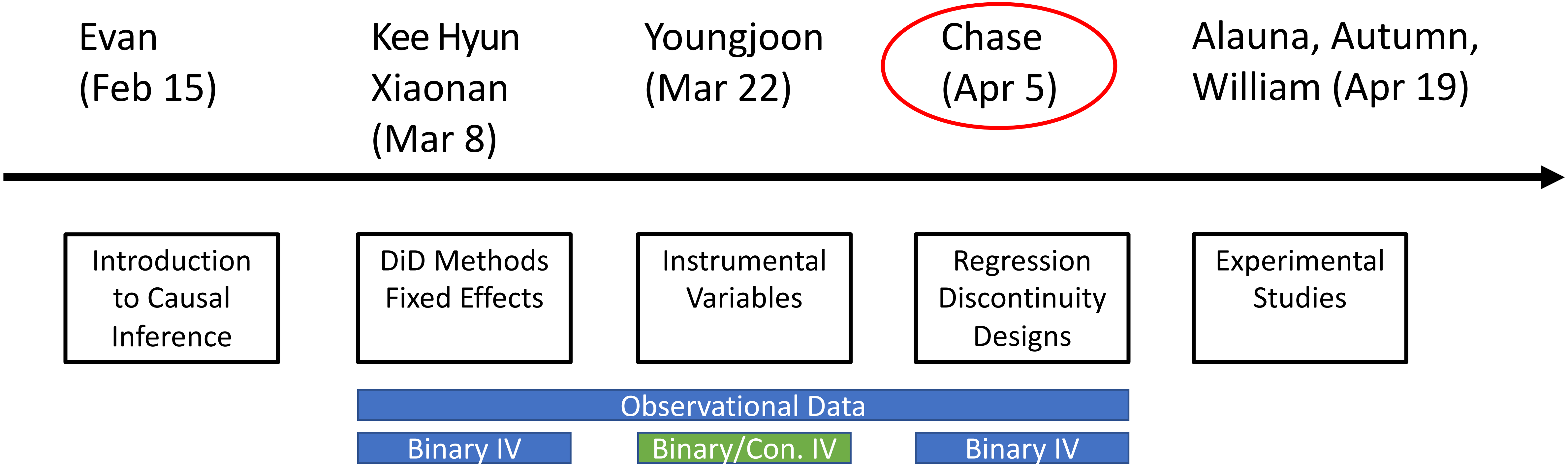
Regression Discontinuity Designs

April 5, 2021

Week 4: Regression Discontinuity Design



Evan (Feb 15)

Kee Hyun Xiaonan (Mar 8)

Youngjoon (Mar 22)

Chase (Apr 5)

Alauna, Autumn, William (Apr 19)

Introduction to Causal Inference

DiD Methods Fixed Effects

Instrumental Variables

Regression Discontinuity Designs

Experimental Studies

Observational Data

Binary IV

Binary/Con. IV

Binary IV

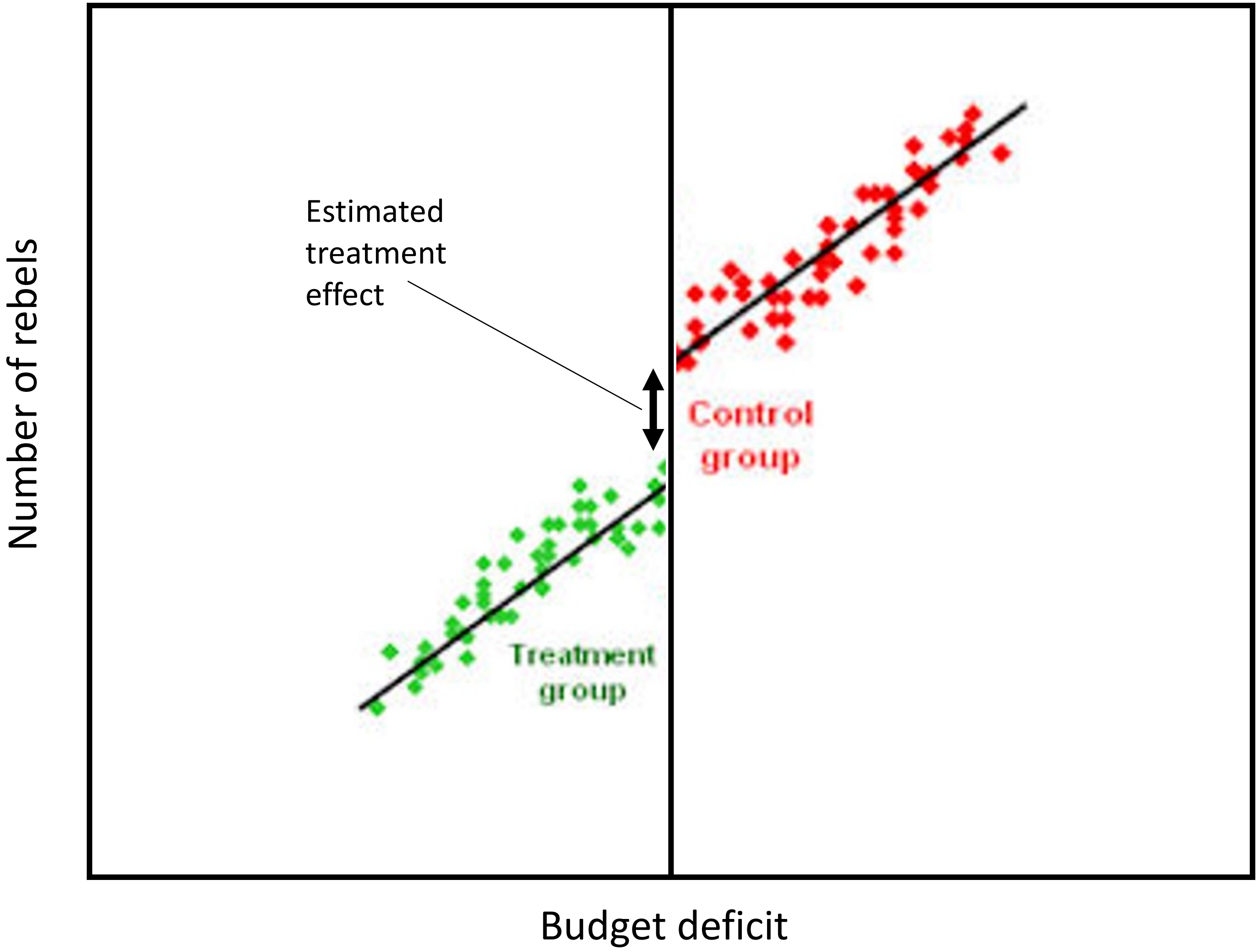
Agenda

1. Introduction to RDD concept
2. When to use RDD
3. Potential issues with RDD
4. Literature examples
5. Example code
6. Further learning

What is RDD?

- RDD uses two samples, one subject to a certain treatment and the other not, to estimate two regression lines and then compares the difference
  - Functions like a quasi experiment
  - The "cutpoint" between the treatment and control groups needs to be random or "as-if" random
  - The method examines a given bandwidth around the cutpoints
  - After a regression line is fit to each group, measure the difference in the intercepts at the cutpoint
    - This is your estimated treatment effect
- This method is most commonly used to examine close elections, but can have a variety of applications

What is RDD?



An IR Example:

Having a budget surplus decreases the number of rebel fighters in conflict states, because these states are more able to acquire funds to pay soldiers.

