Lab Notes

2017-5-26

Quasi-Experimental Research Designs

When a research design uses random assignment (this is different than random selection), then we have a **randomized** or **true experiment**. Some examples include randomized controlled trials, field experiments, lab experiments, and lab-in-the-field experiments.

When we do not have random assignment, but there are still some observations that receive the treatment of interest when others do not, then we may have what is called a **quasi-experiment**. Quasi-experiments are observational studies in that the treatment is not under the full control of the researcher. That is, the researcher does not randomly assign observations to a treatment or control group. But, circumstances can arise when observations *are* assigned to a treatment or control group, and the mechanism by which this occurs may be considered "as good as random". In such cases, we can compare the average outcome between these groups and obtain an estimate that may be considered causal.

A point on semantics: quasi-experiments and natural experiments are similar, and they are sometimes used interchangeably. **Natural experiments** are a type of quasi-experment, when the assignment mechanism occurs "naturally" or in a way that is not under the control of the researcher. This leaves a another subset of quasi-experiments when the treatment is not randomized, but the researcher attempts to create an assignment mechanism that is "as if random".

These can be contrasted with **non-experimental designs**, when all observations either receive the treatment or do not receive the treatment, and there is no distinction between the treatment and control group. We'll be convering two common type types of quasi-experimental research designs: **regression discontinuity designs** and **non-equivalent groups designs**.

Regression Discontinuity Designs

The central idea underlying the regression discontinuity design (RDD) is an **assignment variable**, also called a **running variable**, that determines assignment into a treatment group or control group after some known cutoff point. Each observation falls into the treatment group or control group based on its value of the assignment variable: when an observation has a value of the assignment variable that is greater than or equal to the cutoff, then it is assigned to the treatment group; when it is below the cutoff, then it is assigned to the control group.

If observations are unable to *precisely* manipulate their value of the assignment variable, so that they cannot sort themselves into the treatment group or control group, then the observations just above and below the cutoff point may be considered "as good as randomized". That is, they are good comparisons of one another. The treatment effect can therefore be estimated as the difference in the average outcome just above and just below the cutoff.

Examples

One example is by Card and Shore-Sheppard (2004). They are interested in assessing whether Medicaid improves health outcomes. The problem with merely comparing individuals with and without Medicaid on health outcomes is that those individuals who choose to adopt Medicaid may be also the ones more prone to illness (or vice versa). This is the familiar issue of selection bias. Card and Shore-Shappard therefore exploit a RDD using date of birth as the assignment variable. They observe that effective July 1991, US law

required states to cover children born after September 30, 1983 until their 18th birthday, while those born on or before the date received no such coverage. The authors therefore exploit this cutoff to estimate the difference beteween these groups in whether Medicaid saves lives.

Another example is by Lee (2008) who is interested in the question of whether political incumbents hold an advantage over non-incumbents in elections. This is based on the hypothesis that politicians elected into office make choices that respond to the demands of the electorate, which raises their chances of re-election. In the United States House of Representatives, incumbents were also successful around 90 percent of the time. The issue, however, is that incumbents may be re-elected for many other reasons, and not just due to their incumbency: they may be richer, more charismatic, or Democratic incumbents may be more successful because they are being elected in heavily Democratic districts. Lee therefore uses a RDD to test the incumbency hypothesis by comparing electoral outcomes of candidates who just barely won elections to those who just barely lost election, which he argues that under mild assumptions are comparable on average except in their incumbency status.

The Validity of Regression Discontinuity Designs

One advantage of RDDs is the relatively mild set of assumptions compared to other quasi-experimental designs, and they are potentially more credible compared to designs that use difference-in-differences or instrumental variables. Unlike these approaches, one does not need to assume that the design isolates treatment variation that is "as good as randomized". The validity of the design does not rest of the (untestable) assumption of the exclusion restriction or parallel trends. Rather, this is a consequence of the inability to precisely control the assignment near the cutoff point.

