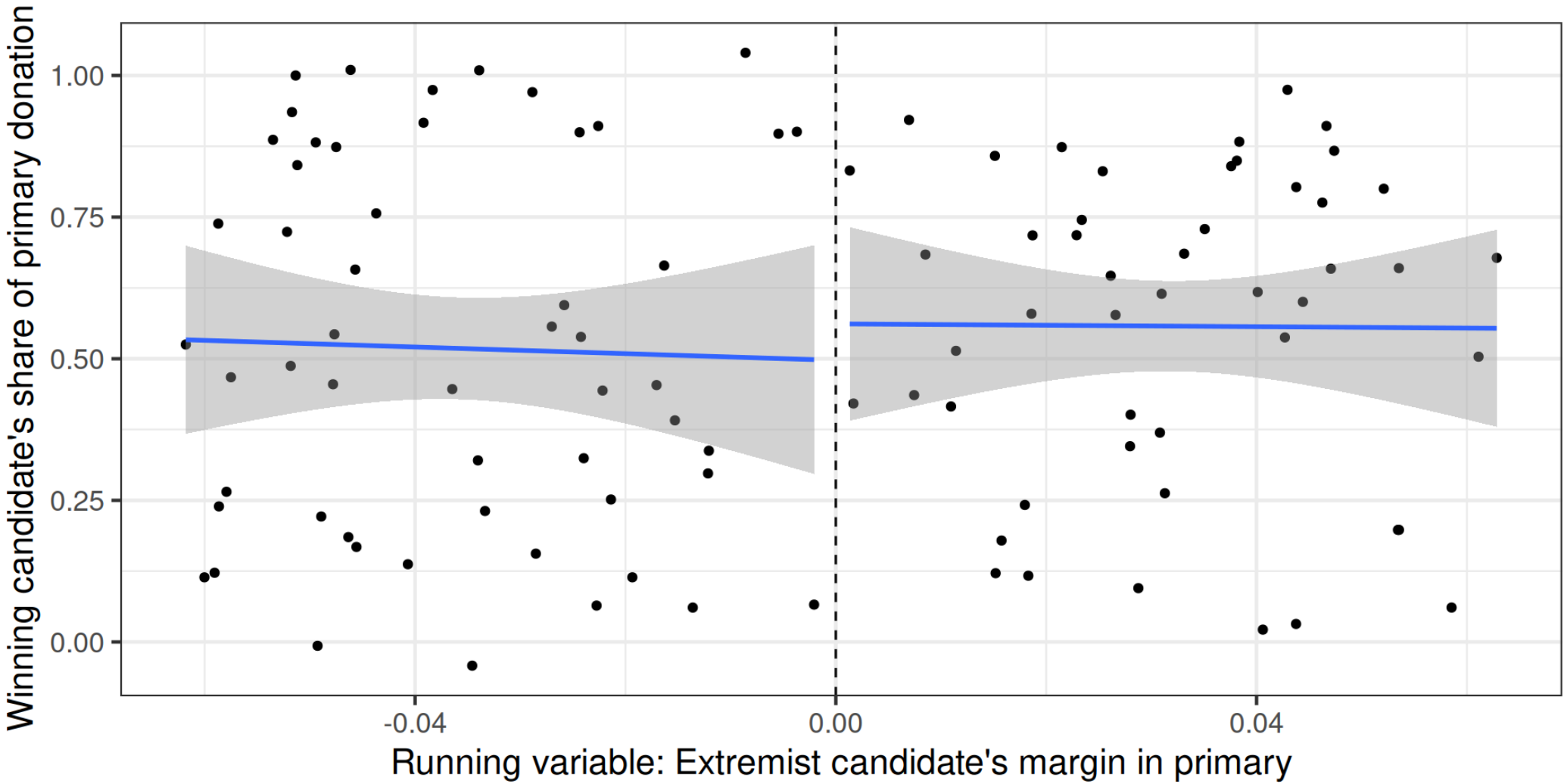
Visual balance check: Incumbency



Visual balance check: Experience



Visual balance check: Fundraising



Wrapping up

What we did today

RDD in practice

1. Starting point: Look at the data!
    - Is there an evident discontinuity in the raw data?
    - How close to the threshold is the relationship roughly linear?
2. Many different ways to estimate RDDs
    - Linear and polynomial via `lm()`
    - Automatic bandwidth selection via `rdrobust()`
    - Use multiple methods to see if results are broadly consistent
3. Assessing balance visually and with statistical tests
    - Ideal is no discontinuity in confounders

Next week

Time to work on your paper drafts — no class sessions next week

But you **must** sign up for a half hour meeting with me

- Use the Google Sheet linked on Brightspace

The more progress you've made beforehand, the better I'll be able to help you