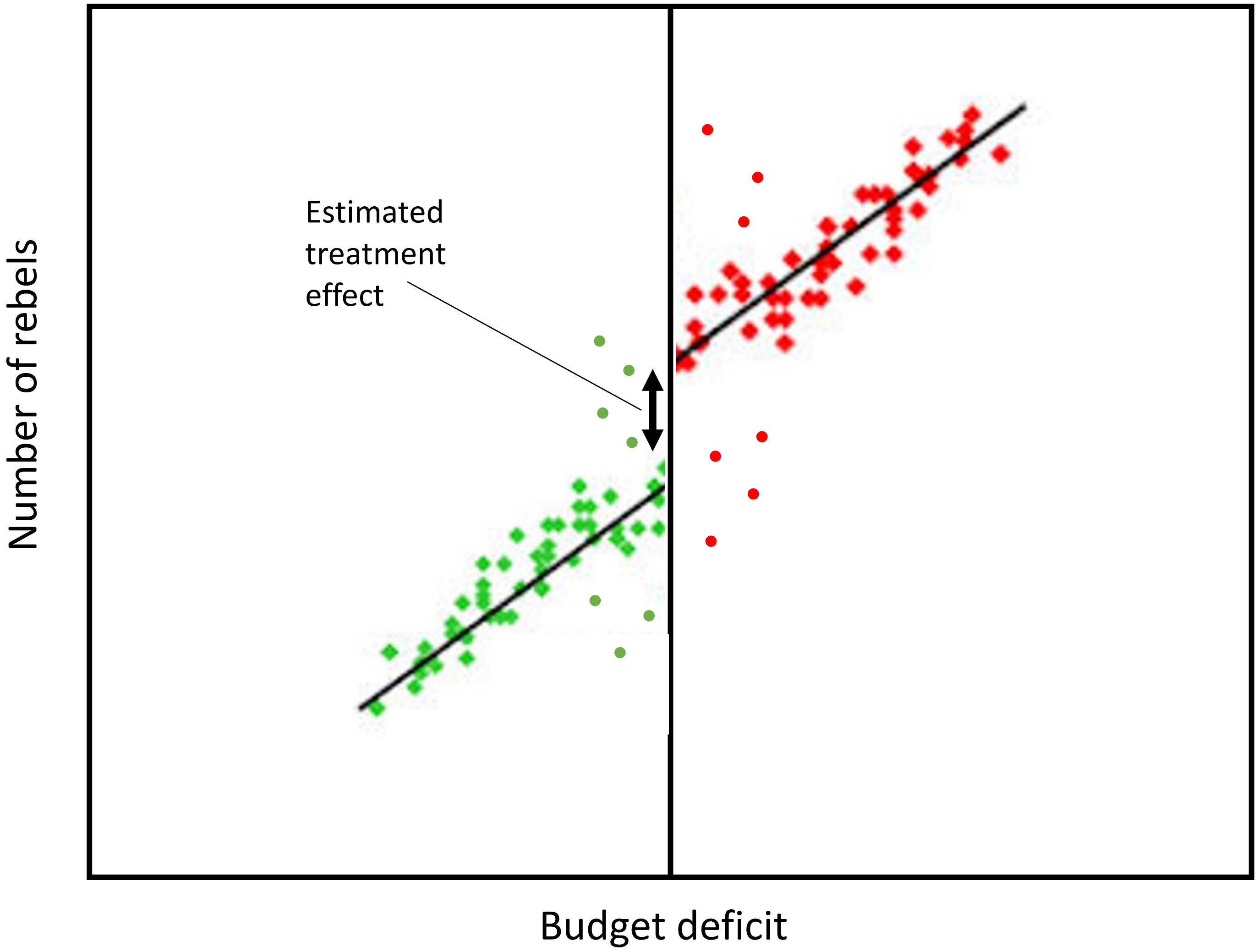
‘Surplus’ is the statistical ‘treatment’

Budget deficit (0 or 1) acts as the cutpoint.

The amount of deficit (or surplus) acts as the x-axis, the number of rebels acts as the y-axis.

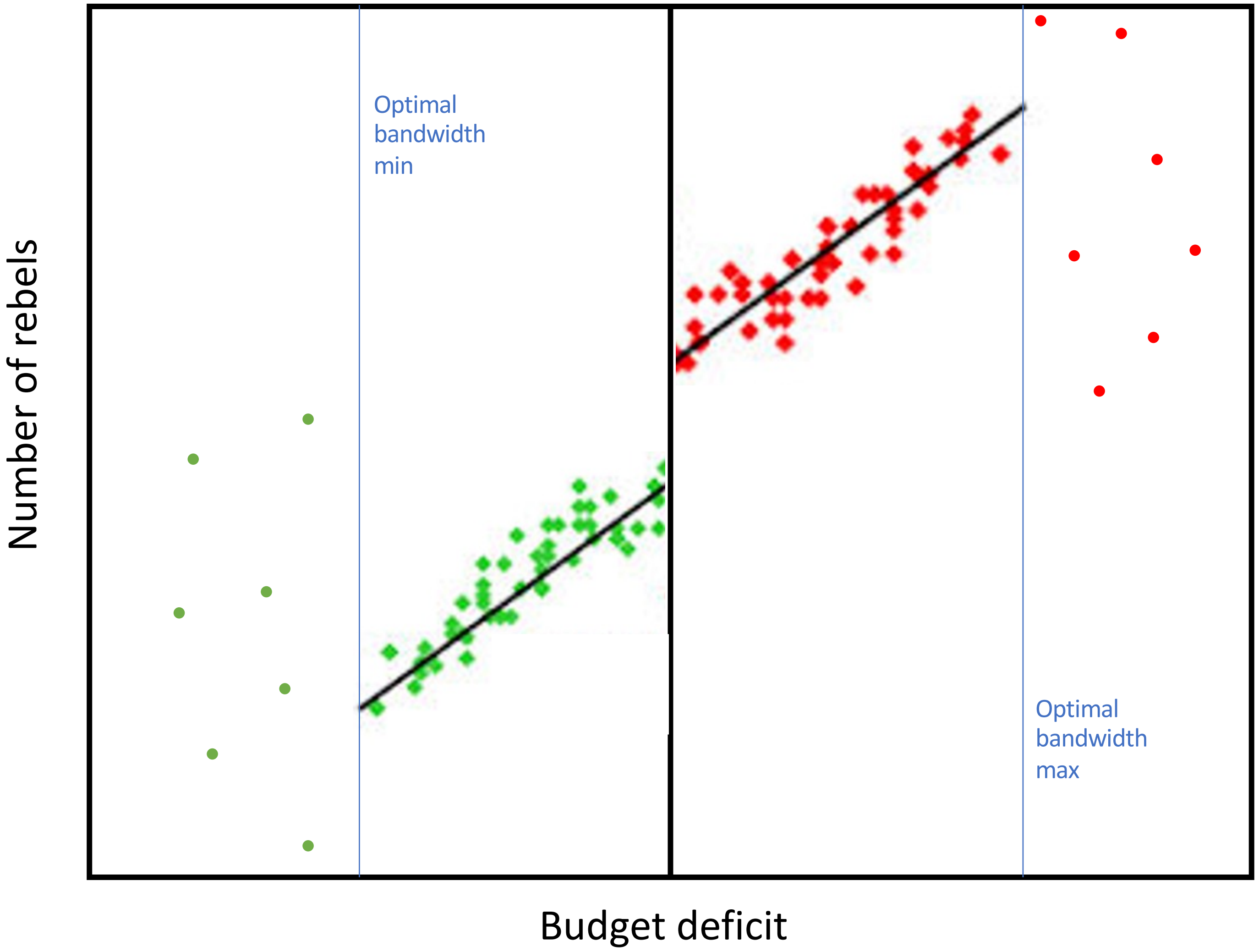
The green points are states with budget surpluses (negative deficits). The red points are states with budget deficits.

What is RDD?



