Tutorial Week 11: Instrumental Variable (IV) Analysis and RDD

Maximilian Birkle   Daniel Lehmann   Henry Lucas

2025-11-18

Contents

1 Setup and Package Loading

```
# Install packages if needed
packages <- c("AER", "haven", "dplyr", "ggplot2", "stargazer",
              "boot", "lmtest", "sandwich", "knitr", "kableExtra",
              "rdrobust", "rddensity", "rdlocrand")

for (pkg in packages) {
  if (!require(pkg, character.only = TRUE)) {
    install.packages(pkg)
    library(pkg, character.only = TRUE)
  }
}
```

2 Load Data

```
# Load the AJR (2001) replication data
ajr_data <- haven::read_dta("acemoglu.dta")
cat("Observations:", nrow(ajr_data), "\n")
```

```
## Observations: 163
```

```r
cat("Variables:", ncol(ajr_data), "\n")
```

```
## Variables: 12
```

3 Problem 1: Instrumental Variable Analysis

3.1 Q1. Writing and Estimating the IV Model (10 pts)

1. **Write down the two-equation IV system** (first stage and second stage).

   Define explicitly:

   - Y_i: outcome
   - T_i: endogenous regressor
   - Z_i: instrument
   - X_i: controls

2. Estimate the following:

   - **First stage:**

   $$T_i = \alpha_1 + \beta_1 Z_i + \mathbf{X}_i' \gamma_1 + \varepsilon_{1i}$$

   - **Reduced form:**

   $$Y_i = \alpha_2 + \delta_1 Z_i + \mathbf{X}_i' \gamma_2 + \varepsilon_{2i}$$

   - **2SLS:**

   $$Y_i = \alpha_3 + \beta_2 T_i + \mathbf{X}_i' \gamma_3 + \varepsilon_{3i}$$

3. Report and interpret:

   - The reduced-form coefficient on Z_i.
   - The first-stage coefficient on Z_i.
   - The **first-stage F-statistic**. Is it above the rule-of-thumb threshold of ≈ 10?
   - The 2SLS estimate of the effect of T_i on Y_i. Is it statistically significant?

3.2 Q2. Randomization and Resampling (10 pts)

3.2.1 Q2(a). Permutation Test

1. **State the null hypothesis** being tested.

2. Conduct a **permutation test** for the IV coefficient:

   - Shuffle the endogenous variable (or fitted values) while holding other variables fixed.
   - Re-estimate the IV model for each permutation.
   - Construct the empirical null distribution.
   - Report the **permutation p-value**.

3. Compare this p-value to the **normal-approximation p-value** from your 2SLS output.

   Discuss any differences and what they imply for small-sample IV inference.

3.2.2 Q2(b). Bootstrap Confidence Intervals

Using bootstrap resampling of observations:

1. Generate bootstrap 2SLS estimates of the coefficient on T_i.
2. Construct three 95% confidence intervals:
   - **Efron percentile**
   - **Bias-corrected (BC—google this one.)**
3. Compare the three CIs:
   - Do they include zero?
   - Are they wider or narrower?
   - What does this imply about the sampling distribution?

3.2.3 Q2(c). Conceptual: Permutation vs Bootstrap

Explain—precisely—what the **permutation test** and the **bootstrap** each measure.

What is held fixed? What is resampled?

Why do they answer conceptually different questions?

3.3 Q3. Instrument Validity and Timing (10 pts)

3.3.1 Q3.1 Causal Priority in AJR's Theory

Why settler mortality must be causally prior to institutions:

[Write your answer here]

Key points to address:

- The logic of instrumental variables requires that $Z \to T \to Y$ (no reverse causation)
- In AJR's theory, high settler mortality → extractive institutions → lower growth
- If institutions could affect mortality rates retroactively, the IV assumption fails
- The exclusion restriction requires mortality affects GDP ONLY through institutions

3.3.2 Q3.2 Timing Problems

Issues with timing of settler mortality measurements:

[Write your answer here]

Address each IV assumption:

Relevance (First-stage):

- Why does measurement timing affect the strength of the first stage?
- Consider data availability and measurement error

Exclusion Restriction:

- If mortality was measured long after colonization, what other channels might exist?
- Could later mortality reflect economic conditions rather than cause them?

Independence:

- Are there confounders that affect both late-measured mortality and outcomes?
- Geographic or climatic factors?

3.3.3 Q3.3 Assessment of AJR Results

Based on your analysis:

Your assessment:

[Write your answer here]

Consider:

- Is the first stage strong enough?
- Are the IV assumptions plausible given the timing issues?
- What are the main threats to validity?
- Would you believe the causal interpretation?

3.4 Q4. Albouy's Critique of AJR (10 pts)

3.4.1 Q4(a). Measurement Problems

Key measurement issues raised by Albouy (2012):

1.
2.

Why they matter for IV validity:

[Your answer]

3.4.2 Q4(b). Violations of IV Assumptions

Relevance:

Albouy's argument:

[Your answer]

Independence:

Albouy's argument:

[Your answer]

Exclusion Restriction:

Albouy's argument:

[Your answer]

3.4.3 Q4(c). Sensitivity and Data Corrections

After Albouy reconstructs and corrects the mortality data:

Effects on:

1. **First stage:** [Your answer - what happens to F-statistic and coefficient?]
2. **Reduced form:** [Your answer - does the relationship weaken?]
3. **2SLS estimates:** [Your answer - how do the causal estimates change?]

What this reveals about stability:

[Your answer - are the AJR findings robust or fragile?]

3.4.4 Q4(d). Interpretation

Do you believe the AJR conclusions still hold?

[Your answer]

Or do the methodological issues undermine the core causal claim?

[Your answer]

Justification:

[Provide clear reasoning based on:

- The strength of Albouy's critique
- Your replication results
- The plausibility of IV assumptions
- The sensitivity of findings to data corrections]

4 Problem 2: Regression Discontinuity Design

4.1 Part A - Data Loading and Setup

```r
# Ensure RDD packages are loaded
library(rdrobust)
library(rddensity)
library(rdlocrand)
library(knitr)
library(kableExtra)

# Load RD data
data <- read.csv("CTV_2020_Sage.csv")

# Define outcome, running variable, and covariates
Y <- data$mv_incpartyfor1
X <- data$mv_incparty

covs <- data[, c("pibpc", "population", "numpar_candidates_eff",
                 "party_DEM_wonlag1_b1", "party_PSDB_wonlag1_b1",
                 "party_PT_wonlag1_b1", "party_PMDB_wonlag1_b1")]
covsnm <- c("GDP per capita", "Population", "No. Effective Parties",
            "DEM Victory t-1", "PSDB Victory t-1", "PT Victory t-1", "PMDB Victory t-1")

cat("RDD data loaded successfully!\n")
```

```
## RDD data loaded successfully!
```

```r
cat("Observations:", nrow(data), "\n")
```

```
## Observations: 27455
```

```r
cat("Running variable (X): Incumbent Party's Margin of Victory\n")
```

```
## Running variable (X): Incumbent Party's Margin of Victory
```

```r
cat("Outcome variable (Y): Incumbent Party Victory at t+1\n")
```

```
## Outcome variable (Y): Incumbent Party Victory at t+1
```

