Problem set #3

2025-03-13

Packages Used:

```
pacman::p_load(tidyverse,gt,foreign,knitr,webshot2,
               reshape2,viridis,ggthemes,stargazer,texreg,
               sandwich,modelsummary,Rmisc,neatStats,gmodels, dplyr, tidyr)
```

Load and Inspect the Data

```
df_ajr <- read.dta("ajr_data.dta")
```

check NA

```
table(is.na(df_ajr$column_name))
```

```
## < table of extent 0 >
```

check NA

```
head(df_ajr, 6)
```

```
##   shortnam logpgp95   avexpr   logem4  lat_abst africa asia other america
## 1      AGO 7.770645 5.363636 5.634789 0.1366667      1    0     0       0
## 2      ARG 9.133459 6.386364 4.232656 0.3777778      0    0     0       1
## 3      AUS 9.897972 9.318182 2.145931 0.3000000      0    0     1       0
## 4      BFA 6.845880 4.454545 5.634789 0.1444445      1    0     0       0
## 5      BGD 6.877296 5.136364 4.268438 0.2666667      0    1     0       0
## 6      BHS 9.285448 7.500000 4.442651 0.2683333      0    0     0       1
```

Setup and naive OLS

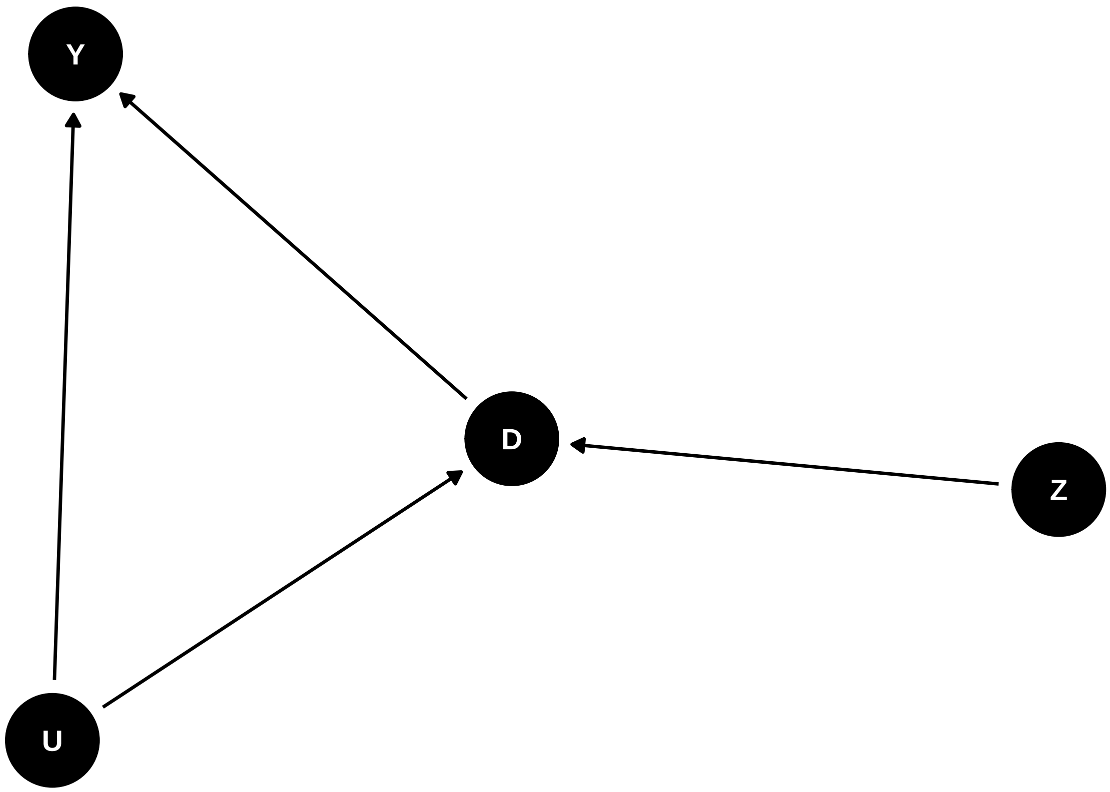
Assuming that their empirical strategy is valid, draw a simple DAG to represent the instrumental variables approach used by AJR. Include a hypothetical unobserved confounder that creates a back-door path between treatment and outcome. Why is it important to include this hypothetical unobserved confounder? What phenomena might the unobserved confounder represent? The reason why to include this confounder is that we can never prove there's no hidden factors that would not affect the outcome variable. In this case, U could be cultural norms, ethnicity, geographical resources…etc that would also shape institutions and long-run economic growth. And it's important to consider these and examine if they're correlated with.

```
library(ggdag)
```

```
##
## Attaching package: 'ggdag'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(dagitty)

dag <- dagitty("dag {
  Z -> D -> Y
  U -> D
  U -> Y
}")

ggdag(dag, text = TRUE) + theme_dag()
```



We will now replicate the main specifications from AJR. Using OLS, estimate the effect of avexpr on loggp95 in two ways, without using instrumental variables regression. First, estimate a linear regression with loggp95 as the dependent variable, and avexpr as the lone regressor (do not include any other covariates). Second, do the same but include, linearly and additively, lat_abst, africa , asia , and other. Present the results in a table, including HC2 robust standard errors. Interpret the direction and statistical significance of the estimates. Why should we be concerned about whether these are good estimates of the causal quantity of interest? Broadly, are these concerns issues of "estimation" or "identification"? There may be unobserved factors that influence both institutions and economic outcomes, leading to biased estimates. The concern with these estimates is whether they truly reflect a causal relationship between institutions and economic development, or if they are simply capturing correlations driven by other factors.This is a problem of identification, because if it was not isolate the exogenous variation in institutions by using IV, then the rsult can not be interpreted as causal or we can not attributed the detected effect to institutions.