The following summarizes the most important points about the validity of RD designs (from Lee and Lemieux 2008):

- Regression discontinuity designs can be invalid if individuals can *precisely* manipulate the assignment variable. Precise control can be distinguished between imprecise control and complete control.
- If individuals are unable to *precisely* manipulate the assignment variable, even though they exert some influence, a consequence of this is that the variation in treatment near the threshold is randomized as though from a randomized experiment. This implies that all observed and unobserved predetermined characteristics will have identical distributions on either side of the cutoff point in the limit, at smaller and smaller neighborhoods of the threshold.
- Regression discontinuity designs can be analyzed and tested like randomized experiments. Baseline covariates should have the same distribution just above and just below the cutoff, and they can be used to test the validity of the RDD.
- Graphical presentations can be helpful and informative, but they should not be tilted toward finding an effect or no effect.
- Nonparametric estimation does not represent a solution to functional form issues. It is helpful to view this as a complement rather than a substitute to parametric estimation.
- Goodness-of-fit and other statistical tests can help to rule out overly restrictive specifications

Estimation

A simple way of implementing RDDs is to estimate two separate regressions on each side of the known cutoff point. The treatment effect can then be computed as the difference between the two intercepts.

The regression on the left hand side of the cutoff point is therefore

$$y_i = \alpha_l + f_l(x_i - c) + e_i$$

α_l is the intercept of this left-side regression. x_i is the assignment variable. It is convenient to subtract c from x_i, so the cutoff is centered at zero. But this does not affect our estimates. f_l is the slope of this regression.

The regression on the right hand side of the cutoff point is as follows.

$$y_i = \alpha_r + f_r(x_i - c) + e_i$$

The treatment effect is therefore estimated as $\alpha_r - \alpha_l$. A more direct way of estimating these regressions is the run a pooled regression on both sides of the cutoff point.

$$y_i = \alpha_l + \tau D_t + f(x_i - c) + e_i$$

In this expression, $\tau = (\alpha_r - \alpha_l)$ and $f(x_i - c) = f_l(x_i - c) + D_t\big(f_r(x_i - c) - f_l(x_i - c)\big)$. τ is now the estimate of the treatment effect.

This is the a common way of estimating the RDD using linear regression. However, the slope on both sides of the cutoff can also be made to vary by including the interaction terms between D_t and x_i.

Let's try to replicate some of the tables and figures in Lee (2008)

```
rm(list=ls())

library(stats)
library(ggplot2)
library(plm)
```

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

```
library(rdd)
```

```
## Loading required package: sandwich

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

## Loading required package: AER

## Loading required package: car

## Loading required package: survival
```

```
data <- read.csv("Lee2008.csv", header=T)

colnames(data)
```

```
##  [1] "statedisdec"     "demsharenext"    "demshareprev"
##  [4] "use"             "demwinprev"      "difdemshare"
##  [7] "difdemshare2"    "difdemshare3"    "difdemshare4"
## [10] "demofficeexp"    "demelectexp"     "othofficeexp"
## [13] "othelectexp"     "right"           "rdifdemshare"
## [16] "rdifdemshare2"   "rdifdemshare3"   "rdifdemshare4"
```

```
## [19] "difdemsharenext"
```

```
#The data set has the following variables:

#demsharenext: Democrat vote share election t+1
#right: Dummy variable of victory in election t
#difdemshare: Democrat vote share election t
#rdifdemshare: Interaction between right and difdemshare

#demwinnext: Democrat win prob. election t+1
#demshareprev: Democrat vote share election t-1
#demwinprev: Democrat win prob. election t-1
#demofficeexp: Democrat political experience
#othofficeexp: Opposition political experience
#demelectexp: Democrat electoral experience
#othelectexp: Opposition electoral experience
```

The outcome variable is *demsharenext* or the vote share of a Democrat in election $t + 1$. Its range is from 0 to 1.

```
summary(data$demsharenext)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.3948  0.5396  0.5532  0.7002  1.0000    2618
```

The assignment variable is $difdemshare$ or the vote share of a Democrat in election t. Its range is from -1 to 1, which is 1 percentage point above and below the 50% cutoff for victory that is centered at zero.

```
summary(data$difdemshare)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.00000 -0.20590  0.07615  0.11860  0.40100  1.00000
```

The treatment variable is *right*. It is a dichotomous variable for whether the candidate won in election t.