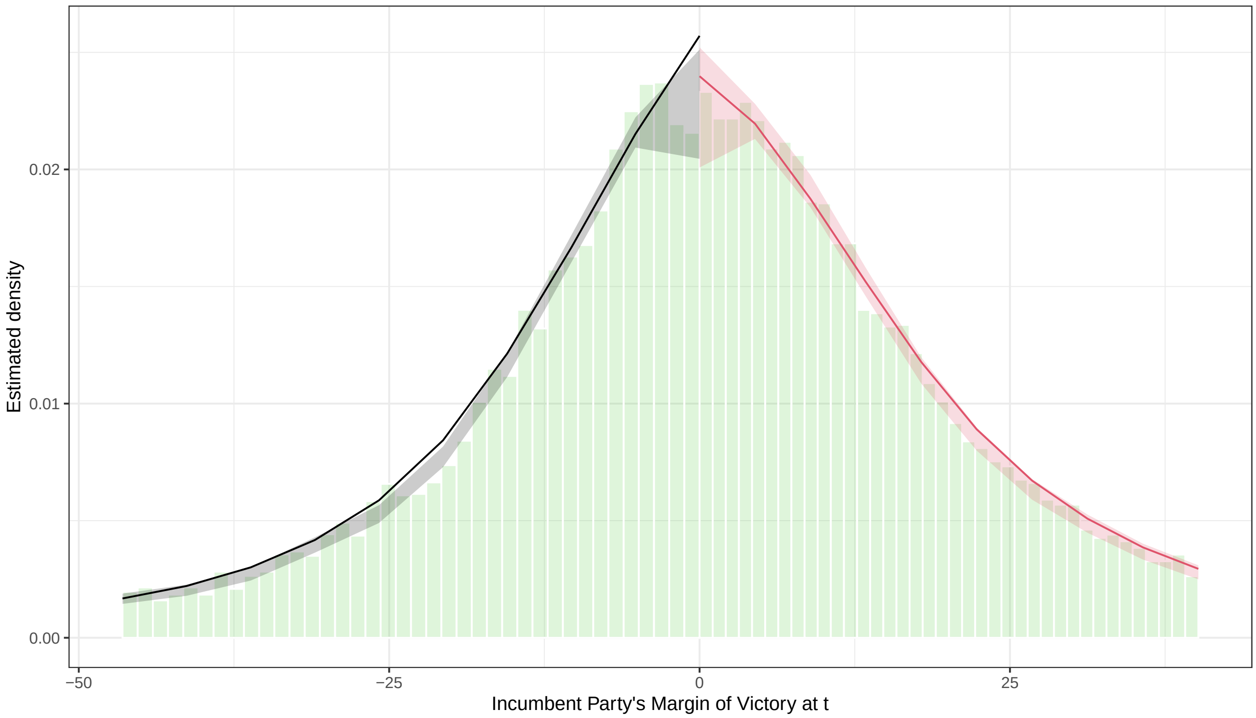
4.2 Part B - Falsification Analysis

4.2.1 Density Test

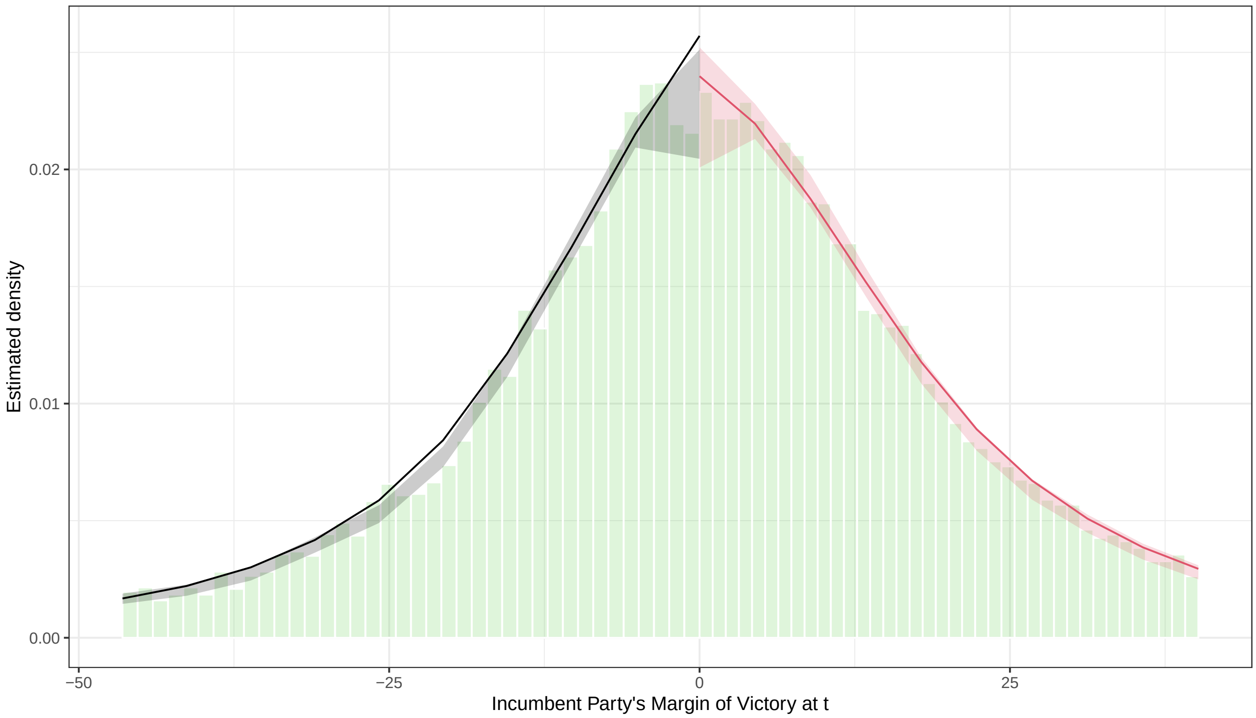
```r
# McCrary density test using rddensity
rddens <- rddensity(X)
summary(rddens)
```

```
##
## Manipulation testing using local polynomial density estimation.
##
## Number of obs =       13308
## Model =               unrestricted
## Kernel =              triangular
## BW method =           estimated
## VCE method =          jackknife
##
## c = 0                 Left of c           Right of c
## Number of obs         6088                7220
## Eff. Number of obs    3852                3590
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           15.493              13.392
##
## Method                T                   P > |T|
## Robust                -0.0757             0.9397

##
## P-values of binomial tests (H0: p=0.5).
##
## Window Length / 2          <c     >=c    P>|T|
## 0.098                      20      26    0.4614
## 0.195                      53      50    0.8439
## 0.293                      80      77    0.8732
## 0.390                     112     114    0.9470
## 0.488                     141     140    1.0000
## 0.585                     180     184    0.8751
## 0.683                     199     213    0.5219
## 0.780                     231     240    0.7125
## 0.878                     254     267    0.5991
## 0.975                     278     296    0.4780
```

```r
# Plot the density
rdplotdensity(rddens, X = data$mv_incparty[!is.na(data$mv_incparty)],
              xlab = "Incumbent Party's Margin of Victory at t",
              ylab = "Estimated density")
```




```
## $Estl
## Call: lpdensity
##
## Sample size                                      6088
## Polynomial order for point estimation    (p=)    2
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    3
## Kernel function                                  triangular
## Scaling factor                                   0.457503569549861
## Bandwidth method                                 user provided
##
## Use summary(...) to show estimates.
##
## $Estr
## Call: lpdensity
##
## Sample size                                      7220
## Polynomial order for point estimation    (p=)    2
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    3
## Kernel function                                  triangular
## Scaling factor                                   0.542571578868265
## Bandwidth method                                 user provided
##
## Use summary(...) to show estimates.
##
## $Estplot
```




```r
# Create summary table for density test
density_results <- data.frame(
  Test = "Manipulation Test",
  Statistic = sprintf("%.4f", rddens$test$t_jk),
  `p-value` = sprintf("%.4f", rddens$test$p_jk),
  Bandwidth = sprintf("%.2f / %.2f", rddens$h$left, rddens$h$right),
  `N (Left)` = format(rddens$N$left, big.mark = ","),
  `N (Right)` = format(rddens$N$right, big.mark = ","),
  `Eff. N (Left)` = format(rddens$N$eff_left, big.mark = ","),
  `Eff. N (Right)` = format(rddens$N$eff_right, big.mark = ","),
  check.names = FALSE
)

kable(density_results,
      booktabs = TRUE,
      align = c("l", "r", "r", "c", "r", "r", "r", "r"),
      caption = "Manipulation Test: Continuity of Density at Threshold") %>%
  kable_styling(full_width = FALSE, position = "center")
```