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)
library(stargazer)

# without other covariates
md_1 <- lm(logpgp95 ~ avexpr, data = df_ajr)
md_2 <- lm(logpgp95 ~ avexpr + lat_abst + africa + asia + other, data = df_ajr)

robust_se1 <- coeftest(md_1, vcov = vcovHC(md_1, type = "HC2"))
robust_se2 <- coeftest(md_2, vcov = vcovHC(md_2, type = "HC2"))

stargazer(md_1, md_2,
          type = "text",
          se = list(robust_se1[,2], robust_se2[,2]),
          omit.stat = c("f", "ser")
)
```

```
##
## ==========================================
##                   Dependent variable:
##               ----------------------------
##                         logpgp95
##                    (1)            (2)
## ------------------------------------------
## avexpr          0.522***      0.401***
##                 (0.050)        (0.066)
##
## lat_abst                       0.875
##                                (0.628)
##
## africa                        -0.881***
##                                (0.154)
##
## asia                           -0.577*
##                                (0.307)
##
## other                          0.107
##                                (0.251)
##
## Constant        4.660***      5.737***
##                 (0.322)        (0.396)
##
## ------------------------------------------
## Observations       64            64
## R2               0.540          0.714
## Adjusted R2      0.533          0.689
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

IV estimates

Now, again using OLS, estimate the effect of logem4 on loggp95 . First, estimate a linear regression with loggp95 as the dependent variable, and logem4 as the lone regressor (do not include any other covariates). Second, do the same but include, linearly and additively, lat_abst, africa , asia , and other . Present the results in a table, including HC2 robust standard errors. Interpret the direction and statistical significance of the estimate of the causal effect. What does this "reduced form" estimator purport to estimate? Under what conditions can we interpret this result as causal?

The table shows a strong negative relationship between settler mortality and economic growth. Higher historical mortality rates are associated with lower modern economic development. This is aligned with AJR's argument, althought the the effect weakens with controls. It captures the total impact of settler mortality on GDP, including both institutional and non-institutional channels, but doesn't directly estimate the effect of institutions. In order to interpret them as causal, settler mortality must be exogenous, affecting GDP only through institutions. If this holds, it serves as a valid instrument; otherwise, the estimate may reflect multiple influences beyond institutions.

```r
md_3 <- lm(logpgp95 ~ logem4, data = df_ajr)
md_4 <- lm(logpgp95 ~ logem4 + lat_abst + africa + asia + other, data = df_ajr)

robust_se3 <- coeftest(md_3, vcov = vcovHC(md_3, type = "HC2"))
robust_se4 <- coeftest(md_4, vcov = vcovHC(md_4, type = "HC2"))

stargazer(md_3, md_4,
          type = "text",
          se = list(robust_se3[,2], robust_se4[,2]),
          omit.stat = c("f", "ser")
)
```

```
##
## =========================================
##                   Dependent variable:
##               ----------------------------
##                         logpgp95
##                   (1)            (2)
## -----------------------------------------
## logem4        -0.573***      -0.377***
##                (0.074)        (0.145)
##
## lat_abst                       1.046
##                               (0.886)
##
## africa                       -0.723***
##                               (0.262)
##
## asia                          -0.525
##                               (0.382)
##
## other                          0.185
##                               (0.257)
##
## Constant       10.731***      9.997***
##                (0.385)        (0.767)
##
## -----------------------------------------
```

```
## Observations        64             64
## R2                0.477          0.584
## Adjusted R2       0.469          0.548
## ==========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

Use instrumental variables regression to estimate the (Conditional) Local Average Treatment Effect (LATE) of avexpr on loggp95, using logem4 as the instrument for avexpr. You may use any function or package of your choice. As before, first include no covariates, and second include linearly and additively lat_abst, africa, asia, and other.

```r
library(AER)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: survival
```

```r
iv_1 <- ivreg(logpgp95 ~ avexpr | logem4, data = df_ajr)
iv_2 <- ivreg(logpgp95 ~ avexpr + lat_abst + africa + asia + other | logem4 + lat_abst + africa + asia +

robust_iv1 <- coeftest(iv_1, vcov = vcovHC(iv_1, type = "HC2"))
robust_iv2 <- coeftest(iv_2, vcov = vcovHC(iv_2, type = "HC2"))

stargazer(iv_1, iv_2,
          type = "text",
          se = list(robust_iv1[,2], robust_iv2[,2]),
          omit.stat = c("f", "ser")
)
```

```
##
## ==========================================
##                   Dependent variable:
##                 ----------------------------
##                         logpgp95
##                    (1)            (2)
## ------------------------------------------
## avexpr           0.944***       1.107**
##                  (0.179)        (0.525)
##
## lat_abst                        -1.178
##                                 (1.915)
##
## africa                          -0.437
##                                 (0.390)
##
```

```
## asia                            -1.047*
##                                 (0.538)
##
## other                           -0.990
##                                 (1.205)
##
## Constant         1.910           1.440
##                 (1.192)         (3.209)
##
## -----------------------------------------
## Observations       64              64
## R2               0.187           0.011
## Adjusted R2      0.174          -0.074
## =========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Report and interpret the F-statistic from a test for weak instrumentation based on the models above. What do you find? F > 10 suggests that logem4 is not a weak IV, which avoid suffering from bias estimate. The positive and statistically significant coefficient on avexpr aligns with the argument that stronger property rights institutions lead to higher economic groowth.

```
first_stage <- lm(avexpr ~ logem4, data = df_ajr)
summary(first_stage)
```

```
##
## Call:
## lm(formula = avexpr ~ logem4, data = df_ajr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6606 -0.9922  0.0280  0.8266  3.3566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.3414     0.6107   15.30  < 2e-16 ***
## logem4       -0.6068     0.1267   -4.79 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.265 on 62 degrees of freedom
## Multiple R-squared:  0.2701, Adjusted R-squared:  0.2584
## F-statistic: 22.95 on 1 and 62 DF,  p-value: 1.077e-05
```

Regression Discontinuity Designs

Load and Inspect the Data

```
df_lee <- read.dta("lee.dta", convert.factors = FALSE)
```

check NA

```
table(is.na(df_lee$column_name))
```

```
## < table of extent 0 >
```

check NA

```r
head(df_lee, 6)
```

```
##   state distnum distid party partname yearel origvote totvote highestvote
## 1     1       1      1   100            1946    82231  175237       93006
## 2     1       1      1   200            1946    93006  175237       93006
## 3     1       1      1   100            1948   127802  233700      127802
## 4     1       1      1   200            1948   103294  233700      127802
## 5     1       1      1   100            1950   134258  231096      134258
## 6     1       1      1   200            1950    96251  231096      134258
##   sechighestvote uniqid officeexp
## 1          82231  15937         0
## 2          82231  19281         0
## 3         103294  23403         0
## 4         103294  19281         1
## 5          96251  23403         1
## 6          96251  25775         0
```

Setup

Create the following three variables: - share_t: Vote share in the current election (candidate's vote share divided by the total number of votes). This is your dependent variable. - margin_tm1: Party's vote margin in the previous election. This is your "forcing" variable representing the party's share of votes cast for the top two candidates in the previous election. Adjust so that the cutpoint lies at 50%. - incumbent: A binary "treatment" indicator that takes '1' if the party won the previous election and '0' if the party did not win. Assume that the candidate with the most votes always wins.