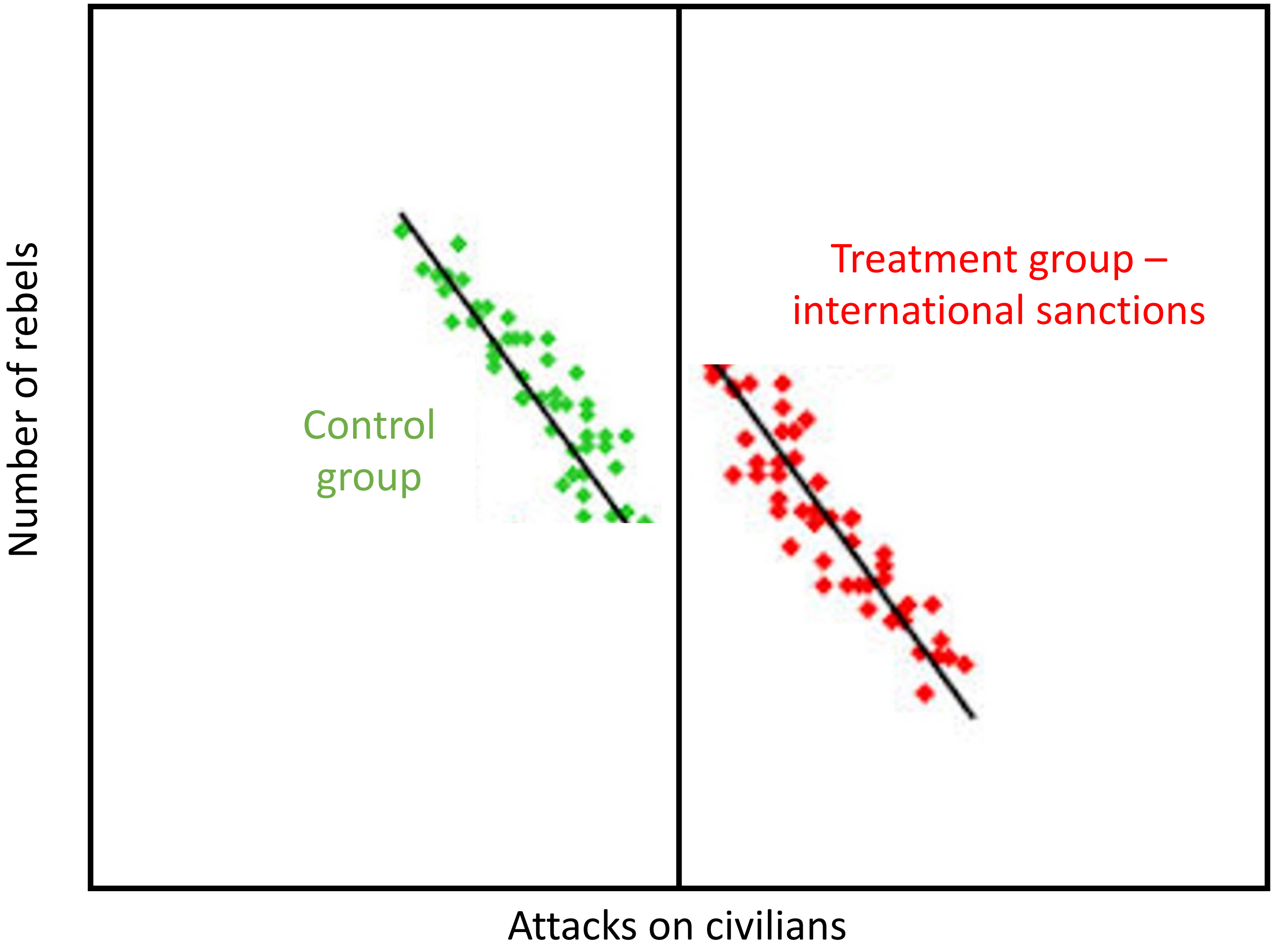
Some points from each group may exist above or below the regression mean for the alternate group; the method is comparing the difference in regression means.

What is RDD?



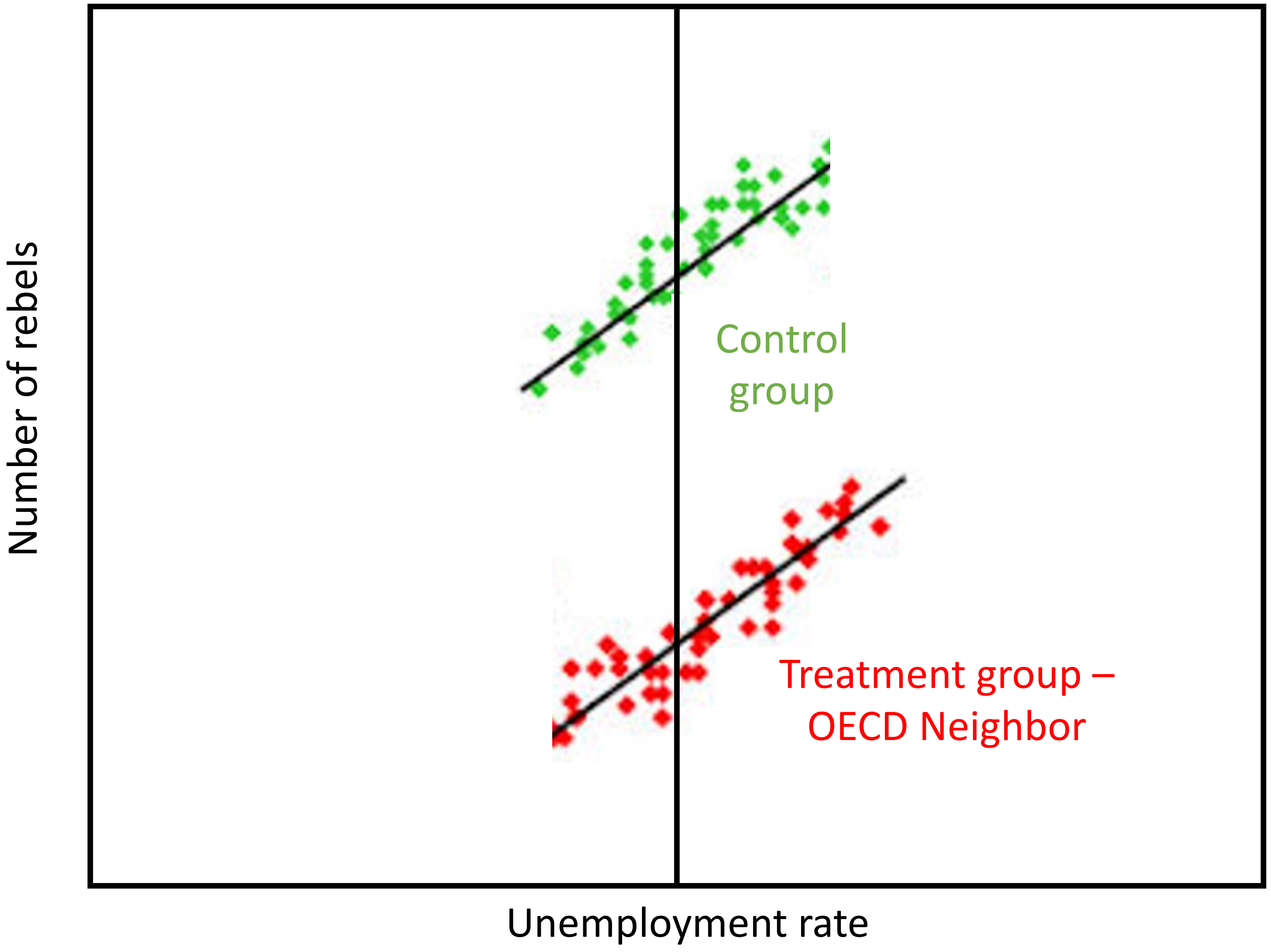
Estimation is occurring within the “optimal bandwidth”, which may not include all values from each group.

What is RDD?



- RDD curves can take many forms
- For example;
  - Treatment group may be on the right
  - The curve could also slope negatively, or be flat
    - Many theoretical base models will be flat
    - Flat RDD is “sharp”, sloped is “fuzzy”
  - The treatment variable may differ from the x-axis variable
    - In some cases, the treatment effect may weaken the related independent variable effect
      - For example, shown here; states that attack civilians are shown on a scale, with the implementation of sanctions for this behavior being the cutpoint, and the x-axis showing the number of attacks
      - This setup is less common for observational data, difficult to establish clean cutpoints

What is RDD?

