Exercise 5 - Replication of Results

Devan Arnold

12/02/2023

Instructions

Complete the exercises below using. Be sure to show all of your work. For this assignment, you should submit a PDF document containing your written answers with the R code embedded in the document.

This work is based on the 2004 paper “Do Voters Affect or Elect Policies? Evidence from the US House” by David S Lee, Enrico Moretti, and Matthew J Butler

---

Task 1 - Paper Summary

As mentioned in the introduction, this exercise will be an attempt to replicate the results of the 2004 paper “Do Voters Affect or Elect Policies? Evidence from the US House” by David S Lee, Enrico Moretti, and Matthew J Butler. This paper sets out to determine the manner in which voters influence policy in the federal government: do they directly elect politicians that then enact policies consistent with their campaign platform, or do voters instead influence how politicians vote?

To this end, the authors set out to compare the score issued by the Americans for Democratic Action (ADA) against the share of votes received during the previous election cycle. The authors utilize Regression Discontinuity Design (RDD) in their analysis to determine the ADA score gap that occurs at the 50% vote share line. Since US elections are essentially a binary choice, the 50% vote share line represents the point under which the measured party loses the election, and above which they win. The gap that occurs here is measured in the paper by gamma (γ), which itself is composed of two parts – the “elect” and “affect” portions. These portions serve as the means by which the authors answer their research question.

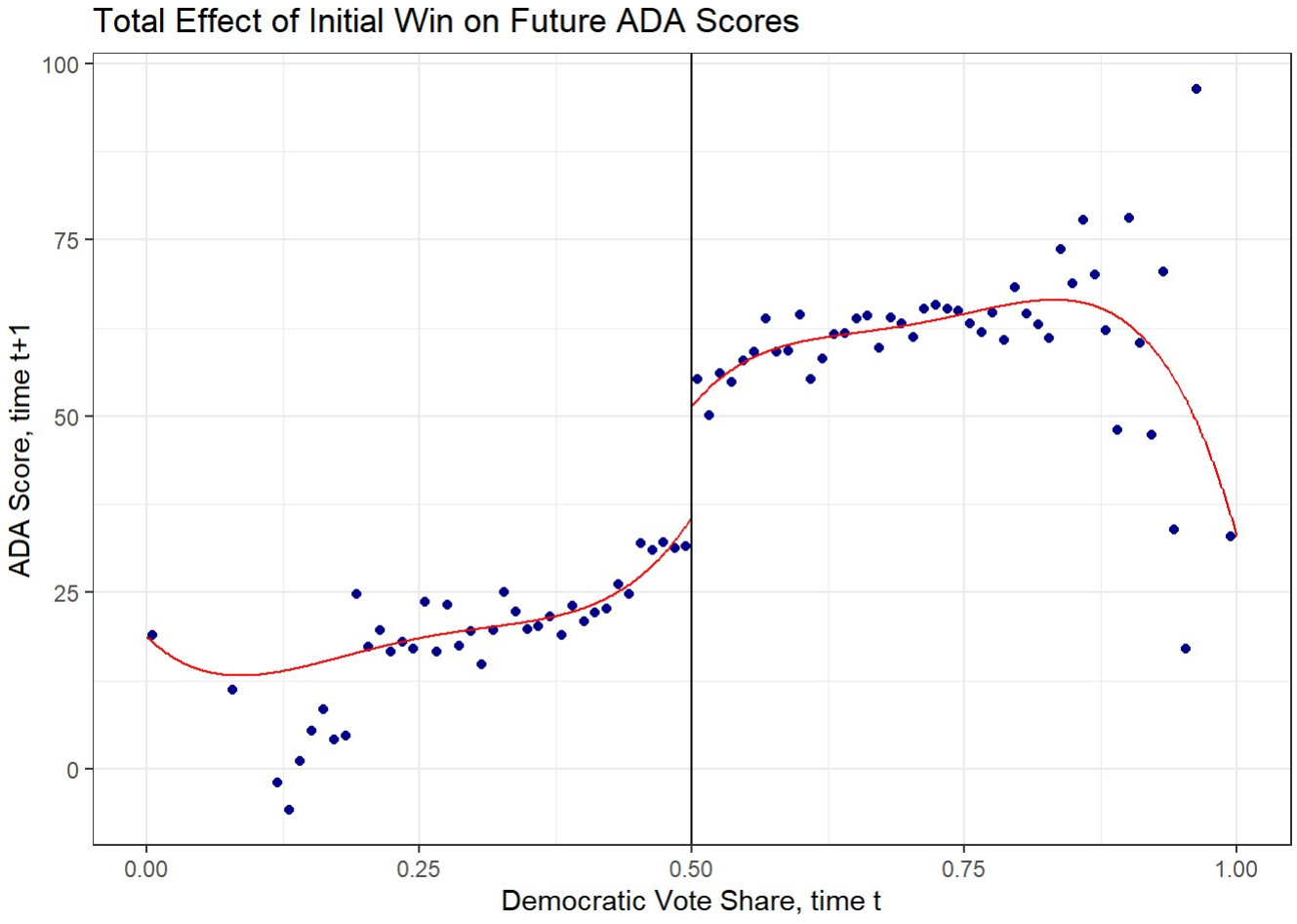
RDD is a good choice of estimation model for this problem for a few reasons. Firstly, as mentioned prior, US elections are in essence binary choices. The authors describe that at the +/-2% vote share level that the winner of an election is essentially random. Secondly, the authors are able to argue that there is no covariance in the demographics of the voting base that explains the discontinuity in the relationship between voting share and ADA score which strengthens the case for RDD. This randomization of outcomes in a rules-based environment and little covariance with other possible confounding influences creates a strong case for an RDD approach.