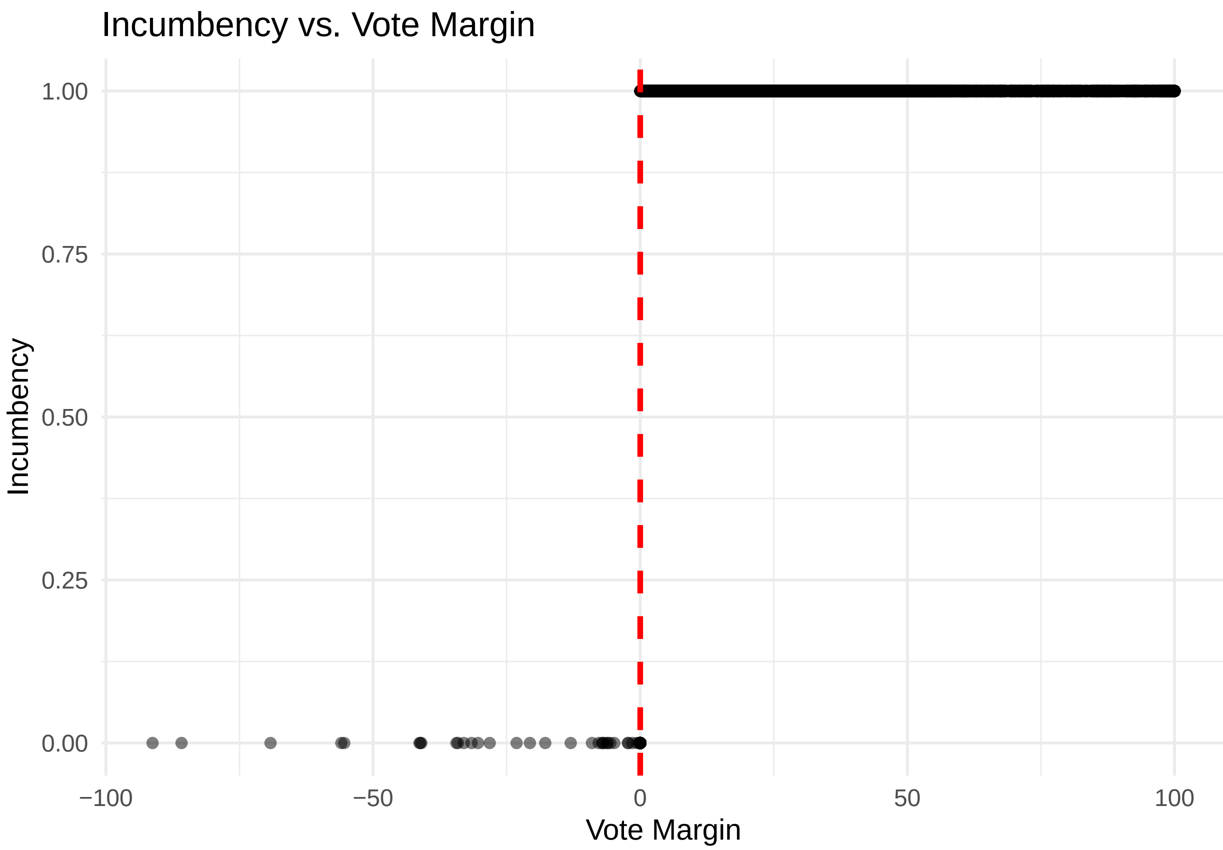
```r
df_lee$share_t <- df_lee$origvote / df_lee$totvote
df_lee$margin_tm1 <- (df_lee$origvote - df_lee$sechighestvote) / (df_lee$origvote + df_lee$sechighestvo
df_lee$incumbent <- ifelse(df_lee$origvote > df_lee$sechighestvote, 1, 0)
```

Test that you constructed the variables correctly by creating a plot with the "treatment" (incumbent) on the y-axis and the forcing variable (margin_tm1) on the x-axis. What kind of RDD is this? Sharp RDD.

```r
# Load necessary libraries
library(ggplot2)

# Scatter plot: Incumbency vs. Previous Vote Margin
ggplot(df_lee, aes(x = margin_tm1, y = incumbent)) +
  geom_point(alpha = 0.5) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(title = "Incumbency vs. Vote Margin",
       x = "Vote Margin",
       y = "Incumbency") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Removed 1612 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



RDD estimates

Now implement the following regression specifications in R , where Y is vote share in election t, X is vote margin in election t-1, and D is incumbency. In each case, report your estimate $\hat{\beta}$ and interpret it with careful reference to the appropriate estimand. For each model, create a scatterplot of X and Y and overlay two fitted curves, one for D=0 and one for D=1. $i.\ Y = \alpha + \beta D + \gamma X + \epsilon$ The result suggests that winning the previous election increases a candidate's vote share in the next election by 17.3 percentage points on average. The LATE for candidates in close elections means the incumbency advantage applies specifically to those who just barely won or lost their previous race, rather than all elections in general. Small SE proves that this effect is precise and unlikely to be due to random chance.

```
library(rdd)
```

```
## Loading required package: Formula
```

```
rdd_1 <- lm(share_t ~ incumbent + margin_tm1, data = df_lee)
robust_rdd1 <- coeftest(rdd_1, vcov = vcovHC(rdd_1, type = "HC2"))

beta_hat <- robust_rdd1[2,1]
se_beta <- robust_rdd1[2,2]