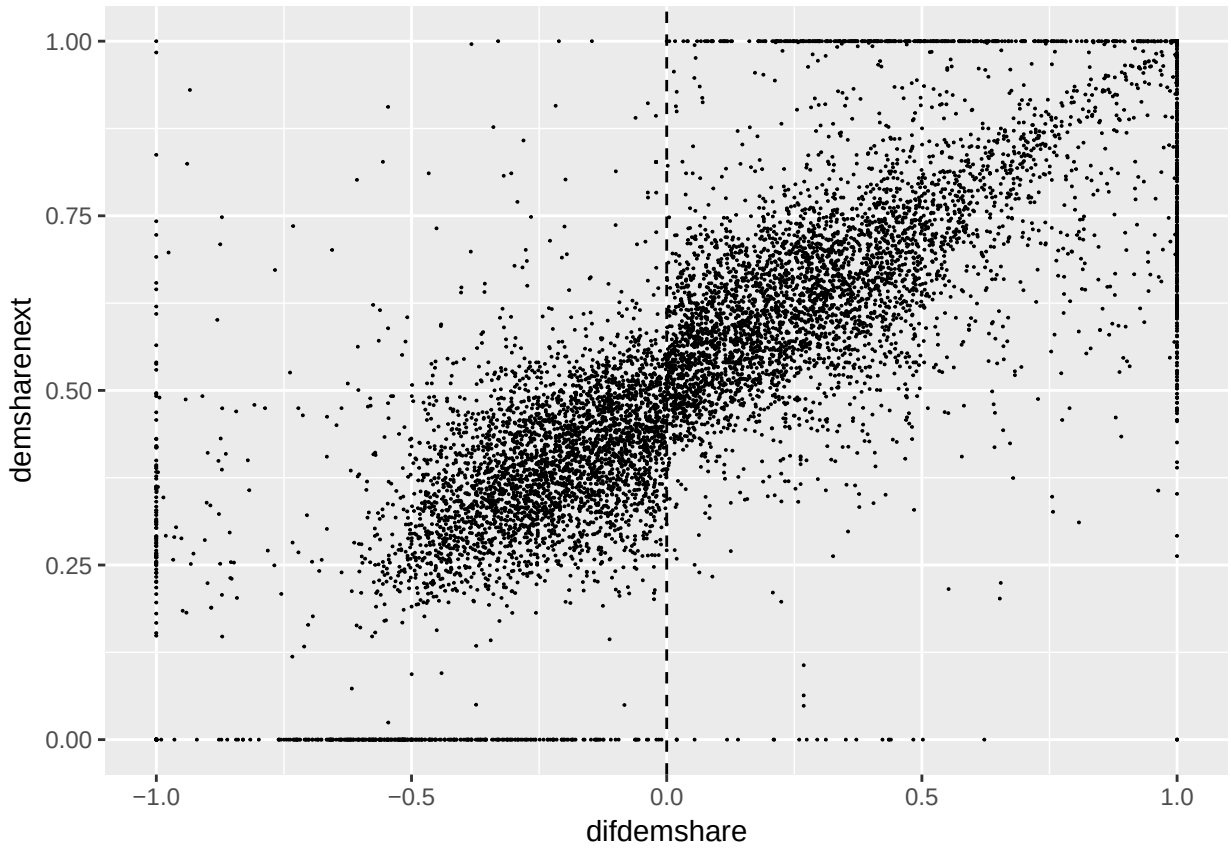
```
summary(data$right)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.5679  1.0000  1.0000       4
```

We can visualize the relationship between the outcome variable and the running variable.

```
ggplot(data, aes(difdemshare, demsharenext))+geom_point(size=0.01)+
  geom_vline(xintercept=0, linetype=2, color="black")
```