Over the next few tasks set out I will attempt to recreate the results that the authors achieved in the original paper.

Task 2 - Vote Share and ADA Score

- **Create your own version of LMB's Figure I.**

```
# Sets the number of bins to equal the number of observed districts
bin_count.2 <- length(unique(plot.df$district))
# Creates an RDD plot for the relationship
rdplot(y=plot.df$score,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_count.2,
       title = "Total Effect of Initial Win on Future ADA Scores",
       x.label = "Democratic Vote Share, time t",
       y.label = "ADA Score, time t+1")
```



- **Obtain your own estimate (or estimates) of γ.**

```
# Creates an RDD object and reports the determined gap
task2 <- rdrobust(y=plot.df$score,x=plot.df$lagdemvoteshare,c=0.5)
task2$Estimate[1]
```

```
## [1] 18.66595
```

- **Compare your estimate to the estimate reported in Column 1 of Table I.**

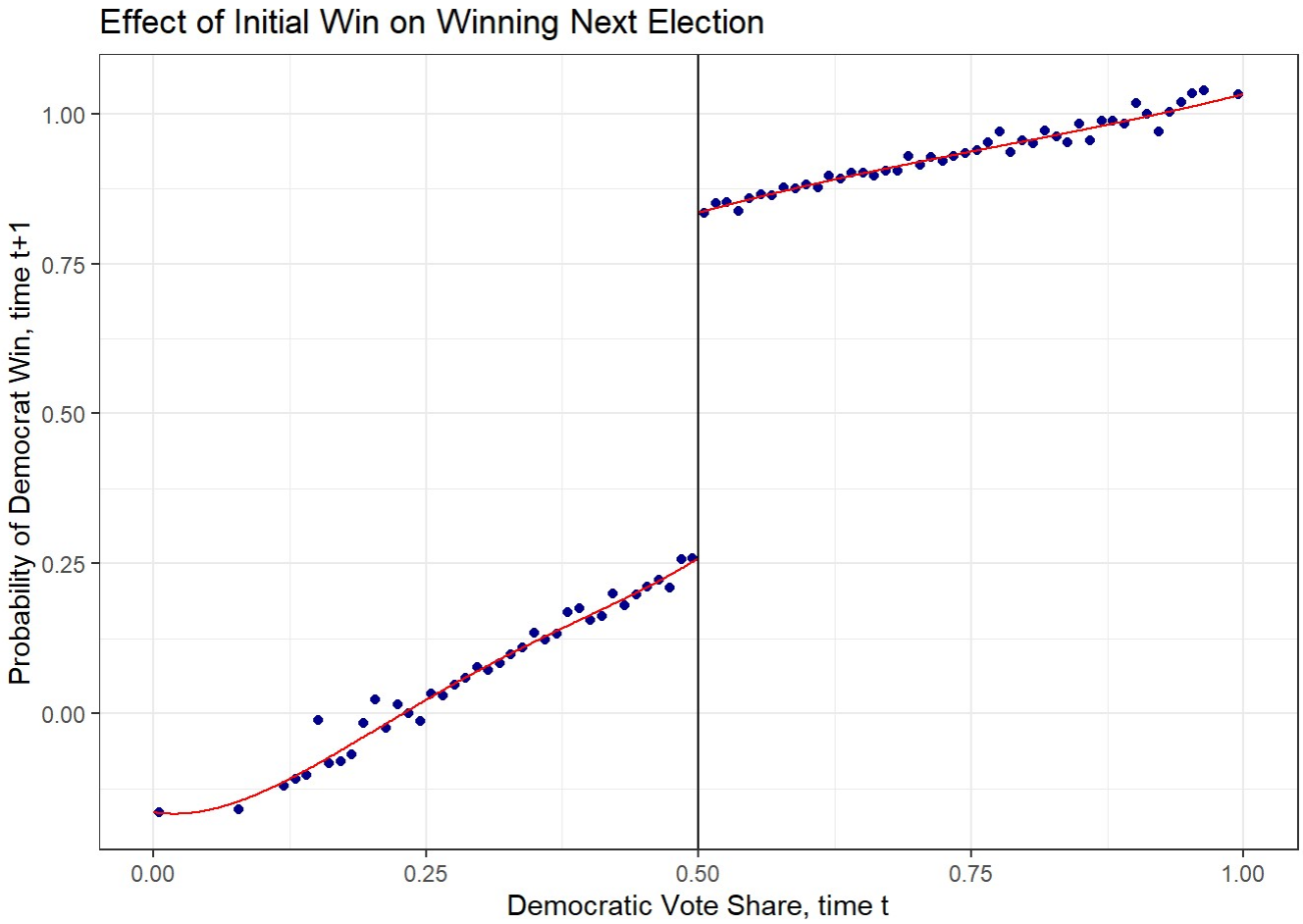
As we can see, the estimate that I generated from the data is 18.7, whereas the authors generated an estimate of 21.2. That puts my estimate approximately 1.3 standard errors (as measured by the authors) below the estimate that they generated.

Task 3 - Incumbancy Effects

- **Create your own version of LMB's Figure IIb**

```
# Creates the predictive model to determine probability of a democrat victory in time
# t+1
probability.D.model <- lm_robust(democrat~lagdemvoteshare + incmbncy + lagdemvoteshare*incmbn