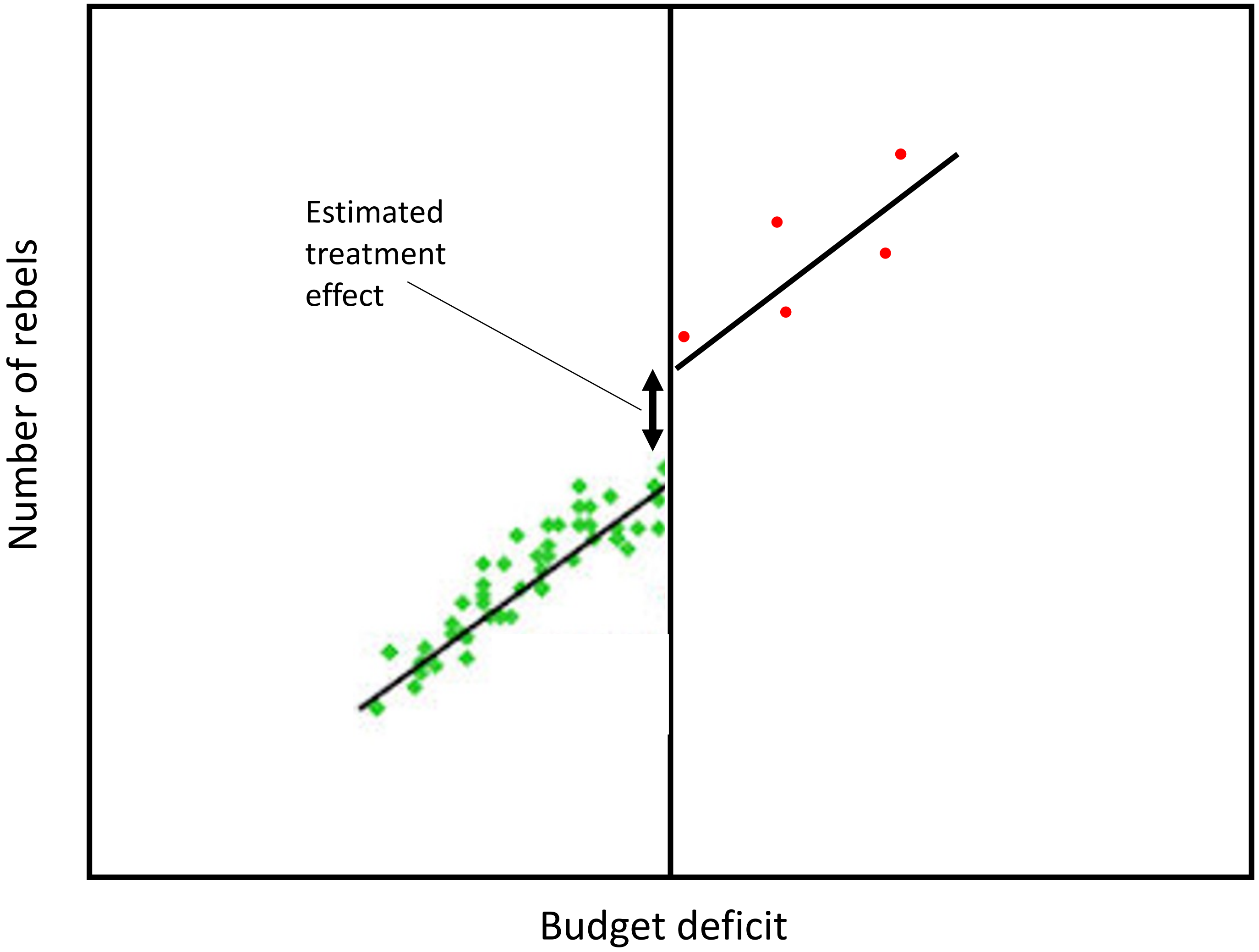


- Another case may use a treatment effect not directly related to the other independent variable (“covariate”)
  - For example, we may use geographic region as the treatment effect
  - In this case, the method is essentially the same as estimating a binary variable coefficient into a model, and then plotting this model in comparison with a model with the variable excluded
    - The primary difference is the “randomness” of the treatment

When should we use RDD?

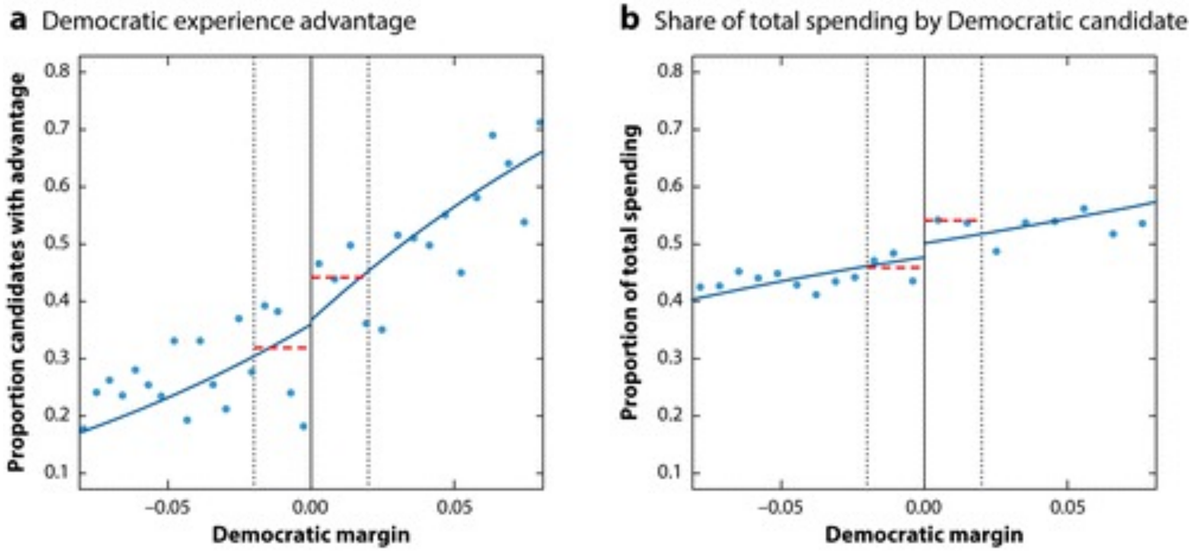
- RDD is most appropriate with a binary independent variable
- Even better when the variable can alternately exist in ‘degrees’ of some kind
  - For example, in election results: win/lose, margin of victory/loss
- RDD is problematic if we are selecting the binary cutoff for a continuous variable
  - For example, unemployment: 10% unemployment is cutpoint
  - This violates the underlying logic of an RDD argument (a binary ‘treatment’) and is also susceptible to selection bias
  - However, it could work in the case of something with a well-accepted “cutoff”, such as a 620 FICO score

When should we use RDD?



- The distribution of cases may limit the effectiveness or statistical significance of this method
  - For example, for very rare events such as conflict or coup
  - In practice, if the cut-off selects only a small number, then the control group may also be constrained. In the same way, using experimental data, more subjects may be sought for the control group if the treatment group takes a significant majority. However, this is usually not possible for observational data.