cat("Estimated Incumbency Effect:", round(beta_hat, 3), "\n")
```

```
## Estimated Incumbency Effect: 0.173
```

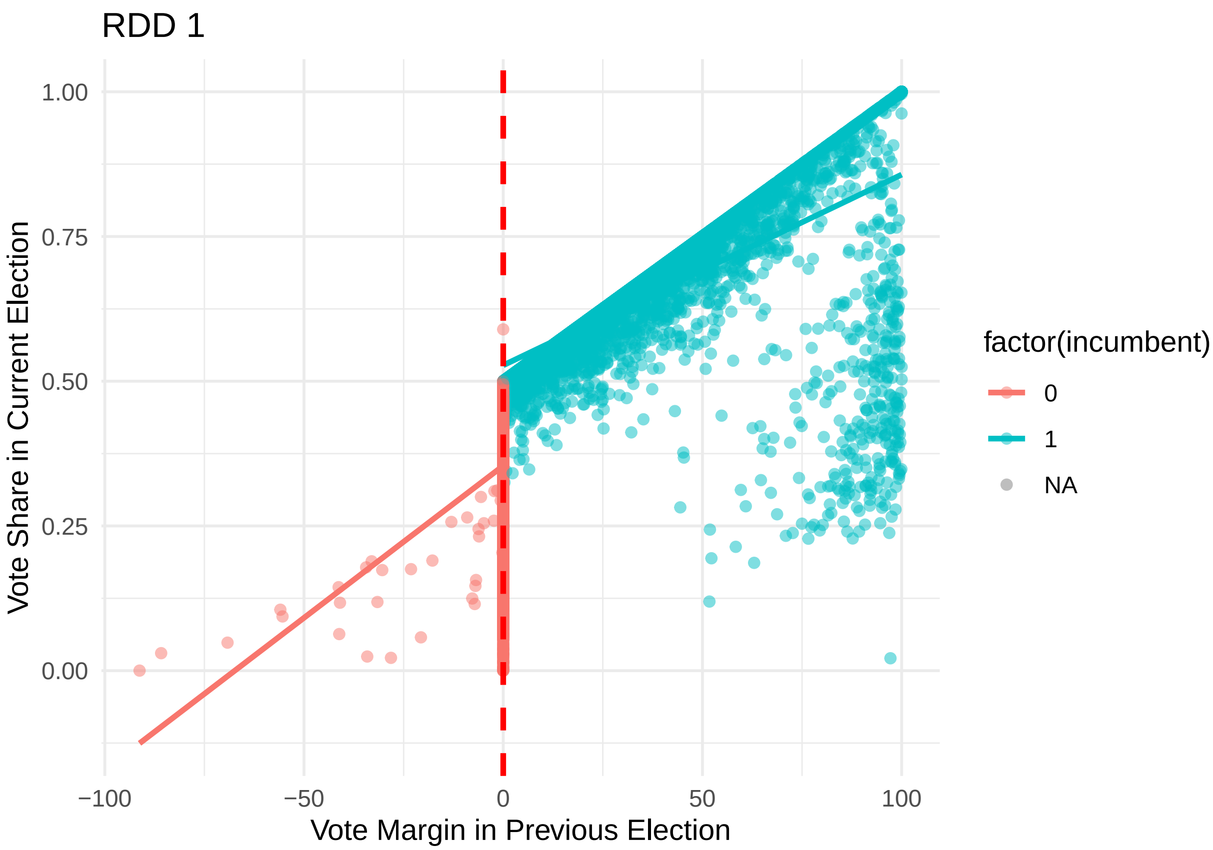
```
cat("SE:", round(se_beta, 3), "\n")
```

```
## SE: 0.002
```

```r
ggplot(df_lee, aes(x = margin_tm1, y = share_t, color = factor(incumbent))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(title = "RDD 1",
       x = "Vote Margin in Previous Election",
       y = "Vote Share in Current Election") +
  theme_minimal()
```

```
## Warning: Removed 1612 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1612 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



ii. $Y = \alpha + \beta D + \gamma X + \delta DX + \epsilon$ Although adding interaction term, beta remains unchanged. This means that the incumbency advantage is fairly stable across different levels of vote margin, at least within the observed range.

```r
rdd_2 <- lm(share_t ~ incumbent * margin_tm1, data = df_lee)
robust_rdd2 <- coeftest(rdd_2, vcov = vcovHC(rdd_2, type = "HC2"))

beta_hat <- robust_rdd2[2,1]
se_beta <- robust_rdd2[2,2]

cat("Estimated Incumbency Effect:", round(beta_hat, 3), "\n")
```

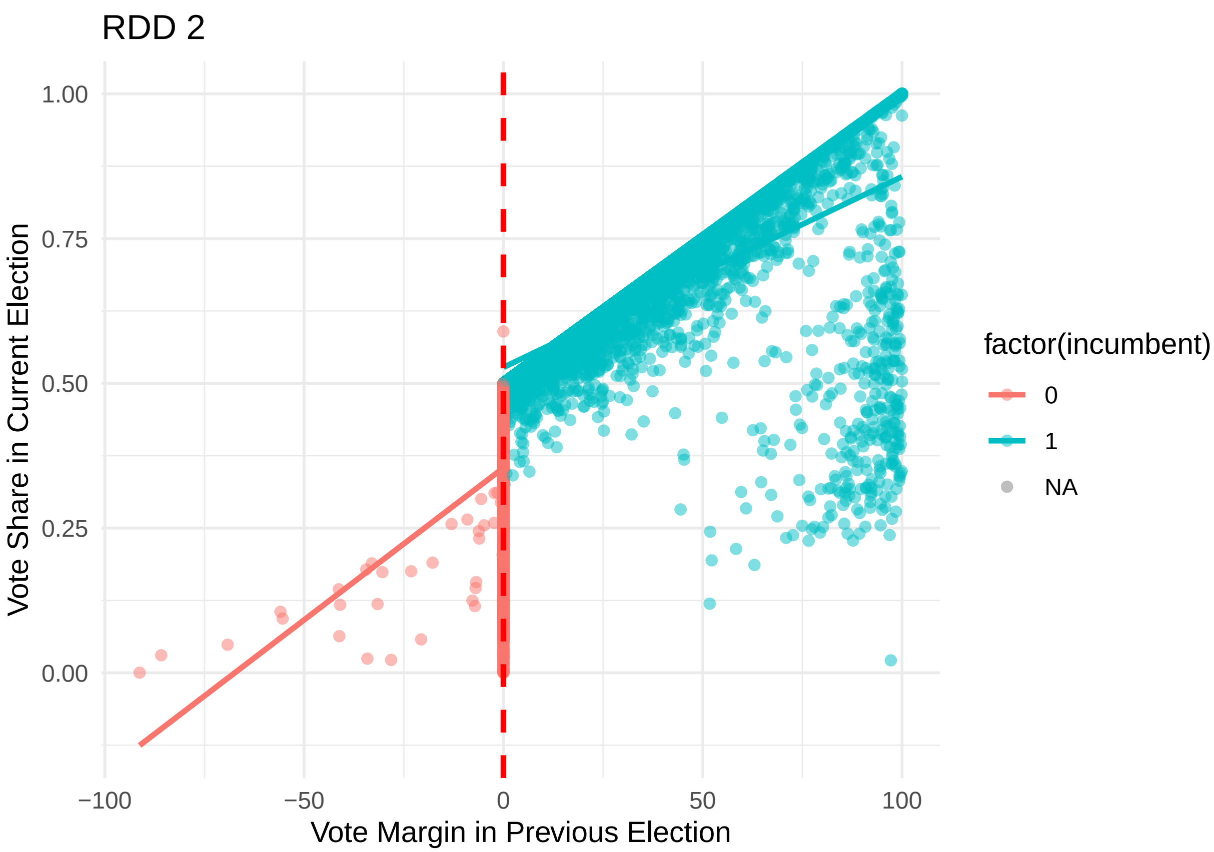
```
## Estimated Incumbency Effect: 0.173
```

```r
cat("SE:", round(se_beta, 3), "\n")
```

```
## SE: 0.002
```

```r
ggplot(df_lee, aes(x = margin_tm1, y = share_t, color = factor(incumbent))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(title = "RDD 2",
       x = "Vote Margin in Previous Election",
       y = "Vote Share in Current Election") +
  theme_minimal()
```

```
## Warning: Removed 1612 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 1612 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



iii. $Y = \alpha + \beta D + \gamma X + \gamma_2 X^2 + \delta DX + \delta_2 DX^2 + \epsilon$ This model accounts for varied effect of incumbency at different parts of the margin distribution. The beta value is slightly lower than previous models, suggesting that the linear models may have overestimated the incumbency advantage by not accounting for this nonlinearity.