cy,plot.df,se_type='HC1')
# Adds predicted values to the plot.df data frame
plot.df <- plot.df %>% mutate("prob_D" = predict(probability.D.model,plot.df))

# Sets the number of bins to equal the number of observed districts
bin_count.3 <- length(unique(plot.df$district))
# Creates an RDD plot for the relationship
rdplot(y=plot.df$prob_D,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_count.3,
       title = "Effect of Initial Win on Winning Next Election",
       x.label = "Democratic Vote Share, time t",
       y.label = "Probability of Democrat Win, time t+1")
```



- **Obtain an estimate (or estimates) of the effect represented in this figure.**

```
# Creates an RDD object and reports the determined gap
task3 <- rdrobust(y=plot.df$prob_D,x=plot.df$lagdemvoteshare,c=0.5)
task3$Estimate[1]
```

```
## [1] 0.568344
```

- **Compare your estimate to the estimate reported in Column 3 of Table I.**

The value estimated using my model is 0.57, whereas the value estimated by the authors is 0.48. My estimate is significantly outside of the range that the authors estimated, and I am sure this is due to the estimation approach that I took. Although I reviewed the estimation methods outlined in Lee 2001 and Lee 2008 (formerly Lee 2003, but it appears that the author had the working paper published since this 2004 paper) I was unable to determine the exact estimation method given in this paper.