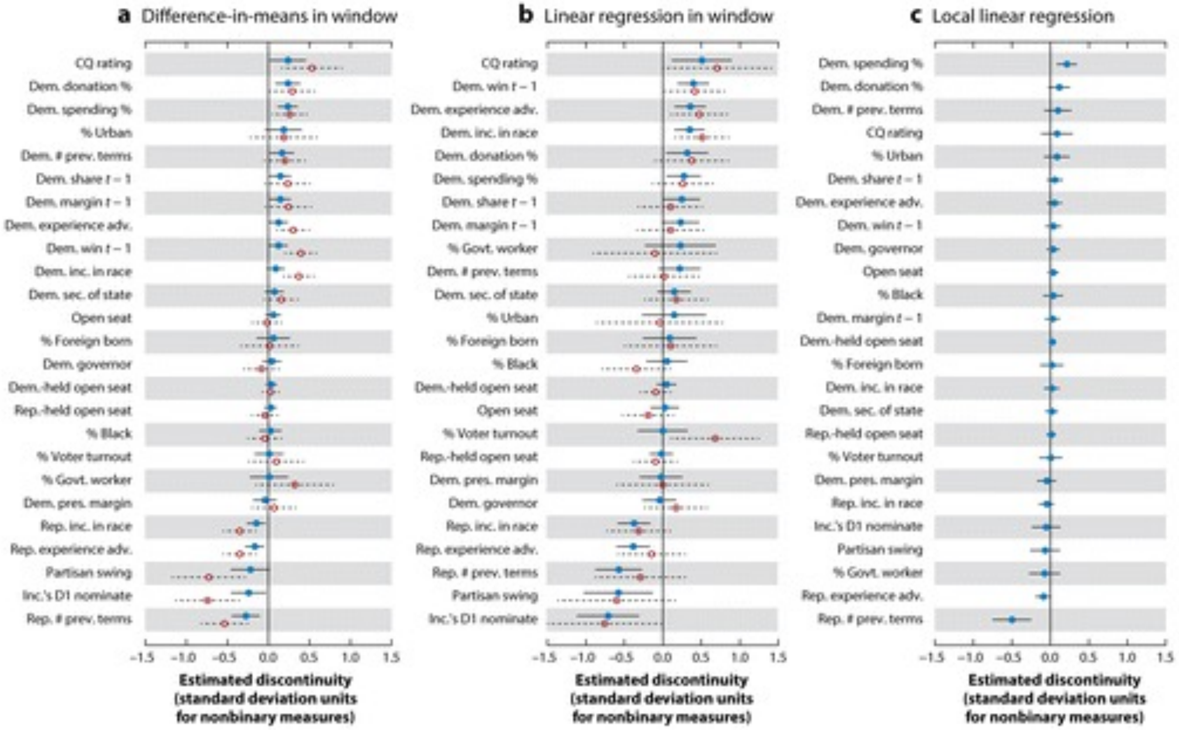
Potential Issues with RDD: Bandwidth selection



de la Cuesta B, Imai K. 2016.
Annu. Rev. Polit. Sci. 19:375–96

- Local randomization assumption: Under the local randomization assumption, also called the as-if-random assumption, the observations below and above the discontinuity threshold, a [−0.02, 0.02] window indicated by dotted lines in this case, are assumed to be identical on average. As a result, the estimated discontinuity is based on two flat lines with no slope (red dashed lines). In contrast, under the continuity assumption, the association with the forcing variable is not assumed to be absent (blue solid lines). The two plots are based on the dataset on US House elections by Caughey & Sekhon (2011) using two pretreatment covariates: the experience advantage of the Democratic candidate (a) and the proportion of total donations given to the Democrat (b). They show that the local randomization assumption can falsely discover a discontinuity (a) or overestimate one (b).
- Solution: test multiple bandwidths

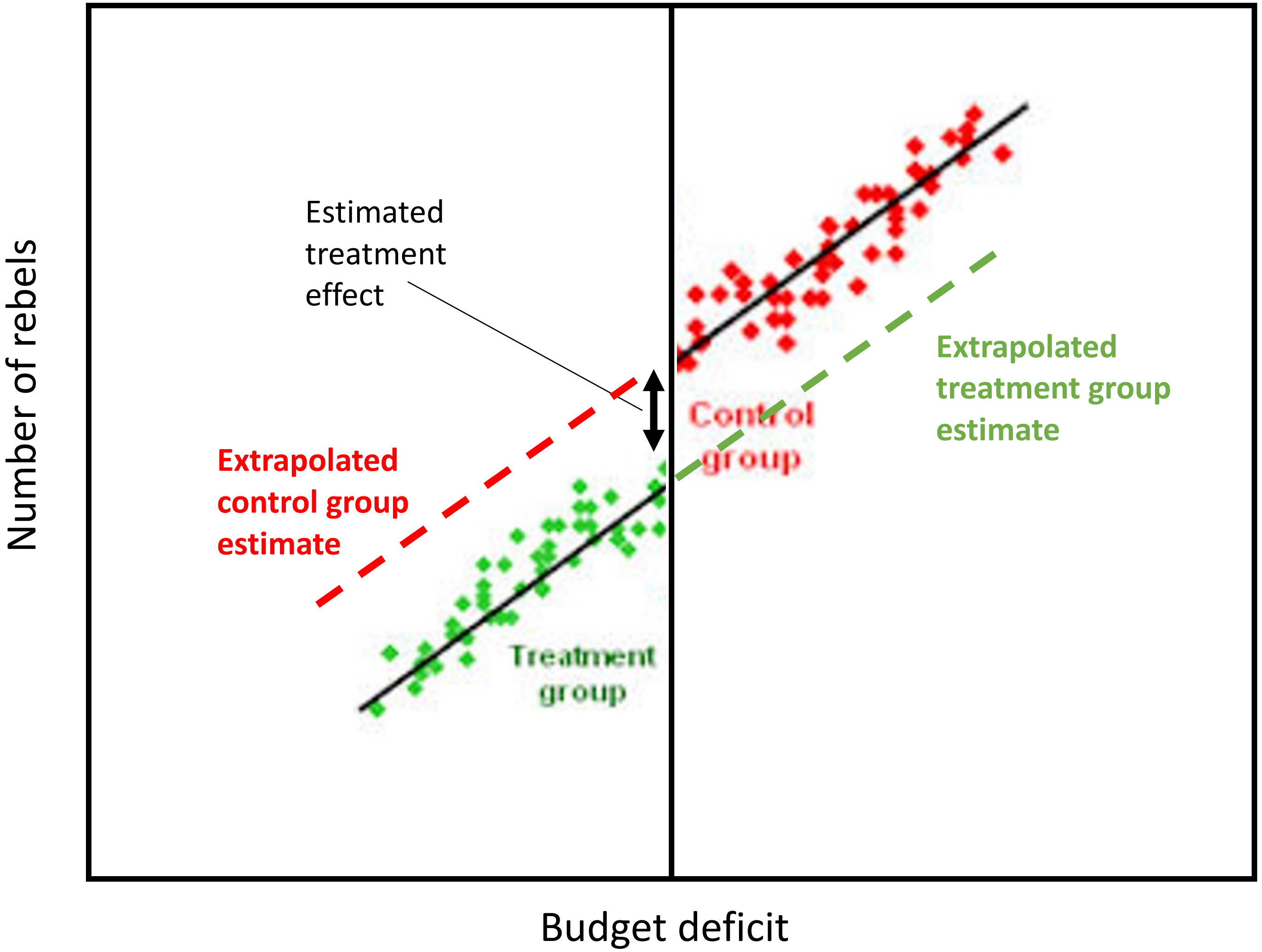
Potential Issues with RDD: Line estimation method



AR de la Cuesta B, Imai K. 2016.
Annu. Rev. Polit. Sci. 19:375–96

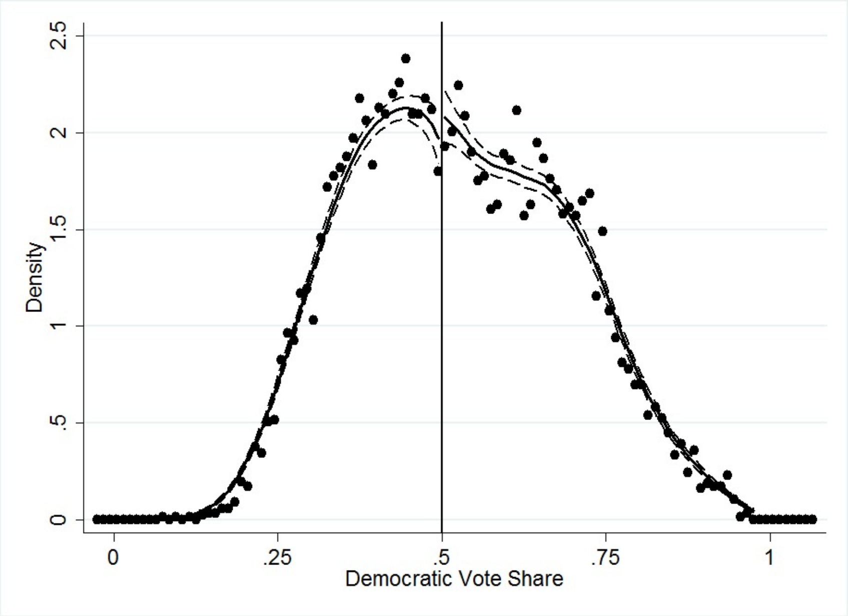
- Figure shows a comparison of estimated discontinuities in pretreatment covariates across three methods. Solid and dashed lines in each panel represent 95% confidence intervals, not corrected for multiplicity. (a) Filled blue circles represent estimates based on the difference-in-means estimator within the 2-percentage-point window on either side of the threshold; open red circles represent estimates within the one-half-percentage-point window. Panel (b) shows the estimates based on the linear regression in the same sets of windows. Panel (c) presents the estimates based on the local linear regression proposed by Calonico et al. (2014). Abbreviations: adv., advantage; CQ, Congressional Quarterly; Dem., Democratic; govt., government; inc., incumbent; pres., president; prev., previous; Rep., Republican; sec., secretary; t, time period.
- Solution: test multiple methods