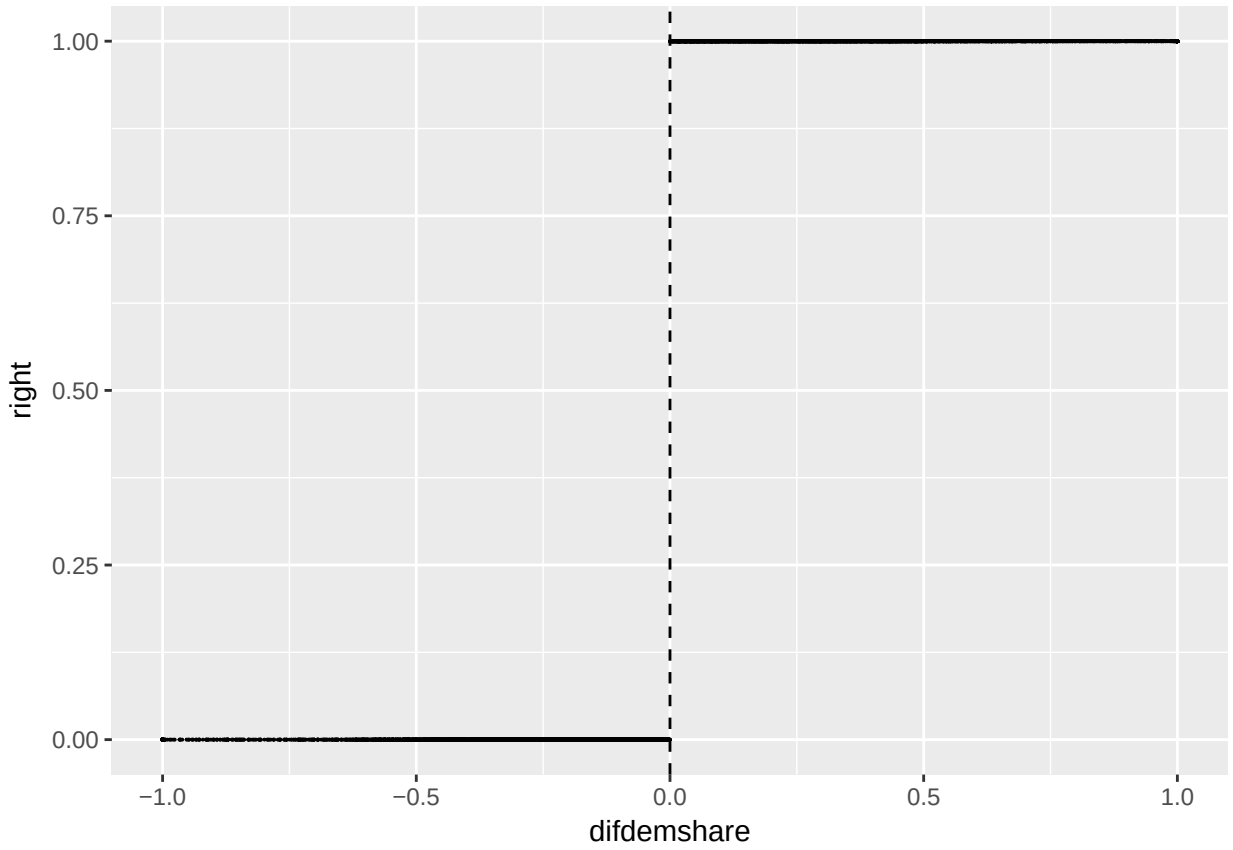
```
## Warning: Removed 2618 rows containing missing values (geom_point).
```



We can also inspect the relationship between the running variable and the treatment variable.

```
ggplot(data, aes(difdemshare, right))+geom_point(size=0.01)+
  geom_vline(xintercept=0, linetype=2, color="black")
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



Using these variables and the RDD set out my Lee, we can estimate the following model.