Task 4 - Tests for Quasi-Random Assignment

- **Create your own version of LMB's Figure 3**

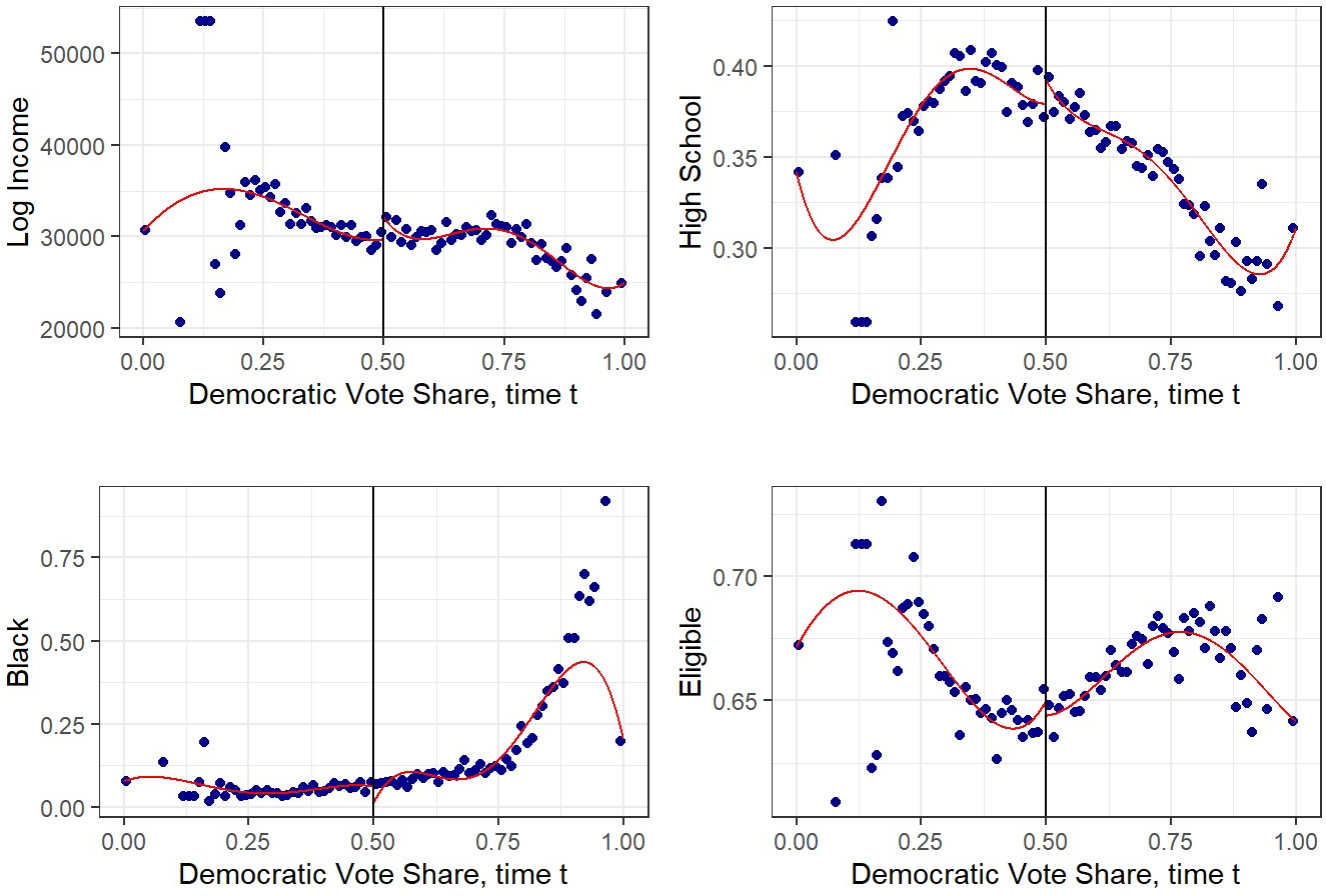
```
# Changes the bin count to match the number of congresses held during the data period
# instead of the number of districts
bin_count.4 <- length(unique(plot.df$district))

# Creates a variable for the percent of population that is voting eligible
plot.df <- plot.df %>% mutate("eligible_percent"=votingpop/totpop)

# Creates RDD plots for each category of interest from the paper
# Income
plot4.1 <- rdplot(y=plot.df$realincome,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_count.4,
                  title="",
                  x.label = "Democratic Vote Share, time t",
                  y.label = "Log Income")
# High School Completion percentage
plot4.2 <- rdplot(y=plot.df$pcthighschl,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_count.4,
                  title="",
                  x.label = "Democratic Vote Share, time t",
                  y.label = "High School")
# Percent of population black
plot4.3 <- rdplot(y=plot.df$pctblack,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_count.4,
                  title="",
                  x.label = "Democratic Vote Share, time t",
                  y.label = "Black")
# Percent of population eligible voters
plot4.4 <- rdplot(y=plot.df$eligible_percent,x=plot.df$lagdemvoteshare,c=0.5,nbins = bin_coun
t.4,
                  title="",
                  x.label = "Democratic Vote Share, time t",
                  y.label = "Eligible")

# Combines the above RDD plots into one plot object, arranged into a 2x2 grid
plot4.fullplot <- grid.arrange(plot4.1$rdplot,plot4.2$rdplot,plot4.3$rdplot,plot4.4$rdplot,nc
ol=2)
```

```
# Outputs the compiled 4 plot figure to the console
plot(plot4.fullplot)
```