Potential Issues with RDD



- External validity and extrapolation: The design is inherently extrapolating the treatment effect along the whole line, although the estimate is being drawn only from the cutpoint
  - The effect could differ at other points in the line
- Including covariates in the model can also complicate this issue
- Solution: Include several robustness tests, estimate full lines if possible

Potential Issues with RDD: Distortion at cut-point

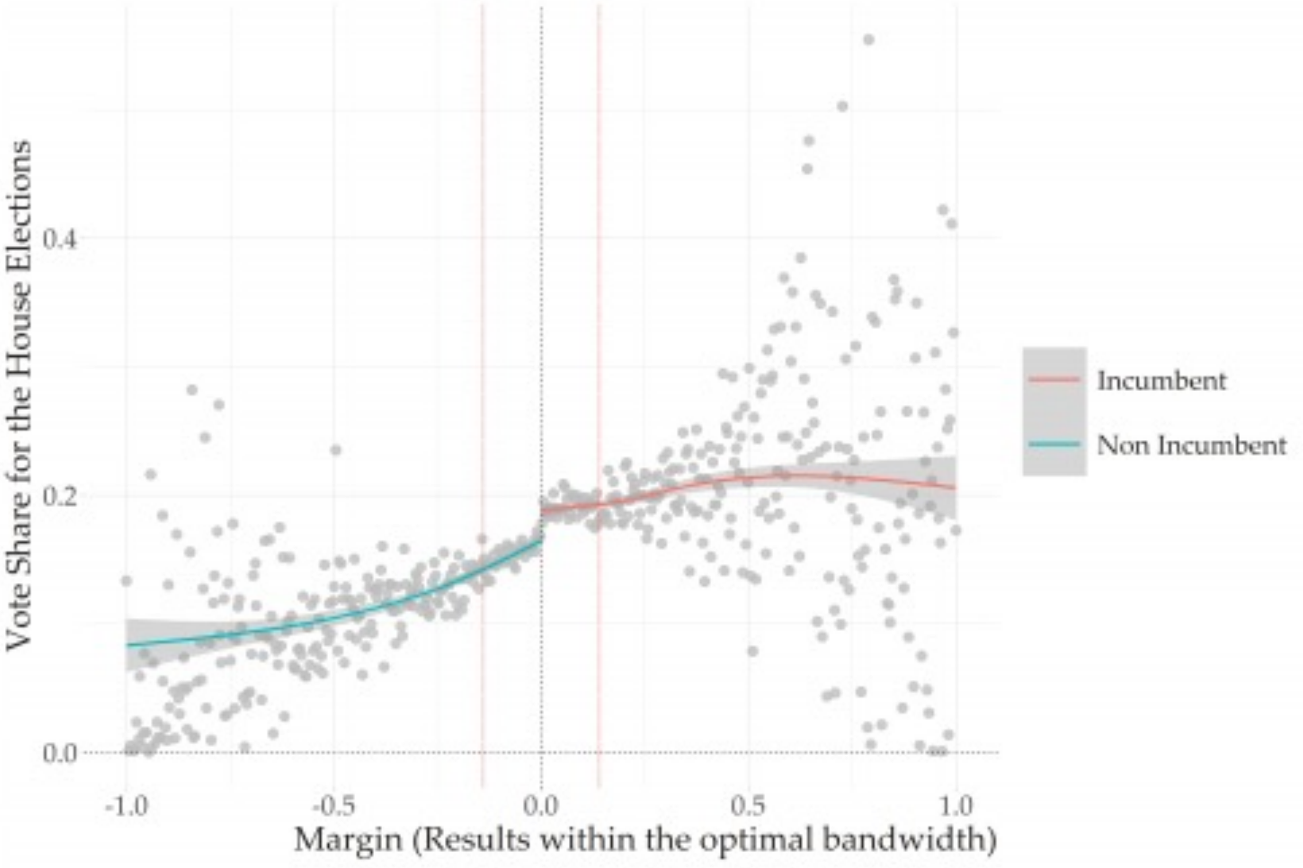


- Distribution density
  - Example: In State X, close elections always result in a recount, and the incumbent party usually wins (i.e. is able to manipulate) the recount
  - Issue: This can have implications for our theory, as the candidates in close elections are manipulating whether they are in the treatment group or not
  - Solution: Observation density test
    - See McCrary (2008)

Literature Example

- Ventura, T. (2021). Do mayors matter? Reverse coattails on congressional elections in Brazil. *Electoral Studies*, *69*, 102242.
- Provides an example of the “close election” genre
- Studies the effect of having an incumbent mayor on a parties’ regional vote performance
  - The effect is hypothesized to be positive

Literature Example: Primary figure



- The y-axis displays the dependent variable, party’s regional vote share
- The x-axis displays the independent variable, party’s mayoral election performance, with the cutpoint being win or loss and the scale showing margin of victory/defeat