```r
rdd_3 <- lm(share_t ~ incumbent * margin_tm1 + I(margin_tm1^2) + I(incumbent * margin_tm1^2), data = df_
robust_rdd3 <- coeftest(rdd_3, vcov = vcovHC(rdd_3, type = "HC2"))

beta_hat <- robust_rdd3[2,1]
se_beta <- robust_rdd3[2,2]
```

```r
cat("Estimated Incumbency Effect:", round(beta_hat, 3), "\n")
```

```
## Estimated Incumbency Effect: 0.121
```

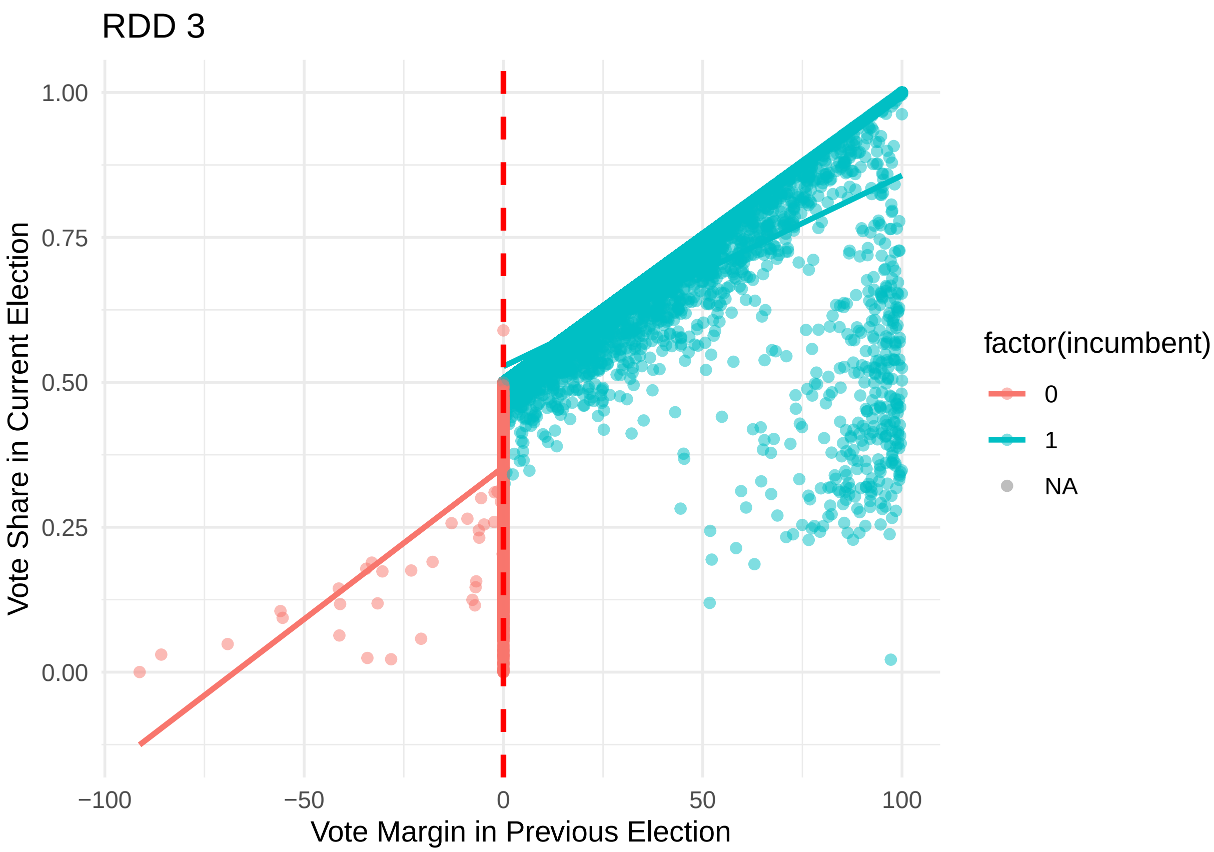
```r
cat("SE:", round(se_beta, 3), "\n")
```

```
## SE: 0.002
```

```r
ggplot(df_lee, aes(x = margin_tm1, y = share_t, color = factor(incumbent))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(title = "RDD 3",
       x = "Vote Margin in Previous Election",
       y = "Vote Share in Current Election") +
  theme_minimal()
```

```
## Warning: Removed 1612 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1612 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



iv. **A local linear regression with a triangular kernel. Local linear regression works by fitting a straight line in a local data space, defined by some point $X = X_0$ and a bandwidth around X_0. The value of $\hat{Y}(X_0)$ at point $X = X_0$ is then evaluated, and the process is repeated for each X_0. The result is a smoothed conditional expectation function $E[Y \mid X]$. To implement this in `R`, use either the package `rdd`, and choose the Imbens-Kalyanamaran optimal bandwidth *or* the Calonico, Cattaneo, and Titiunik (CCT) optimal bandwidth from the `rdrobust` package for the**

bandwidth. Report the estimate of β and the optimal bandwidth. This model focuses only on close elections. Its beta is lower than in the parametric models, suggesting that parametric models may have overestimated the incumbency advantage

```
library(rdrobust)
rdd_result <- rdrobust(y = df_lee$share_t, x = df_lee$margin_tm1, kernel = "triangular")
```

```
## Warning in rdrobust(y = df_lee$share_t, x = df_lee$margin_tm1, kernel =
## "triangular"): Mass points detected in the running variable.
```

```
beta_estimate <- rdd_result$coef[1]
optimal_bandwidth <- rdd_result$bws[1]

cat("Beta Estimae:", beta_estimate, "\n")
```