- **Replicate the associated columns in Table 2**

```
# Creates the ln(realincome) variable in the plot.df data frame
plot.df <- plot.df %>%  mutate("ln_realincome" = log(realincome))

# Creates linear regression models for the conditions listed in the paper
task4.col1.model <- plot.df %>% lm_robust(formula = democrat~ log(realincome) + pcthighschl +
pctblack + eligible_percent + totpop + factor(district) + factor(congress) + factor(district)
*factor(congress),se_type = "HC1")
task4.col2.model <- plot.df %>% filter(demvoteshare >= 0.25 & demvoteshare <= 0.75) %>%
              lm_robust(formula = democrat~ log(realincome) + pcthighschl + pctblack + eligib
le_percent + factor(district) + totpop + factor(congress) + factor(district)*factor(congres
s),se_type = "HC1")
task4.col3.model <- plot.df %>% filter(demvoteshare >= 0.40 & demvoteshare <= 0.60) %>%
              lm_robust(formula = democrat~ log(realincome) + pcthighschl + pctblack + eligib
le_percent + factor(district) + totpop + factor(congress) + factor(district)*factor(congres
s),se_type = "HC1")
task4.col4.model <- plot.df %>% filter(demvoteshare >= 0.45 & demvoteshare <= 0.55) %>%
              lm_robust(formula = democrat~ log(realincome) + pcthighschl + pctblack + eligib
le_percent + factor(district) + totpop + factor(congress) + factor(district)*factor(congres
s),se_type = "HC1")
task4.col5.model <- plot.df %>% filter(demvoteshare >= 0.48 & demvoteshare <= 0.52) %>%
              lm_robust(formula = democrat~ log(realincome) + pcthighschl + pctblack + eligib
le_percent + factor(district) + totpop + factor(congress) + factor(district)*factor(congres
s),se_type = "HC1")
# Creates the RDD estimates for each variable of interest
task4.col6.row1 <- rdrobust(y=plot.df$ln_realincome,x=plot.df$lagdemvoteshare,c=0.5,p=4)
task4.col6.row2 <- rdrobust(y=plot.df$pcthighschl,x=plot.df$lagdemvoteshare,c=0.5,p=4)
task4.col6.row3 <- rdrobust(y=plot.df$pctblack,x=plot.df$lagdemvoteshare,c=0.5,p=4)
task4.col6.row4 <- rdrobust(y=plot.df$eligible_percent,x=plot.df$lagdemvoteshare,c=0.5,p=4)
# Creates vectors of the coefficients of interest, as well as the number of observations
task4.col1 <- c(task4.col1.model$coefficients[2],task4.col1.model$coefficients[3],task4.col1.
model$coefficients[4],task4.col1.model$coefficients[5],task4.col1.model$nobs)
task4.col2 <- c(task4.col2.model$coefficients[2],task4.col2.model$coefficients[3],task4.col2.
model$coefficients[4],task4.col2.model$coefficients[5],task4.col2.model$nobs)
task4.col3 <- c(task4.col3.model$coefficients[2],task4.col3.model$coefficients[3],task4.col3.
model$coefficients[4],task4.col3.model$coefficients[5],task4.col3.model$nobs)
task4.col4 <- c(task4.col4.model$coefficients[2],task4.col4.model$coefficients[3],task4.col4.
model$coefficients[4],task4.col4.model$coefficients[5],task4.col4.model$nobs)
task4.col5 <- c(task4.col5.model$coefficients[2],task4.col5.model$coefficients[3],task4.col5.
model$coefficients[4],task4.col5.model$coefficients[5],task4.col5.model$nobs)
task4.col6 <- c(task4.col6.row1$Estimate[1],task4.col6.row2$Estimate[1],task4.col6.row3$Estim
ate[1],task4.col6.row4$Estimate[1],task4.col1.model$nobs)


# Creates and stores a data to combine above information
task4.models <- data.frame()
task4.models <- cbind(task4.col1, task4.col2, task4.col3, task4.col4, task4.col5, task4.col6)
# Applies names to rows and columns
colnames(task4.models) <- c("All","+/-0.25","+/-0.10","+/-0.05","+/-0.02","Polynomial")
rownames(task4.models) <- c("Log Income"," Percent HS Grad","Percent Population Black","Perce
nt Eligible Voters","Num. Obs")
```

```
# Creates kable object to display the results
task4.table <- kable(task4.models, align = "c", digits = 3)
task4.table <- task4.table %>%  kableExtra::row_spec(row = 5,bold = T,extra_css = "border-to
p: 1px solid")
```

```
# Creates the kable table of results
kable_classic_2(task4.table)
```