Literature Example: Primary table

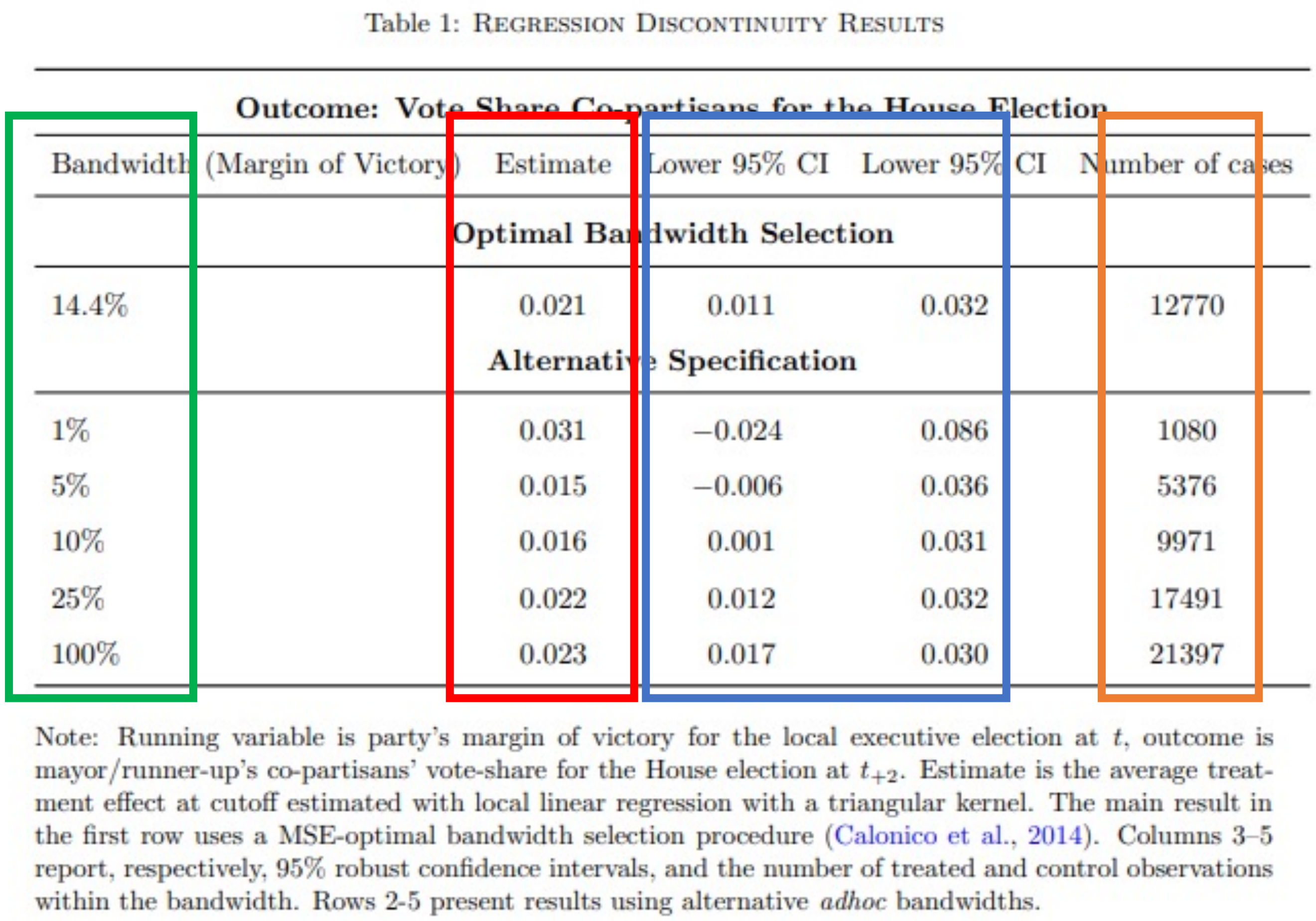


Table 1: Regression Discontinuity Results

Outcome: Vote Share Co-partisans for the House Election				
Bandwidth (Margin of Victory)	Estimate	Lower 95% CI	Lower 95% CI	Number of cases
Optimal Bandwidth Selection				
14.4%	0.021	0.011	0.032	12770
Alternative Specification				
1%	0.031	−0.024	0.086	1080
5%	0.015	−0.006	0.036	5376
10%	0.016	0.001	0.031	9971
25%	0.022	0.012	0.032	17491
100%	0.023	0.017	0.030	21397

Note: Running variable is party's margin of victory for the local executive election at t, outcome is mayor/runner-up's co-partisans' vote-share for the House election at t_{+2}. Estimate is the average treatment effect at cutoff estimated with local linear regression with a triangular kernel. The main result in the first row uses a MSE-optimal bandwidth selection procedure (Calonico et al., 2014). Columns 3–5 report, respectively, 95% robust confidence intervals, and the number of treated and control observations within the bandwidth. Rows 2-5 present results using alternative *adhoc* bandwidths.

Optimal bandwidth selection

Treatment effect estimate (interpreted as percentages)

Confidence interval

Sample size

Literature Example: Conceptual Check

- This model works well because we have a clear binary variable:
  - Party mayoral candidate either won (1) or lost (0)
  - Can also be expressed as a vote margin

- We have a clear theory about how this would impact the party's vote share in the general election:
  - An incumbent mayor would be expected to help ("reverse coattails")

- Alternative example: If our theory was about how spending in the mayoral race would affect the party's chances in the general, RDD would be less appropriate
  - There is no clear cut-point on which we could divide two regression groups

Code: Literature Replication (Table Results)

```
# Defininin the variables
y<- df$vs_party_fed
r <- 100*df$vote_margin_share
band <- list("opt", 1,5,10, 25, 100)

# Model local linear
res <- sim_rd(y=y, x=100*df$vote_margin_share,
          vary = band, p=1)

# Selecting the Results
d_fig <- map_df(res, extract)

# Table 1 --------------------------------------------------------


d_fig %<>%  mutate_if(is.numeric, round, digits=3) %<>%
  mutate(h=paste(c(1, 5, 10, 14.4, 25, 100),"\\%", sep="")) # add by hand the optimal value label


colnms  = c("Bandwidth (Margin of Victory)", "Estimate", "Lower 95\\% CI",
         "Lower 95\\% CI","Number of cases")
```

```
table<- Hmisc::latex(d_fig, file=here("outputs","table_1.tex"),
          title="",
          table.env=FALSE,
          center="none",
          col.just = c("l",rep("c",ncol(d_fig)-1)),
          colheads=colnms,
          booktabs = TRUE,
          rowname = NULL,
          caption = "\\textsc{Regression Discontinuity Results}",
          n.cgroup=5,
          cgroup = "\\textsc{Outcome: Vote Share Co-partisans for the House Election}")
```

- Primary RDD package required: ‘rdrobust’
- Also requires several other packages
- Define variables and cut-point needed for model
- Run model (d_fig) using sim_rd function
- Create a formatted Latex table using the model
- Link to full replication code

Code: Literature Replication (Figure)

```
# Treatment ---------------------------------------------------------
d$treat <- ifelse(d$vote_margin_share <0, "Non Incumbent", "Incumbent")
df <- d


# Figure 2 ------------------------------------------------------------
cut <- cut(df$vote_margin_share,500, include.lowest = TRUE)
tmp <- aggregate(y, by=list(cut = cut), FUN=mean, na.rm=T)
tmp1 <- aggregate((df$vote_margin_share), by=list(cut = cut), FUN=mean, na.rm=T)
data <- data.frame(margin = tmp1$x, y = tmp$x)


ggplot() +
  geom_point(data = data[data$margin <0 ,], aes(margin, y),na.rm=T,size=3, color = 'gray', alpha=.8) +
  geom_point(data = data[data$margin >0 ,], aes(margin, y),na.rm=T,size=3, color = 'gray', alpha=.8) +
  stat_smooth(data = df,
        aes(vote_margin_share, vs_party_fed, group=treat, color=treat)) +
  geom_hline(yintercept=0, linetype="dotted") +
  xlab("Margin (Results within the optimal bandwidth)") +
  geom_vline(xintercept=0, linetype="dotted") +
  ylab("Vote Share for the House Elections") +
  scale_color_discrete(name="") +
  geom_vline(xintercept=-0.14, linetype="F1", color="tomato2", alpha=.5) +
  geom_vline(xintercept=0.14, linetype="F1", color="tomato2", alpha=.5)
```