Table 1: Manipulation Test: Continuity of Density at Threshold

Test	Statistic	p-value	Bandwidth	N (Left)	N (Right)	Eff. N (Left)	Eff. N (Right)
Manipulation Test	-0.0757	0.9397	15.49 / 13.39	6,088	7,220	3,852	3,590

Note: Test statistic based on local polynomial density estimation with triangular kernel. Null hypothesis: No discontinuity in density at threshold (no manipulation). High p-value indicates no evidence of manipulation around the cutoff.

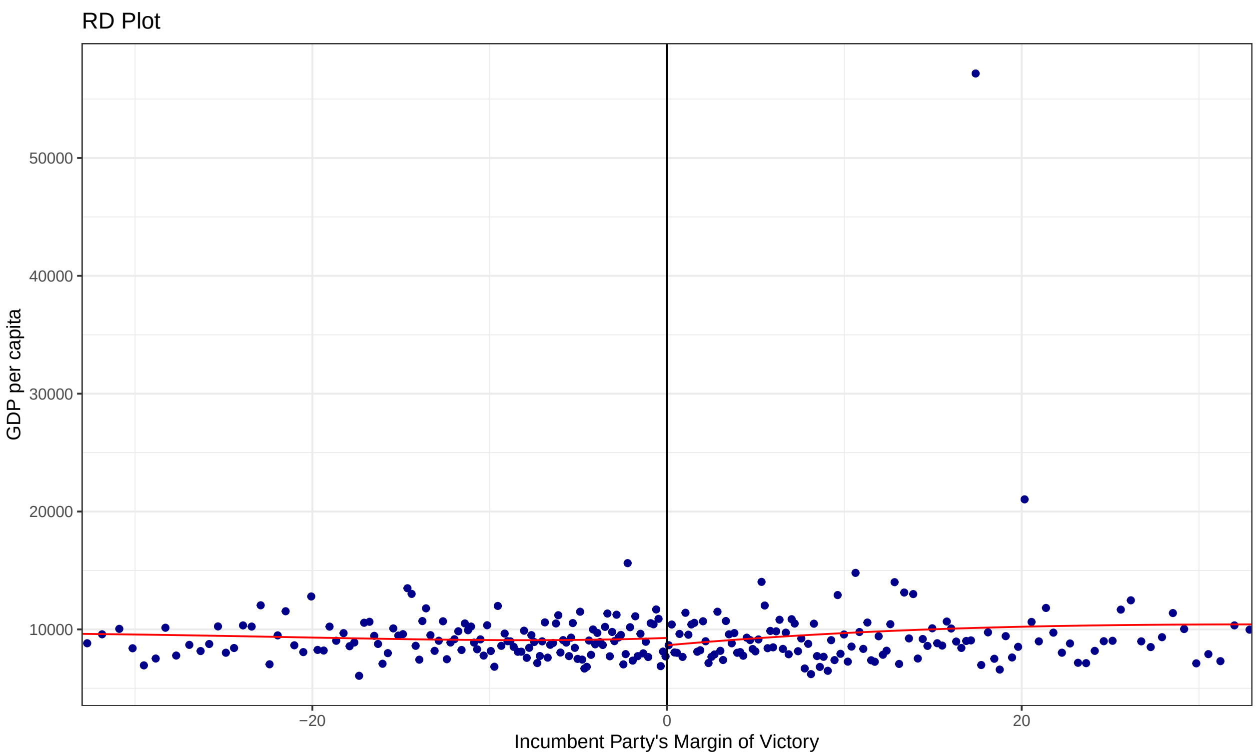
Interpretation:

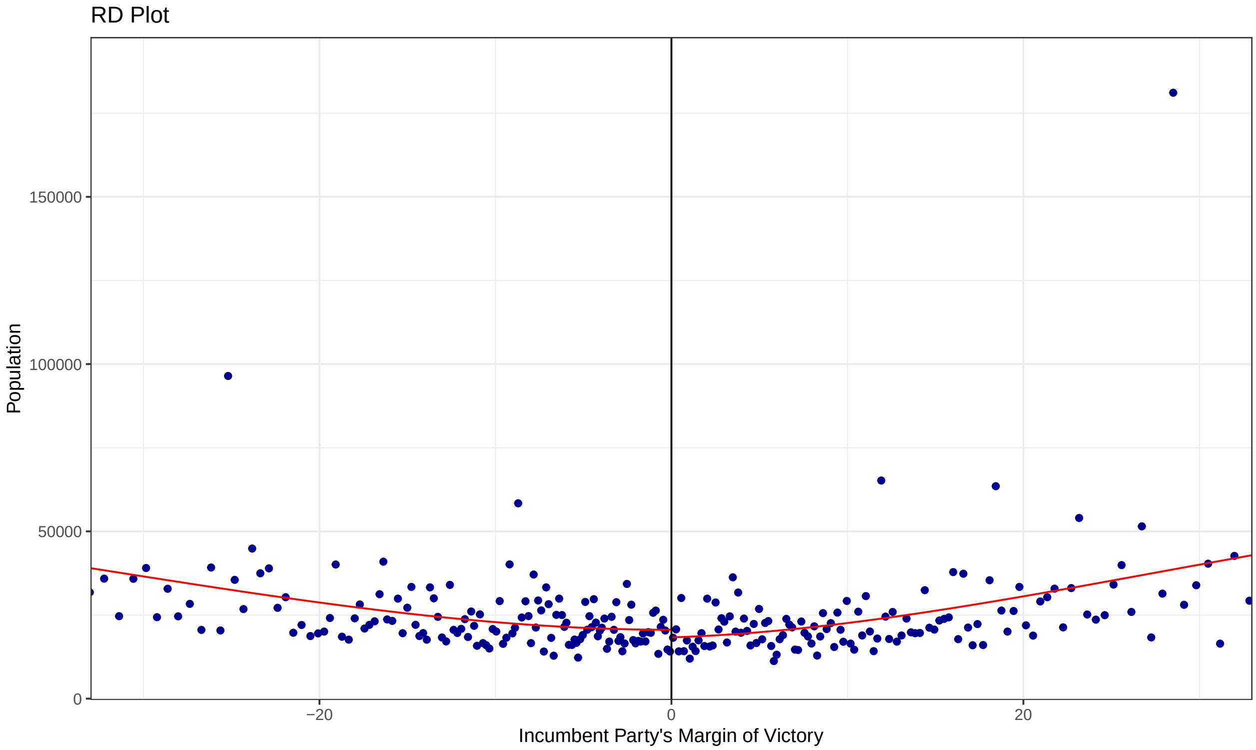
The density test checks whether there is evidence of manipulation around the threshold (margin of victory =

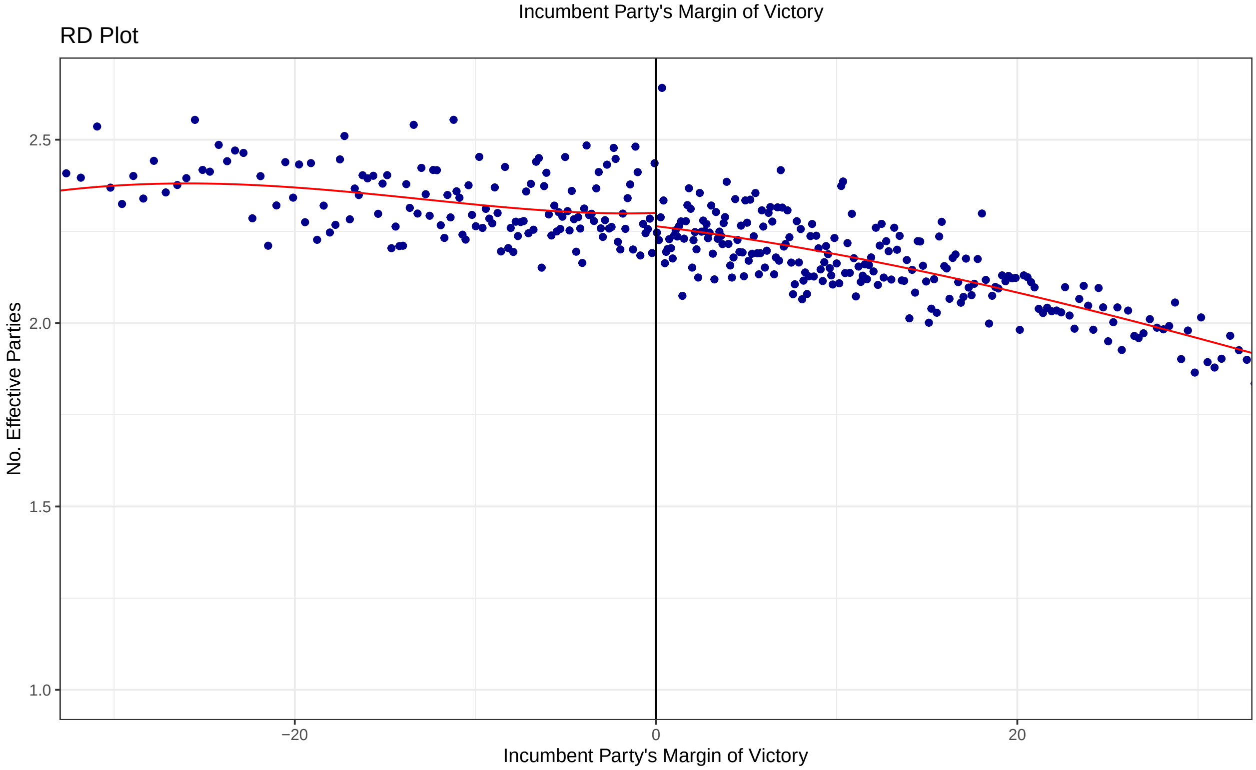
0). A significant discontinuity in the density would suggest that parties can manipulate their vote margins to barely win elections, which would violate the RD identifying assumptions.

4.2.2 Covariate Balance Tests

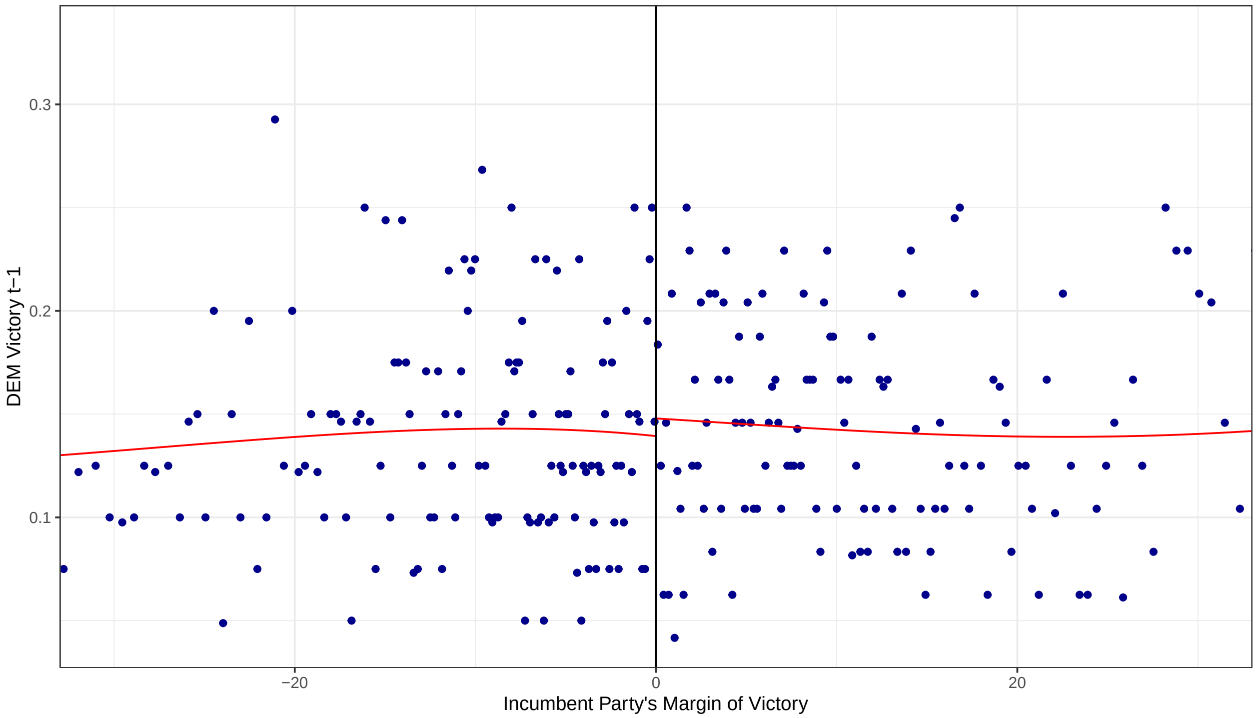
```
# Initialize lists to store results
balance_results <- list()

# Test for balance in covariates at the threshold
for(c in 1:ncol(covs)){
  cat("\n")
  rdr_cov <- rdrobust(covs[,c], X)
  balance_results[[c]] <- rdr_cov

  # Create RD plot for this covariate
  rdplot(covs[,c], X,
         y.label = covsnm[c],
         x.label = "Incumbent Party's Margin of Victory",
         x.lim = c(-30, 30),
         binselect = "qsmv")
}
```