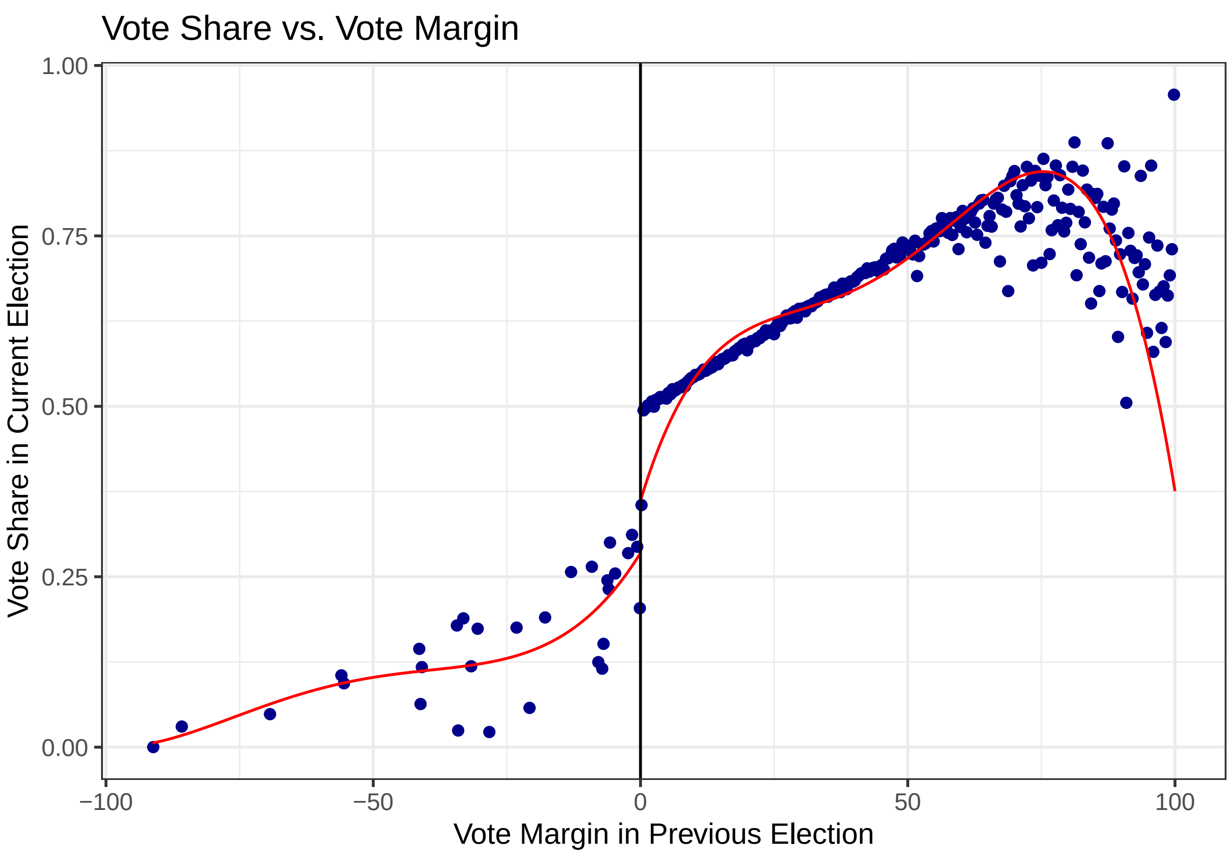
```
## Beta Estimae: 0.07296903
```

```
cat("Optimal Bandwidth:", optimal_bandwidth, "\n")
```

```
## Optimal Bandwidth: 13.56322
```

```
rdplot(df_lee$share_t, df_lee$margin_tm1, c = 0, binselect = "esmv", kernel = "triangular",
       title = "Vote Share vs. Vote Margin",
       x.label = "Vote Margin in Previous Election",
       y.label = "Vote Share in Current Election")
```

```
## [1] "Mass points detected in the running variable."
```



Do your results depend on the functional form of the regression? Why? Yes, the results depend on the regression model used because different models make different assumptions about how vote margin affects vote share. Adding interaction terms lets incumbency effects change based on how close the previous election was. OLS looks at the whole dataset, which can introduce bias if the relationship isn't the same everywhere. Local regression RDD focuses only on close elections, where the effect of incumbency is more credible. If results change a lot between models, it means the choice of regression matters, and we should check for nonlinearity to avoid misleading conclusions.