```
M0 <- lm(demsharenext ~ difdemshare + right, data=data)
summary(M0)
```

```
##
## Call:
## lm(formula = demsharenext ~ difdemshare + right, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89167 -0.06707  0.00288  0.07603  0.87929
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.454269   0.002667  170.34   <2e-16 ***
## difdemshare 0.333564   0.005123   65.11   <2e-16 ***
## right       0.103834   0.004695   22.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1395 on 9171 degrees of freedom
##   (2621 observations deleted due to missingness)
## Multiple R-squared:  0.6593, Adjusted R-squared:  0.6592
## F-statistic:  8874 on 2 and 9171 DF,  p-value: < 2.2e-16
```

How would you interpret the results?

Lee builds on this simple model to include an interaction term between the treatment variable and the running

variable and several polynomial terms of both the running variable and the interaction. These additional terms account for the functional form of the regression line on both sides of the cutoff. The following reults are from Table 3 in Lee (2001, 2008).

```
model <- formula(demsharenext ~ difdemshare + difdemshare2 + difdemshare3
                 + difdemshare4 + rdifdemshare + rdifdemshare2 +
                   rdifdemshare3 + rdifdemshare4 + right)

column1 <- lm(model, data=data)
summary(column1)
```

```
##
## Call:
## lm(formula = model, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88471 -0.06379  0.00376  0.07412  0.73013
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.444504   0.008419  52.800  < 2e-16 ***
## difdemshare    0.156080   0.134775   1.158  0.24686
## difdemshare2  -0.217491   0.627933  -0.346  0.72908
## difdemshare3   1.244400   1.055540   1.179  0.23846
## difdemshare4   1.451814   0.561324   2.586  0.00971 **
## rdifdemshare   0.324514   0.179084   1.812  0.07001 .
## rdifdemshare2 -0.132766   0.809315  -0.164  0.86970
## rdifdemshare3 -0.659819   1.321828  -0.499  0.61767
## rdifdemshare4 -1.817207   0.688640  -2.639  0.00833 **
## right          0.090681   0.011436   7.929 2.46e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1373 on 9164 degrees of freedom
##   (2621 observations deleted due to missingness)
## Multiple R-squared:  0.6704, Adjusted R-squared:  0.6701
## F-statistic:  2071 on 9 and 9164 DF,  p-value: < 2.2e-16
```

```
column2 <- lm(update(model, "~.+demshareprev+demwinprev"), data=data)
summary(column2)
```

```
##
## Call:
## lm(formula = update(model, "~.+demshareprev+demwinprev"), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80882 -0.06203  0.00132  0.06688  0.63704
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.320082   0.010840  29.527  < 2e-16 ***
## difdemshare    0.403924   0.152170   2.654 0.007963 **
## difdemshare2   1.251316   0.710651   1.761 0.078319 .
```

```
## difdemshare3   3.355114   1.196873   2.803 0.005074 **
## difdemshare4   2.387916   0.636841   3.750 0.000179 ***
## rdifdemshare  -0.059694   0.200646  -0.298 0.766086
## rdifdemshare2 -1.565522   0.914585  -1.712 0.086994 .
## rdifdemshare3 -2.754913   1.487911  -1.852 0.064138 .
## rdifdemshare4 -2.781310   0.776439  -3.582 0.000343 ***
## right          0.077701   0.012685   6.125 9.58e-10 ***
## demshareprev   0.292511   0.013131  22.277  < 2e-16 ***
## demwinprev    -0.016553   0.006624  -2.499 0.012474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.13 on 6546 degrees of freedom
##   (5237 observations deleted due to missingness)
## Multiple R-squared:  0.7098, Adjusted R-squared:  0.7093
## F-statistic:  1455 on 11 and 6546 DF,  p-value: < 2.2e-16
```

```
column3 <- lm(update(model, "~.+demofficeexp+othofficeexp"), data=data)
summary(column3)
```