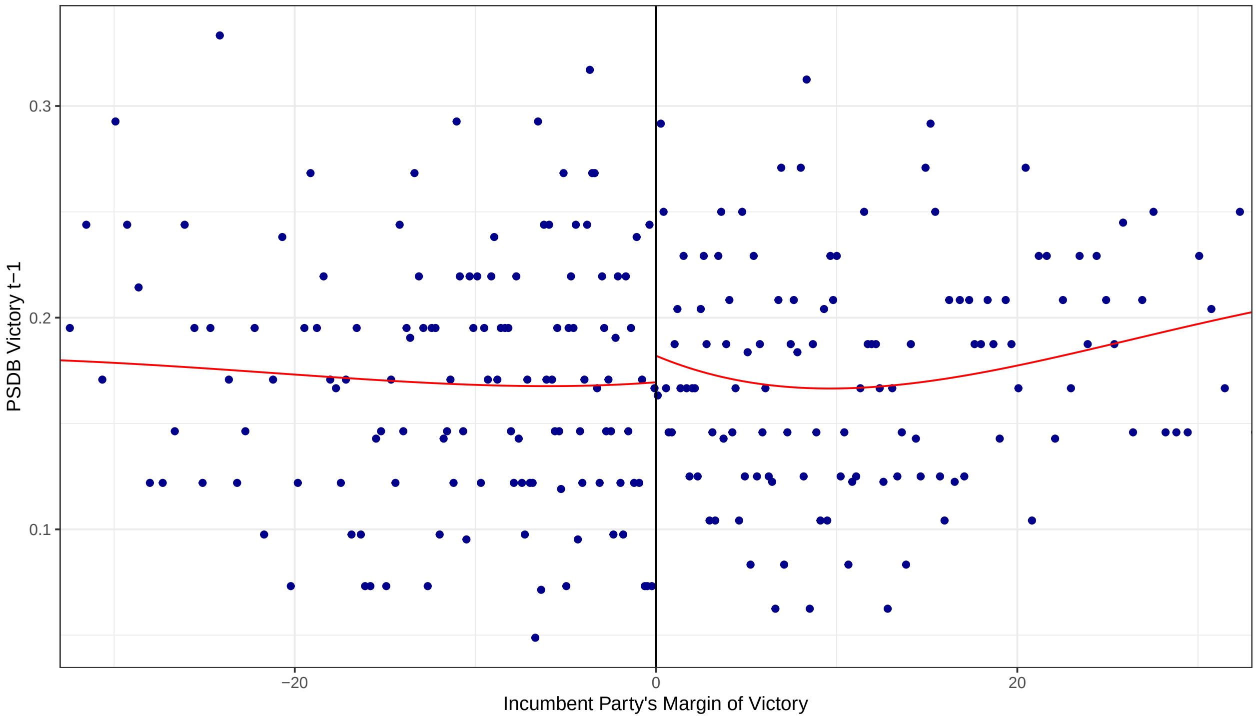

RD Plot
Population
150000
100000
50000
0
−20
0
20
Incumbent Party's Margin of Victory

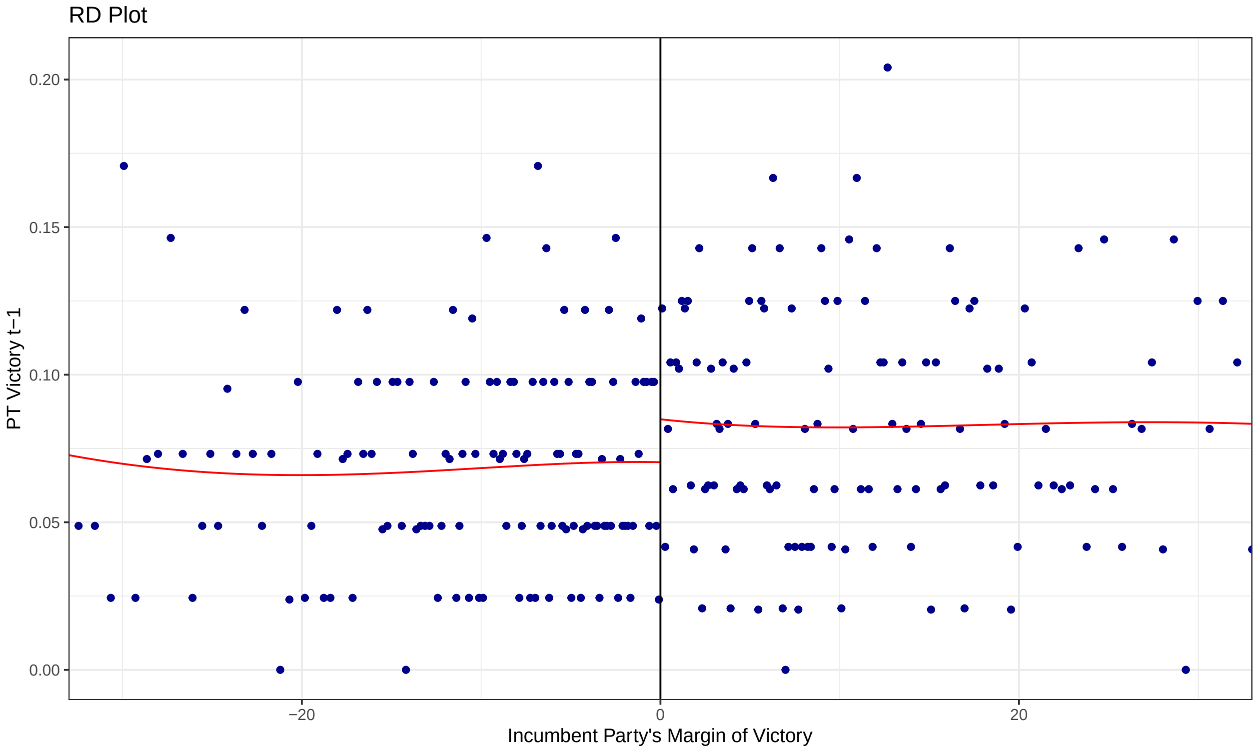

RD Plot
No. Effective Parties
2.5
2.0
1.5
1.0
−20
0
20
Incumbent Party's Margin of Victory