Robustness

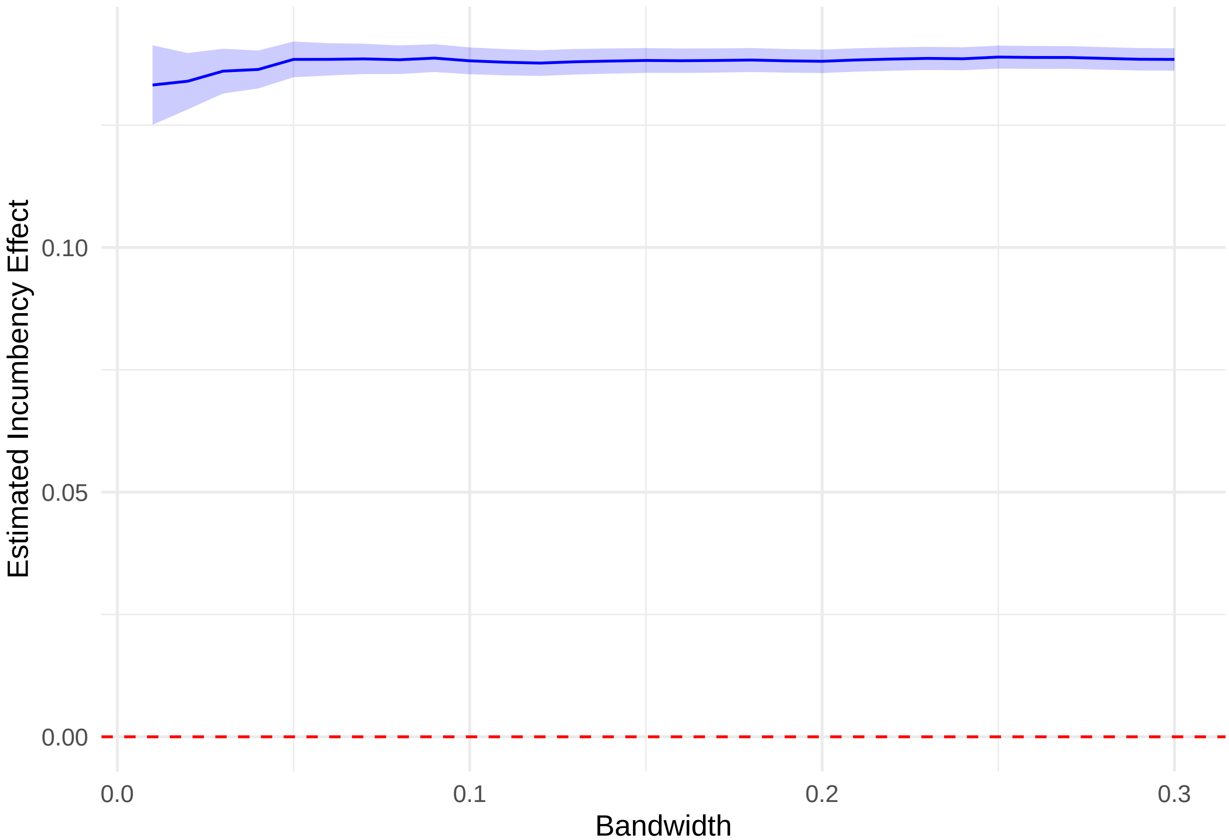
For most of the previous section, you used the whole dataset to fit the model, with the exception of(iv), where you used an optimal bandwidth chosen by an algorithm. Now let's see if the results are robust to different bandwidths around the discontinuity. Use bandwidth sizes from 0.01 to 0.3, in increments of 0.01. For each bandwith, trim the data on either side of the threshold and fit the model from (ii) on the trimmed dataset. Plot the coefficients for all bandwidth sizes with 95% confidence intervals. What do you conclude about the robustness of the results? The estimated effect remains consistently around 0.13–0.14 across all bandwidth sizes. So, the incumbency advantage is not driven by specific bandwidth choices. Although a wider CI presented at very small bandwidths due to fewer observations, the overall window remain tight and do not cross zero, further proved vallidate the incumbency effect.

```r
bandwidths <- seq(0.01, 0.3, by = 0.01)

beta_estimates <- c()
lower_ci <- c()
upper_ci <- c()

for (bw in bandwidths) {

  trimmed_data <- df_lee %>% filter(abs(margin_tm1) <= bw * 100)

  rdd_model <- lm(share_t ~ incumbent * margin_tm1, data = trimmed_data)
  robust_se <- coeftest(rdd_model, vcov = vcovHC(rdd_model, type = "HC2"))

  beta_hat <- robust_se[2,1]
  beta_se <- robust_se[2,2]

  beta_estimates <- c(beta_estimates, beta_hat)
  lower_ci <- c(lower_ci, beta_hat - 1.96 * beta_se)
  upper_ci <- c(upper_ci, beta_hat + 1.96 * beta_se)
}

results_df <- data.frame(
  Bandwidth = bandwidths,
  Beta = beta_estimates,
  Lower_CI = lower_ci,
  Upper_CI = upper_ci
)

results_df
```

```
##    Bandwidth      Beta  Lower_CI  Upper_CI
## 1       0.01 0.1331938 0.1250773 0.1413103
## 2       0.02 0.1339744 0.1282178 0.1397310
## 3       0.03 0.1360285 0.1314511 0.1406060
## 4       0.04 0.1363731 0.1324935 0.1402528
## 5       0.05 0.1384270 0.1347755 0.1420784
```

```
## 6         0.06 0.1384385 0.1351425 0.1417344
## 7         0.07 0.1385305 0.1354324 0.1416287
## 8         0.08 0.1383578 0.1354289 0.1412867
## 9         0.09 0.1386914 0.1358469 0.1415358
## 10        0.10 0.1381383 0.1353977 0.1408789
## 11        0.11 0.1378499 0.1351685 0.1405313
## 12        0.12 0.1376750 0.1350445 0.1403055
## 13        0.13 0.1379509 0.1353423 0.1405595
## 14        0.14 0.1380926 0.1355218 0.1406633
## 15        0.15 0.1382092 0.1356756 0.1407427
## 16        0.16 0.1381569 0.1356590 0.1406549
## 17        0.17 0.1382097 0.1357365 0.1406829
## 18        0.18 0.1382963 0.1358509 0.1407417
## 19        0.19 0.1381353 0.1357185 0.1405522
## 20        0.20 0.1380421 0.1356495 0.1404348
## 21        0.21 0.1383206 0.1359313 0.1407100
## 22        0.22 0.1384993 0.1361214 0.1408772
## 23        0.23 0.1386343 0.1362691 0.1409996
## 24        0.24 0.1385585 0.1362086 0.1409084
## 25        0.25 0.1389008 0.1365456 0.1412560
## 26        0.26 0.1388284 0.1364892 0.1411675
## 27        0.27 0.1388223 0.1364977 0.1411469
## 28        0.28 0.1386355 0.1363292 0.1409418
## 29        0.29 0.1384557 0.1361646 0.1407469
## 30        0.30 0.1384308 0.1361455 0.1407161
```