```
##
## Call:
## lm(formula = update(model, "~.+demofficeexp+othofficeexp"), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89203 -0.06432  0.00367  0.07343  0.72725
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.4501384  0.0085012  52.950  < 2e-16 ***
## difdemshare    0.1646730  0.1346762   1.223  0.22146
## difdemshare2  -0.1548394  0.6270374  -0.247  0.80496
## difdemshare3   1.3562191  1.0541598   1.287  0.19829
## difdemshare4   1.5098525  0.5606320   2.693  0.00709 **
## rdifdemshare   0.3384021  0.1788422   1.892  0.05850 .
## rdifdemshare2 -0.1915849  0.8082619  -0.237  0.81264
## rdifdemshare3 -0.8374795  1.3203750  -0.634  0.52592
## rdifdemshare4 -1.8314914  0.6876047  -2.664  0.00774 **
## right          0.0909352  0.0114185   7.964 1.87e-15 ***
## demofficeexp  -0.0025256  0.0004986  -5.066 4.15e-07 ***
## othofficeexp  -0.0020931  0.0008129  -2.575  0.01004 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1371 on 9162 degrees of freedom
##   (2621 observations deleted due to missingness)
## Multiple R-squared:  0.6715, Adjusted R-squared:  0.6711
## F-statistic:  1703 on 11 and 9162 DF,  p-value: < 2.2e-16
```

```
column4 <- lm(update(model, "~.+demelectexp+othelectexp"), data=data)
summary(column4)
```

```
##
## Call:
```

```
## lm(formula = update(model, "~.+demelectexp+othelectexp"), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89242 -0.06414  0.00351  0.07355  0.72777
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.4509345  0.0085242  52.900  < 2e-16 ***
## difdemshare    0.1684626  0.1346492   1.251  0.21092
## difdemshare2  -0.1481654  0.6269933  -0.236  0.81320
## difdemshare3   1.3620442  1.0541033   1.292  0.19634
## difdemshare4   1.5115268  0.5605986   2.696  0.00702 **
## rdifdemshare   0.3354580  0.1788232   1.876  0.06070 .
## rdifdemshare2 -0.2026518  0.8081579  -0.251  0.80201
## rdifdemshare3 -0.8364825  1.3202519  -0.634  0.52637
## rdifdemshare4 -1.8369253  0.6875666  -2.672  0.00756 **
## right          0.0909384  0.0114173   7.965 1.85e-15 ***
## demelectexp   -0.0025757  0.0004886  -5.271 1.38e-07 ***
## othelectexp   -0.0018842  0.0007904  -2.384  0.01716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1371 on 9162 degrees of freedom
##   (2621 observations deleted due to missingness)
## Multiple R-squared:  0.6716, Adjusted R-squared:  0.6712
## F-statistic:  1703 on 11 and 9162 DF,  p-value: < 2.2e-16
```

```
column5 <- lm(update(model, "~.+demshareprev+demwinprev+demofficeexp+othofficeexp+demelectexp+othelecte
summary(column5)
```

```
##
## Call:
## lm(formula = update(model, "~.+demshareprev+demwinprev+demofficeexp+othofficeexp+demelectexp+othelec
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79619 -0.06210  0.00087  0.06676  0.63754
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.3133731  0.0110241  28.426  < 2e-16 ***
## difdemshare    0.4246366  0.1517819   2.798 0.005162 **
## difdemshare2   1.2295484  0.7087108   1.735 0.082804 .
## difdemshare3   3.2363198  1.1940593   2.710 0.006739 **
## difdemshare4   2.3054638  0.6354696   3.628 0.000288 ***
## rdifdemshare  -0.0477862  0.2001348  -0.239 0.811292
## rdifdemshare2 -1.5623918  0.9118695  -1.713 0.086687 .
## rdifdemshare3 -2.6939645  1.4846261  -1.815 0.069635 .
## rdifdemshare4 -2.6494509  0.7745157  -3.421 0.000628 ***
## right          0.0775140  0.0126472   6.129 9.36e-10 ***
## demshareprev   0.2975720  0.0131227  22.676  < 2e-16 ***
## demwinprev    -0.0064850  0.0069019  -0.940 0.347456
## demofficeexp  -0.0001356  0.0034794  -0.039 0.968907
```