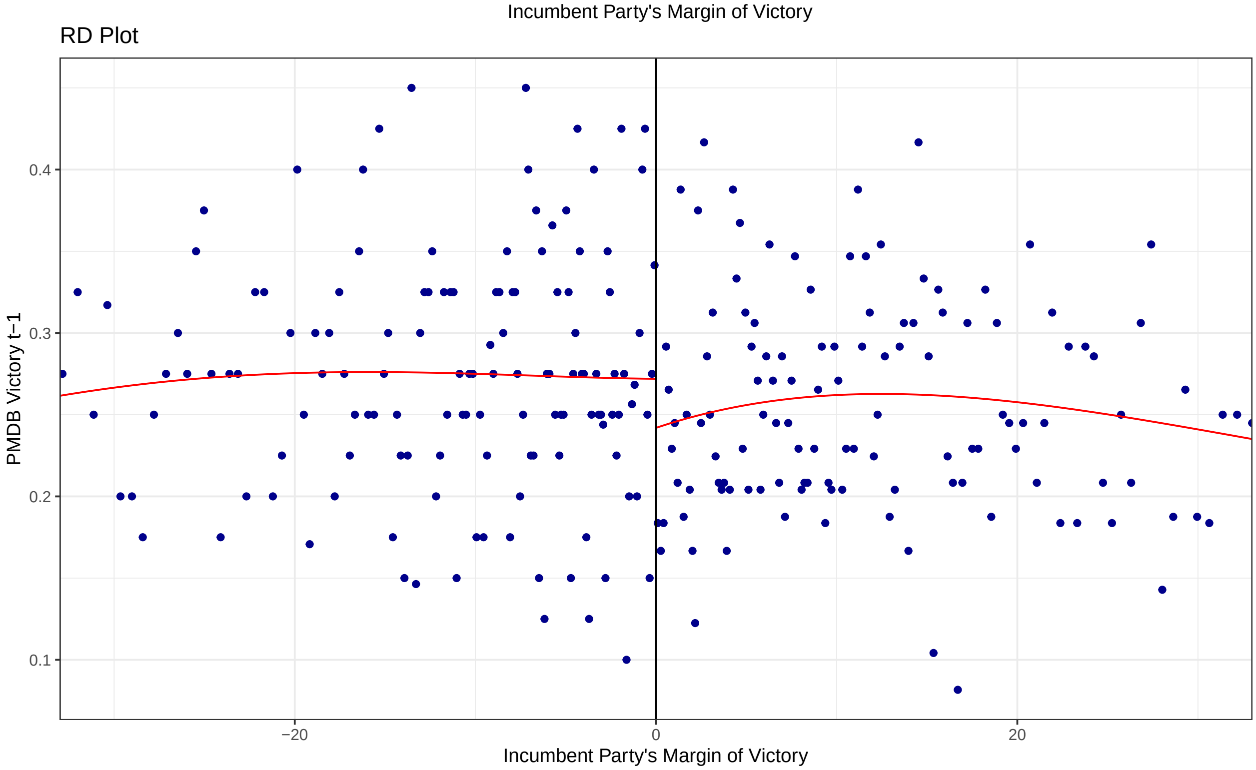
RD Plot



RD Plot








```r
# Create a summary table of balance tests
balance_table <- data.frame(
  Covariate = covsnm,
  Coefficient = sprintf("%.4f", sapply(balance_results, function(x) x$coef[1])),
  `Robust SE` = sprintf("%.4f", sapply(balance_results, function(x) x$se[3])),
  `t-statistic` = sprintf("%.3f", sapply(balance_results, function(x) x$z[3])),
  `p-value` = sprintf("%.3f", sapply(balance_results, function(x) x$pv[3])),
```

```
  `N (Left)` = format(sapply(balance_results, function(x) x$N_h[1]), big.mark = ","),
  `N (Right)` = format(sapply(balance_results, function(x) x$N_h[2]), big.mark = ","),
  check.names = FALSE
)

kable(balance_table,
      booktabs = TRUE,
      align = c("l", "r", "r", "r", "r", "r", "r"),
      caption = "Covariate Balance Tests at Threshold") %>%
  kable_styling(full_width = FALSE, position = "center")
```

Table 2: Covariate Balance Tests at Threshold

Covariate	Coefficient	Robust SE	t-statistic	p-value	N (Left)	N (Right)
GDP per capita	-149.9175	560.7565	-0.190	0.849	3,626	3,740
Population	-22.5594	2124.3938	0.177	0.860	2,514	2,543
No. Effective Parties	-0.0363	0.0301	-1.368	0.171	3,312	3,409
DEM Victory t-1	0.0049	0.0180	0.119	0.905	4,026	4,183
PSDB Victory t-1	0.0228	0.0199	1.439	0.150	3,800	3,914
PT Victory t-1	0.0131	0.0135	1.000	0.317	4,407	4,645
PMDB Victory t-1	-0.0347	0.0234	-1.707	0.088	3,837	3,962

Note: Robust bias-corrected RD estimates using MSE-optimal bandwidth selection. Standard errors calculated using nearest-neighbor variance estimator.