	All	+/-0.25	+/-0.10	+/-0.05	+/-0.02	Polynomial
Log Income	-0.078	0.037	0.262	0.286	0.457	0.067
Percent HS Grad	-0.694	-0.621	-0.380	-0.925	-1.865	0.013
Percent Population Black	0.812	0.754	0.272	-0.117	0.028	0.000
Percent Eligible Voters	1.743	2.267	1.421	0.614	0.240	-0.016
Num. Obs	**9258.000**	**6924.000**	**2763.000**	**1382.000**	**553.000**	**9258.000**

- **Discuss the purpose of this figure and how it supports (or does not) the rest of your analysis**

As discussed in my summary of the paper, the lack of discontinuity present in these potential confounding factors strengthens the argument for the use of Regression Discontinuity Design in this context. This composite figure is meant to illustrate that fact.

Task 5 - Conclusions

In this exercise, I took steps to review the work done by the authors, to understand the methods that they implemented, and to deconstruct their methods for replication. This required careful reading of the paper, consultation of external resources from both the course and outside of it, and I created the regression models to replicate the results of the authors.

In the end, my efforts to replicate this paper were mixed. My numerical estimations were way off base, but I was able to recreate the plots used by the authors for the most part. Personally, I think that this creates an interesting problem: economic papers have to explain what they did, why it matters, what their results are, and the implications of those results but recreating the work is hard. In other sciences, my understanding of the purpose of lab reports is to document the hypothesis of the researcher, the methods of the test, the steps the researchers took, and their results. However, in economics (and in this class as well) we have learned how direct replication and inclusion of steps for direct replication are discouraged at the institutional level. Economics has a hard time being taken seriously by other sciences (we have a perception problem), and for me this exercise highlighted that. I can see the results of the author, but explicit statement of methods is excluded from the paper.

I do not know how I could have done this task better, and while I can read the results of the authors, I cannot replicate their findings. This speaks to either my lack of skill in this area, the difficulty of reproduction via inference, a wider issue in economic literature, or some combination of the three.