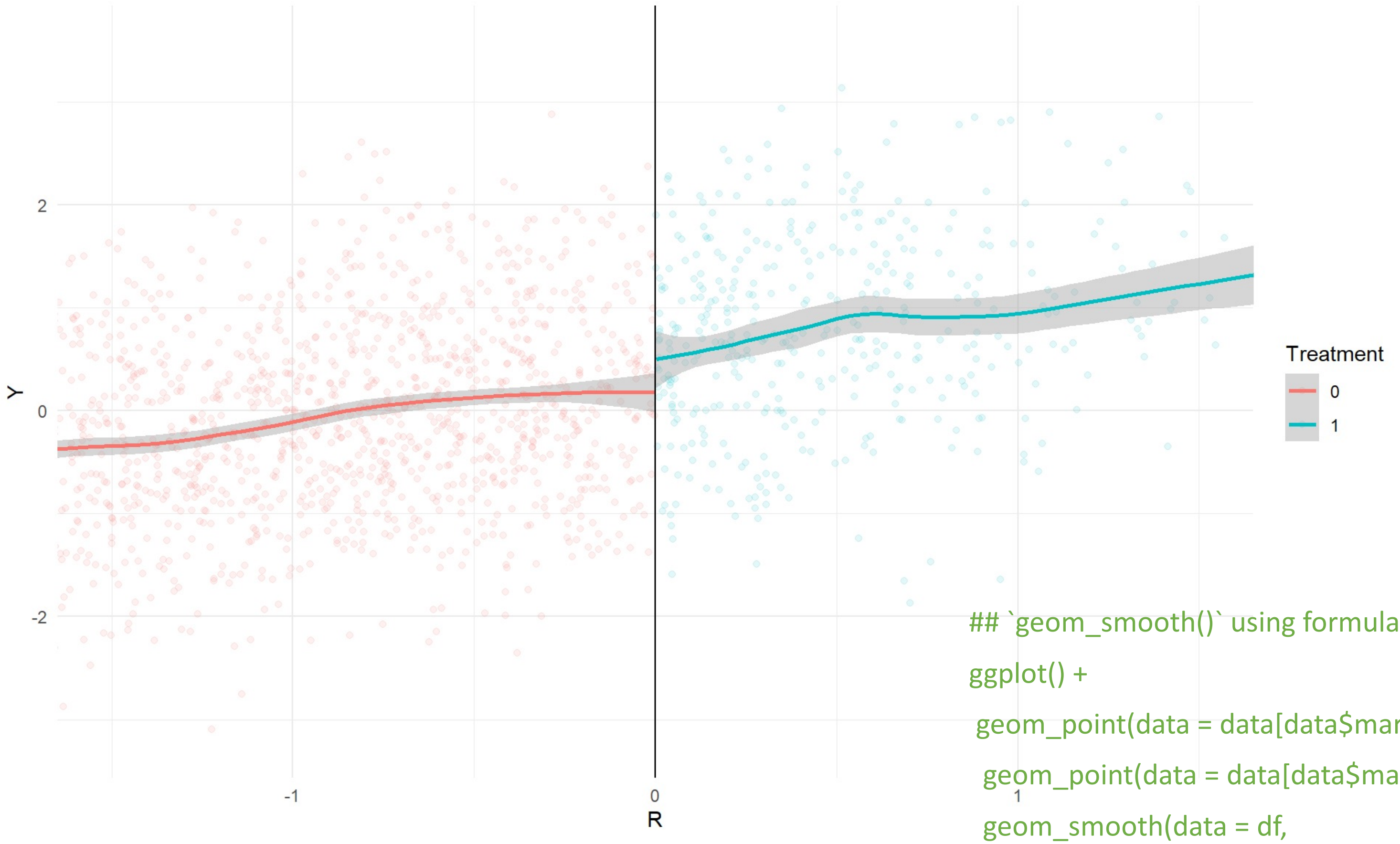
- Primary package required: ggplot
- Also requires several other packages
- Establishment of binary variable and cutpoint
- Establish group designations
- Construction of the figure (ggplot)
- Link to full replication code

Code: General Example (Table Results)

```
library(rdd)

bw <- with(RD_data, IKbandwidth(R, Y, cutpoint = 0))

rdd_simple <- RDestimate(Y ~ R, data = RD_data, cutpoint = 0, bw = bw)

summary(rdd_simple)
##
## Call:
## RDestimate(formula = Y ~ R, data = RD_data, cutpoint = 0, bw = bw)
##
## Type:
## sharp
##
## Estimates:
##           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
## LATE      1.0894     1177          0.3035    0.11323     2.680    0.007355  **
## Half-BW   0.5447     611           0.2308    0.15471     1.492    0.135722
## Double-BW 2.1787     1832          0.2699    0.08968     3.010    0.002613  **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## F-statistics:
##           F       Num. DoF  Denom. DoF  p
## LATE      37.73   3         1173        0.000e+00
## Half-BW   12.64   3         607         1.006e-07
## Double-BW 104.74  3         1828        0.000e+00
```

- Another code option is ‘rdd’
- Establishment of the cutpoint
- Construction of model
- Call model results
- Tutorial

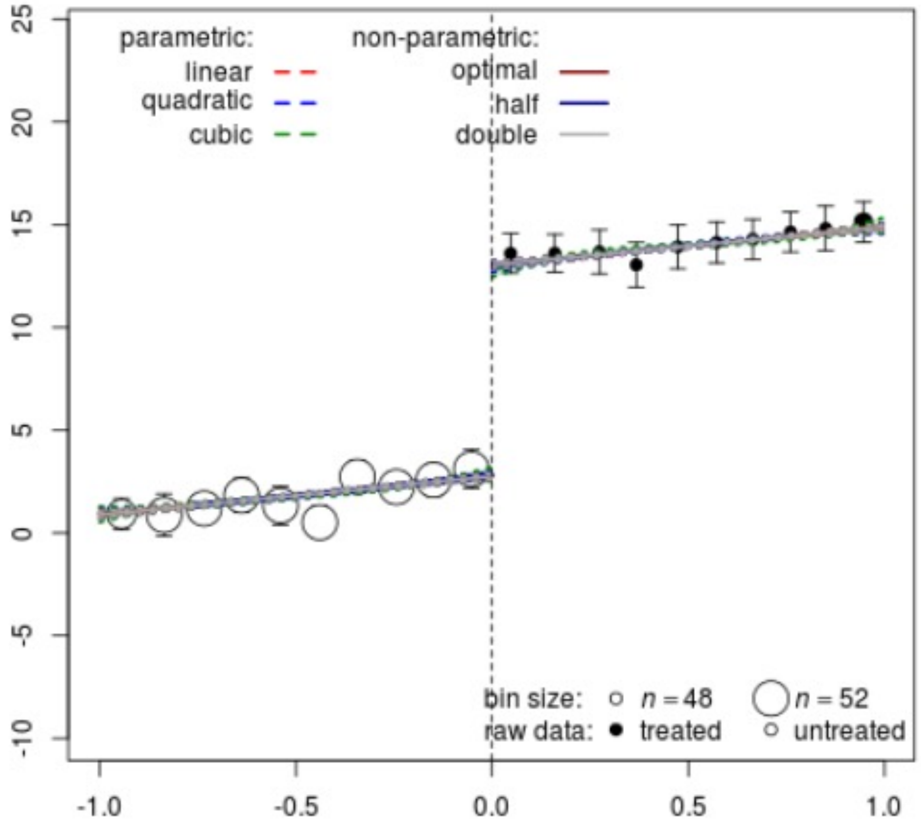
Code: General Example (Figure)



- Use ggplot to create RDD figure
- Uses example data from literature paper
- Figure itself can be found here: Tutorial

```
## `geom_smooth()` using formula 'y ~ x’
ggplot() +
 geom_point(data = data[data$margin <0 ,], aes(margin, y),na.rm=T,size=3, color = 'gray', alpha=.8) +
  geom_point(data = data[data$margin >0 ,], aes(margin, y),na.rm=T,size=3, color = 'gray', alpha=.8) +
  geom_smooth(data = df,
              aes(vote_margin_share, vs_party_fed, group=treat, color=treat))
```

Code: General Example 2 (Figure)



- Another simple example using the rdd package instead of ggplot
- Function utilized is plot.RD
- Tutorial

```
dat <- data.frame(x = runif(1000, -1, 1), cov = rnorm(1000))

dat$tr <- as.integer(dat$x >= 0)

dat$y <- 3 + 2 * dat$x + 3 * dat$cov + 10 * (dat$x >= 0) + rnorm(1000)

rd <- rd_est(y ~ x + tr | cov, data = dat, cutpoint = 0, t.design = "geq")

plot(rd)
```

For further learning

1. Imbens and Lemieux 2008 – General methods article on the use of the RDD method with observational data

2. Gelman and Imbens 2019 – Discusses the degree of line fit, recommends linear or other low-order line fits

3. Keele and Titiunik 2015 – Examines geographic boundaries as random cutpoints in constructing natural experiments

4. de la Cuesta and Imai 2016 – Annual Review article discussing close election application

Another great RDD source for R (and conceptual): https://rpubs.com/cuborican/RDD

Summary

- RDD examines two groups, one exposed to a treatment and the other not, to estimate an effect of that treatment
    - The method is “native” to the experimental world, but here we apply it to observational data
- RDD is a good way to examine binary independent variables that can also vary in degrees
    - It can also be used as a way to examine “natural experiments”
- Key points of concern come around the assumptions made for the cutpoint , the optimal bandwidths, and the line estimation method
- The most common application looks at election data

References

1. De la Cuesta, B., & Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, *19*, 375-396.
2. Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447-456.
3. Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, *142*(2), 615-635.
4. Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*(1), 127-155.
5. Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281-355.
6. McCrary (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test". Journal of Econometrics. 142 (2): 698–714.
7. Ventura, T. (2021). Do mayors matter? Reverse coattails on congressional elections in Brazil. *Electoral Studies*, *69*, 102242.

Note: Special thanks to Evan Jones and Tiago Ventura for assistance with this presentation