```
## othofficeexp  -0.0002070  0.0040220  -0.051 0.958954
## demelectexp   -0.0029448  0.0033989  -0.866 0.386302
## othelectexp    0.0032936  0.0038736   0.850 0.395203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1296 on 6542 degrees of freedom
##   (5237 observations deleted due to missingness)
## Multiple R-squared:  0.7117, Adjusted R-squared:  0.7111
## F-statistic:  1077 on 15 and 6542 DF,  p-value: < 2.2e-16
```

```
demsharenext <- lm(demsharenext ~ demshareprev+demwinprev+demofficeexp+
     othofficeexp+demelectexp+othelectexp, data=na.omit(data))

demsharenextres <- predict(demsharenext)
col6data <- cbind(na.omit(data), demsharenextres)

column6 <- lm(update(model, "demsharenextres~."), data=col6data)
summary(column6)
```

```
##
## Call:
## lm(formula = update(model, "demsharenextres~."), data = col6data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61848 -0.04946 -0.00177  0.05333  0.60482
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.493203   0.007269  67.854  < 2e-16 ***
## difdemshare    0.625612   0.118099   5.297 1.21e-07 ***
## difdemshare2   1.321511   0.553704   2.387  0.01703 *
## difdemshare3   2.249311   0.933126   2.411  0.01596 *
## difdemshare4   1.402808   0.496313   2.826  0.00472 **
## rdifdemshare   0.434284   0.156290   2.779  0.00547 **
## rdifdemshare2 -3.851094   0.708729  -5.434 5.72e-08 ***
## rdifdemshare3  0.931693   1.159783   0.803  0.42181
## rdifdemshare4 -2.807518   0.604533  -4.644 3.48e-06 ***
## right         -0.004168   0.009895  -0.421  0.67360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1014 on 6548 degrees of freedom
## Multiple R-squared:  0.7021, Adjusted R-squared:  0.7017
## F-statistic:  1715 on 9 and 6548 DF,  p-value: < 2.2e-16
```

```
column7 <- lm(update(model, "difdemsharenext~.+ demwinprev+demofficeexp+othofficeexp+demelectexp+othele
summary(column7)
```

```
##
## Call:
## lm(formula = update(model, "difdemsharenext~.+ demwinprev+demofficeexp+othofficeexp+demelectexp+othel
##     data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81074 -0.06843  0.00311  0.06649  0.98459
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.043665   0.011757   3.714 0.000206 ***
## difdemshare    0.487552   0.181991   2.679 0.007403 **
## difdemshare2   1.333946   0.849786   1.570 0.116523
## difdemshare3   1.794364   1.431389   1.254 0.210039
## difdemshare4   0.897761   0.761316   1.179 0.238352
## rdifdemshare   0.023430   0.239969   0.098 0.922223
## rdifdemshare2 -2.321091   1.093257  -2.123 0.033783 *
## rdifdemshare3 -1.072422   1.779791  -0.603 0.546826
## rdifdemshare4 -1.074403   0.928023  -1.158 0.247016
## right          0.078791   0.015165   5.196  2.1e-07 ***
## demwinprev    -0.175078   0.007364 -23.775  < 2e-16 ***
## demofficeexp  -0.001843   0.004172  -0.442 0.658714
## othofficeexp  -0.007823   0.004820  -1.623 0.104604
## demelectexp   -0.003060   0.004076  -0.751 0.452765
## othelectexp    0.011258   0.004641   2.426 0.015308 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1554 on 6543 degrees of freedom
##   (5237 observations deleted due to missingness)
## Multiple R-squared:  0.1265, Adjusted R-squared:  0.1246
## F-statistic: 67.68 on 14 and 6543 DF,  p-value: < 2.2e-16
```

```
column8 <- lm(update(model, "demshareprev~.+ demwinprev+demwinprev+demofficeexp+othofficeexp+demelectexp
summary(column8)
```