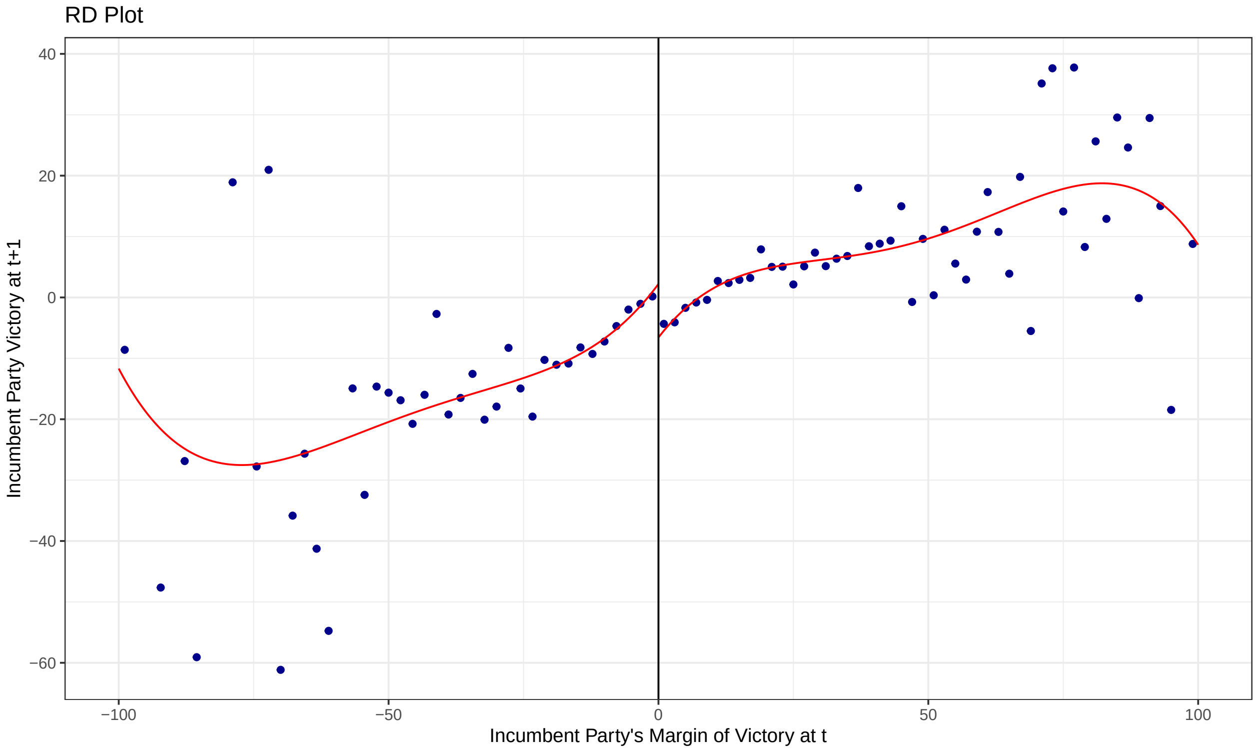
Interpretation:

Covariate balance tests check whether pre-treatment characteristics are continuous at the threshold. If covariates jump discontinuously at the cutoff, this suggests the treatment is not "as-if" randomly assigned, which would undermine the validity of the RD design.

4.3 Part C - Outcome Analysis

4.3.1 RD Plot

```
# Create RD plot of outcome variable
rdplot(Y, X,
       y.label = "Incumbent Party Victory at t+1",
       x.label = "Incumbent Party's Margin of Victory at t")
```



4.3.2 Continuity-Based RDD Analysis

```
# Main RDD estimate without covariates
rdr <- rdrobust(Y, X)

# RDD estimate with covariates
rdrcovs <- rdrobust(Y, X, covs = covs)

# Create a comprehensive results table
results_table <- data.frame(
  Specification = c("Without Covariates", "With Covariates"),
  Coefficient = sprintf("%.4f", c(rdr$coef[1], rdrcovs$coef[1])),
  `Conv. SE` = sprintf("%.4f", c(rdr$se[1], rdrcovs$se[1])),
  `Robust SE` = sprintf("%.4f", c(rdr$se[3], rdrcovs$se[3])),
  `p-value` = sprintf("%.3f", c(rdr$pv[3], rdrcovs$pv[3])),
  `95% CI` = sprintf("[%.3f, %.3f]", c(rdr$ci[3,1], rdrcovs$ci[3,1]),
                                      c(rdr$ci[3,2], rdrcovs$ci[3,2])),
  `BW (L/R)` = sprintf("%.2f / %.2f", c(rdr$bws[1], rdrcovs$bws[1]),
                                      c(rdr$bws[2], rdrcovs$bws[2])),
  `N (Left)` = format(c(rdr$N_h[1], rdrcovs$N_h[1]), big.mark = ","),
  `N (Right)` = format(c(rdr$N_h[2], rdrcovs$N_h[2]), big.mark = ","),
  check.names = FALSE
)

kable(results_table,
      booktabs = TRUE,
      align = c("l", "r", "r", "r", "r", "c", "c", "r", "r"),
      caption = "Regression Discontinuity Estimates: Effect of Incumbent Party Victory on Future Electo
  kable_styling(full_width = FALSE, position = "center")
```

Table 3: Regression Discontinuity Estimates: Effect of Incumbent Party Victory on Future Electoral Success

Specification	Coefficient	Conv. SE	Robust SE	p-value	95% CI	BW (L/R)	N (Left)	N (Ri
Without Covariates	-6.2813	1.6008	1.8571	0.000	[-10.223, -2.944]	15.29 / 27.51	1,533	1
With Covariates	-6.1061	1.5782	1.8433	0.001	[-9.881, -2.655]	14.45 / 25.25	1,481	1

```r
# Print detailed summary for reference
cat("\n=== DETAILED RDD ESTIMATES ===\n\n")
```

```
##
## === DETAILED RDD ESTIMATES ===
```

```r
cat("--- Without Covariates ---\n")
```

```
## --- Without Covariates ---
```

```r
print(summary(rdr))
```

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 5463
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 2220         3243
## Eff. Number of Obs.            1533         1740
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                  15.291       15.291
## BW bias (b)                  27.509       27.509
## rho (h/b)                     0.556        0.556
## Unique Obs.                    2213         3119
##
## =====================================================================
##                    Point    Robust Inference
##                 Estimate         z     P>|z|      [ 95% C.I. ]
## ---------------------------------------------------------------------
##      RD Effect    -6.281    -3.545     0.000   [-10.223 , -2.944]
## =====================================================================
## NULL
```

```r
cat("\n--- With Covariates ---\n")
```

```
##
## --- With Covariates ---
```

```r
print(summary(rdrcovs))
```

```
## Covariate-adjusted Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 5460
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
```