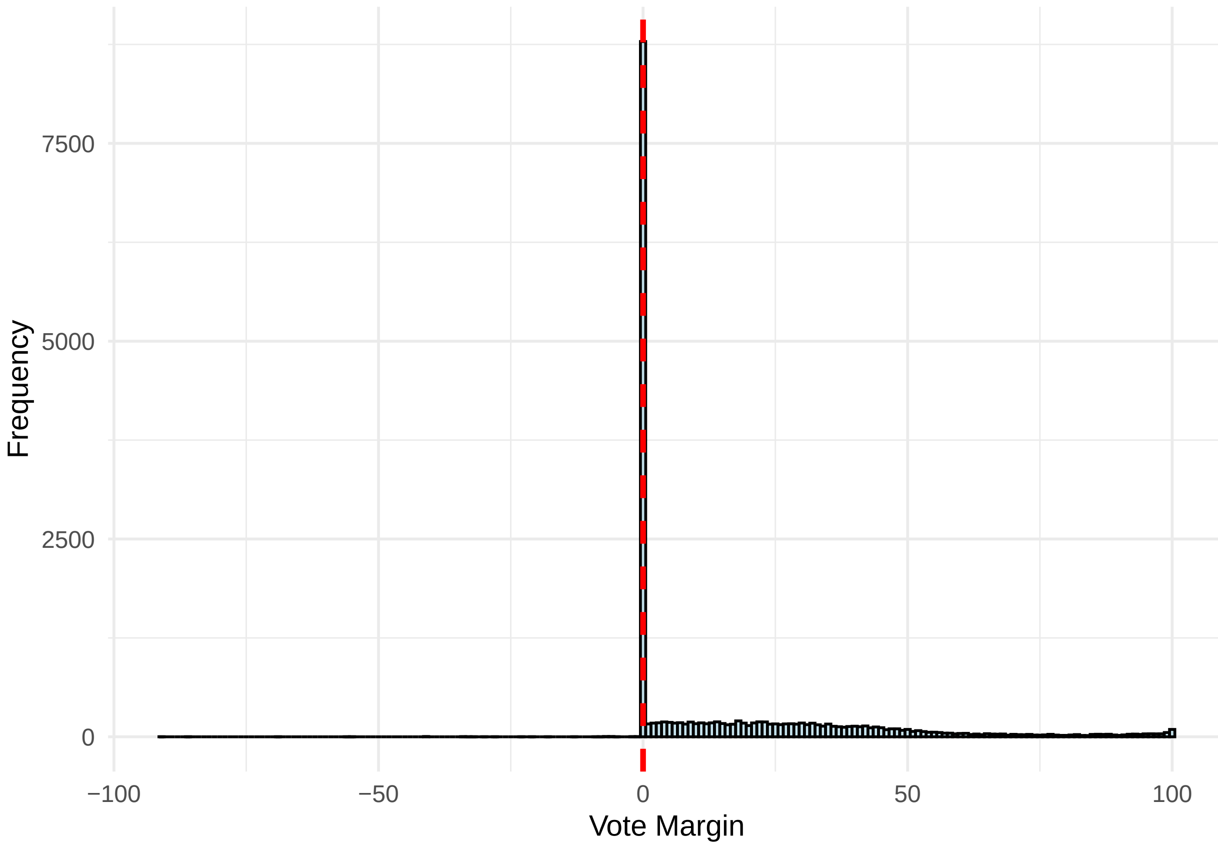
```r
ggplot(results_df, aes(x = Bandwidth, y = Beta)) +
  geom_line(color = "blue") +
  geom_ribbon(aes(ymin = Lower_CI, ymax = Upper_CI), alpha = 0.2, fill = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Bandwidth",
       y = "Estimated Incumbency Effect") +
  theme_minimal()
```



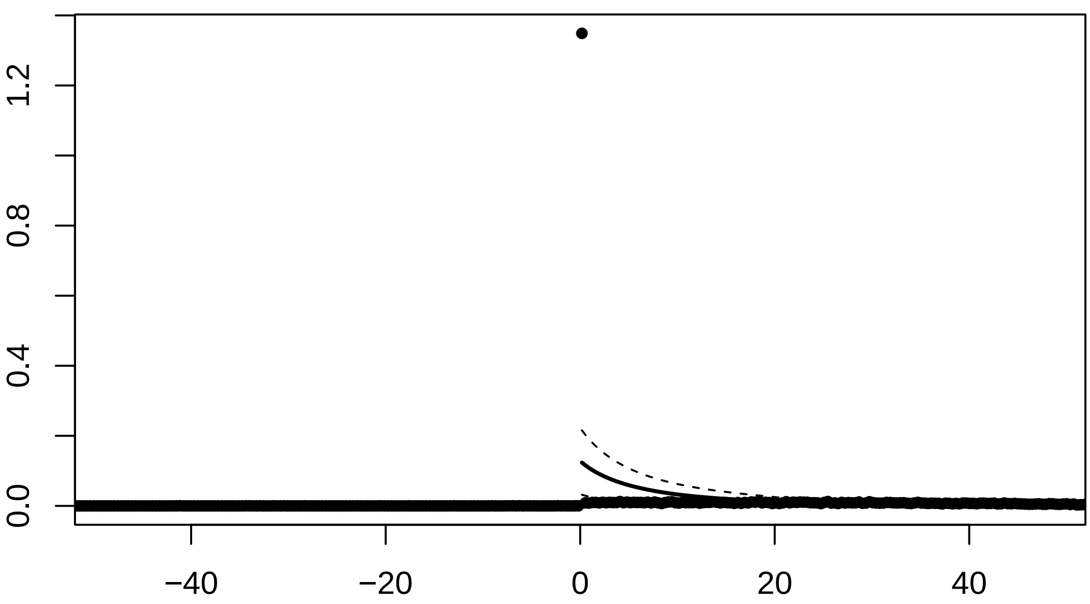
Assess the plausibility of the identification assumption for sharp RDD in this application by examining the density of the forcing variable around the cutoff. First, create a histogram of the forcing variable using bins of one percentage point. Second, conduct a formal test of the difference in density around the cutoff using the DCdensity() function in the rdd package and report the value from the test. Why is this analysis a good diagnostic for assessing the assumption? What can you say about the plausibility of the assumption in this case? From the histogram, we've seen that it's not smooth around zero but rather an extreme spike at the cutoff. The McCrary test validated that there's manipulation or sorting at the cutoff. The p-value is very small, meaning there's a significant discontinuity in the distribution of vote margins at 0%, and confirming that the RDD identification assumption is violated.

```
ggplot(df_lee, aes(x = margin_tm1)) +
  geom_histogram(binwidth = 1, color = "black", fill = "lightblue", alpha = 0.7) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", size = 1) +
  labs(x = "Vote Margin",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Removed 1612 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
DCdensity(df_lee$margin_tm1, cutpoint = 0)
```




```
## [1] 1.471312e-97
```

What does the RDD identification assumption have to say about how the officeexp variable should look near the threshold? You do not need to actually implement this test. Hypothetically, if an observed covariate failed to behave as expected, how would that the interpretation of your results be affected? Would your results necessarily be invalidated? The RDD assumption predicts that officeexp should be continuous at the cutoff. If it's not and rather it jumps at the cutoff, this suggests that incumbents and non-incumbents were already different before treatment, which could indicate sorting, manipulation, or omiunobserved variable bias. This would raise concerns about whether incumbency is truly random near the threshold. If the imbalance is small and explainable, results may still be valid but require additional robustness checks.