```
##
## Number of Obs.                 2218         3242
## Eff. Number of Obs.            1481         1672
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                  14.451       14.451
## BW bias (b)                  25.248       25.248
## rho (h/b)                     0.572        0.572
## Unique Obs.                    2211         3118
##
## =====================================================================
##                    Point    Robust Inference
##                 Estimate         z     P>|z|      [ 95% C.I. ]
## ---------------------------------------------------------------------
##      RD Effect    -6.106    -3.401     0.001    [-9.881 , -2.655]
## =====================================================================
## NULL
```

Notes: Dependent variable is Incumbent Party Victory at t+1 (in percentage points). Robust bias-corrected confidence intervals and p-values reported. MSE-optimal bandwidth selection with triangular kernel. BW (L/R) shows left/right bandwidths. Covariates include GDP per capita, Population, No. Effective Parties, and DEM/PSDB/PT/PMDB Victory at t-1.

Interpretation:

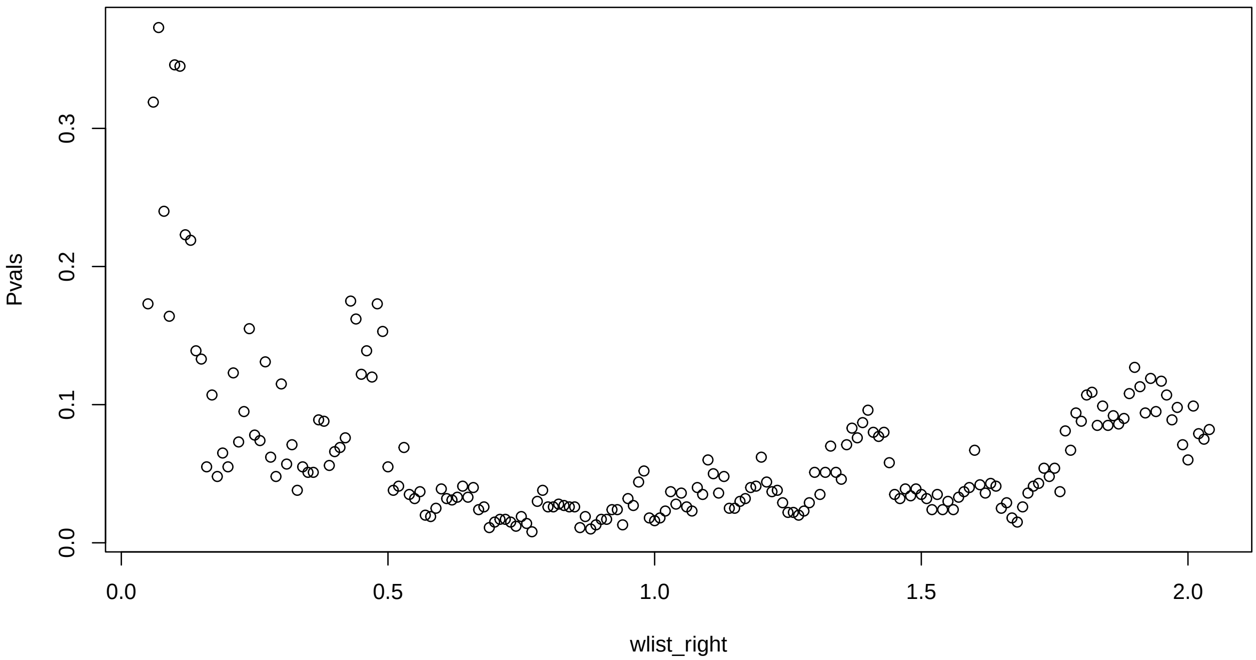
The RDD estimates show the causal effect of barely winning an election (vs. barely losing) on the probability of the incumbent party winning the next election. This tests the "incumbency curse" hypothesis - that winning may actually hurt a party's chances in the next election due to weak parties and term limits in Brazilian municipalities.

Compare the estimates with and without covariates. If they are similar, this provides additional evidence that the RD design is valid (covariates should not matter much if treatment is as-if random near the threshold).

4.4 Part D - Local Randomization Approach

```r
# Window selection for local randomization
rdwin <- rdwinselect(X, covs, wmin = 0.05, wstep = 0.01, nwindows = 200,
                     seed = 765, plot = TRUE, quietly = TRUE)
```

```
## Mass points detected in running variable
## You may use wmasspoints option for constructing windows at each mass point
```




```r
# Use selected window (or manually choose)
w <- 0.15

# Randomization inference
rdrand <- rdrandinf(Y, X, wl = -w, wr = w, reps = 1000, seed = 765)
```

```
##
## Selected window = [-0.15;0.15]
##
## Running randomization-based test...
## Randomization-based test complete.
##
##
## Number of obs      =          5463
## Order of poly      =             0
## Kernel type        =       uniform
## Reps               =          1000
## Window             =    set by user
## H0:           tau  =         0.000
## Randomization      = fixed margins
##
## Cutoff c =     0.000   Left of c  Right of c
##        Number of obs        2220        3243
##   Eff. number of obs          19          20
##      Mean of outcome       0.631      -9.361
##      S.d. of outcome      17.733      16.610
##               Window      -0.150       0.150
##
## ================================================================================
##                                   Finite sample           Large sample
##                                ------------------   ----------------------------
##           Statistic          T         P>|T|        P>|T|     Power vs d =   8.867
## ================================================================================
```

```
##      Diff. in means     -9.992        0.076        0.070                        0.363
## ==================================================================================
```

```r
# Create local randomization results table
local_rand_table <- data.frame(
  Approach = c("Continuity-Based", "Local Randomization"),
  Window = c(sprintf("±%.2f / ±%.2f", rdr$bws[1], rdr$bws[2]),
             sprintf("±%.2f", w)),
  Coefficient = sprintf("%.4f", c(rdr$coef[1], rdrand$obs.stat)),
  `Robust SE` = c(sprintf("%.4f", rdr$se[3]), "-"),
  `p-value` = sprintf("%.3f", c(rdr$pv[3], rdrand$p.value)),
  `95% CI` = c(sprintf("[%.3f, %.3f]", rdr$ci[3,1], rdr$ci[3,2]), "-"),
  `N (Left)` = format(c(rdr$N_h[1], rdrand$sumstats[1]), big.mark = ","),
  `N (Right)` = format(c(rdr$N_h[2], rdrand$sumstats[2]), big.mark = ","),
  Method = c("Asymptotic", "Permutation"),
  check.names = FALSE
)

kable(local_rand_table,
      booktabs = TRUE,
      align = c("l", "c", "r", "r", "r", "c", "r", "r", "l"),
      caption = "Comparison of RDD Approaches: Continuity-Based vs. Local Randomization") %>%
  kable_styling(full_width = FALSE, position = "center")
```

Table 4: Comparison of RDD Approaches: Continuity-Based vs. Local Randomization

Approach	Window	Coefficient	Robust SE	p-value	95% CI	N (Left)	N (Right)
Continuity-Based	±15.29 / ±27.51	-6.2813	1.8571	0.000	[-10.223, -2.944]	1,533	1,740
Local Randomization	±0.15	-9.9923	—	0.076	—	2,220	19

```r
cat("\n=== LOCAL RANDOMIZATION INFERENCE (Detailed) ===\n")
```

```
##
## === LOCAL RANDOMIZATION INFERENCE (Detailed) ===
```

```r
print(summary(rdrand))
```

```
##             Length Class  Mode
## sumstats    10     -none- numeric
## obs.stat     1     -none- numeric
## p.value      1     -none- numeric
## asy.pvalue   1     -none- numeric
## window       2     -none- numeric
```

Notes: Dependent variable is Incumbent Party Victory at t+1 (in percentage points). Local randomization uses 1,000 permutations with fixed margins assumption. Continuity-based approach uses MSE-optimal bandwidth with robust bias-correction. Local randomization assumes as-if random assignment within the narrow window.

Interpretation:

The local randomization approach assumes that units very close to the threshold (within a narrow window) are essentially randomly assigned to treatment. This provides an alternative inference method that doesn't rely on asymptotic approximations and may be more appropriate with discrete running variables.

4.5 Part E - Assessment

Key Findings:

1. **Density Test:** [Interpret whether there is evidence of manipulation]
2. **Covariate Balance:** [Summarize whether covariates are balanced]
3. **RDD Estimates:** [State the main findings about the incumbency effect]
4. **Robustness:** [Assess whether estimates are stable across specifications]

Overall Validity:

[Your assessment of whether the RDD design is credible in this context and whether the findings support the "incumbency curse" hypothesis]