```
##
## Call:
## lm(formula = update(model, "demshareprev~.+ demwinprev+demwinprev+demofficeexp+othofficeexp+demelecte
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87670 -0.06901  0.00235  0.07401  0.61594
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.3642752  0.0078010  46.696  < 2e-16 ***
## difdemshare   -0.4319814  0.1175950  -3.673 0.000241 ***
## difdemshare2  -1.7536736  0.5382426  -3.258 0.001126 **
## difdemshare3  -0.5614989  0.8915255  -0.630 0.528829
## difdemshare4   0.6644616  0.4695062   1.415 0.157034
## rdifdemshare   0.2198975  0.1569633   1.401 0.161263
## rdifdemshare2  2.9837141  0.7072918   4.219 2.48e-05 ***
## rdifdemshare3 -0.2736101  1.1329840  -0.241 0.809177
## rdifdemshare4 -0.5897849  0.5864380  -1.006 0.314583
## right          0.0158990  0.0101125   1.572 0.115934
## demwinprev     0.2373865  0.0049072  48.376  < 2e-16 ***
## demofficeexp   0.0027142  0.0027590   0.984 0.325262
```

```
## othofficeexp   0.0069370  0.0030071   2.307 0.021085 *
## demelectexp   -0.0002418  0.0026989  -0.090 0.928612
## othelectexp   -0.0074310  0.0028881  -2.573 0.010098 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1208 on 9160 degrees of freedom
##   (2620 observations deleted due to missingness)
## Multiple R-squared:  0.7366, Adjusted R-squared:  0.7362
## F-statistic:  1829 on 14 and 9160 DF,  p-value: < 2.2e-16
```

Testing the Validity of Regression Discontinuity Designs

There are several diagnostics that can be used to test the validity of RD designs. These diagnostics are used to show whether the observations around the cutoff point are comparable to one another or the density of the observations around the cutoff point is continuous.

Lee (2008) uses the former in Table 1.

```
LeeTable1 <- read.csv("table_one_final.csv")

LeeTable1 <- cbind(LeeTable1[,1], round(LeeTable1[,2:9], digits = 3))

colnames(LeeTable1) <- c("Variable", "All - winner" , "All - loser", "|Margin|<.5 - winner", "|Margin|<

LeeTable1
```

```
##          Variable All - winner All - loser |Margin|<.5 - winner
## 1   demsharenextm        0.698       0.347                0.629
## 2  demsharenextse        0.003       0.003                0.003
## 3  demsharenextsd        0.179       0.150                0.145
## 4     demwinnextm        0.909       0.094                0.878
## 5    demwinnextse        0.004       0.005                0.006
## 6    demwinnextsd        0.276       0.285                0.315
## 7   demshareprevm        0.681       0.368                0.607
## 8  demshareprevse        0.003       0.003                0.003
## 9  demshareprevsd        0.189       0.153                0.152
## 10    demwinprevm        0.889       0.109                0.842
## 11   demwinprevse        0.005       0.006                0.007
## 12   demwinprevsd        0.310       0.306                0.360
## 13  demofficeexpm        3.812       0.261                3.550
## 14 demofficeexpse        0.061       0.025                0.074
## 15 demofficeexpsd        3.766       1.293                3.746
## 16  othofficeexpm        0.245       2.876                0.350
## 17 othofficeexpse        0.018       0.054                0.025
## 18 othofficeexpsd        1.084       2.802                1.262
## 19   demelectexpm        3.945       0.464                3.727
## 20  demelectexpse        0.061       0.028                0.075
## 21  demelectexpsd        3.787       1.457                3.773
## 22   othelectexpm        0.400       3.007                0.528
## 23  othelectexpse        0.019       0.054                0.027
## 24  othelectexpsd        1.189       2.838                1.357
## 25            obs     3818.000    2740.000             2546.000
##    |Margin|<.5 - loser |Margin|<.05 - winner |Margin|<.05 - loser
```

```
## 1                0.372                0.542                0.446
## 2                0.003                0.006                0.006
## 3                0.124                0.116                0.107
## 4                0.100                0.681                0.202
## 5                0.006                0.026                0.023
## 6                0.294                0.458                0.396
## 7                0.391                0.501                0.474
## 8                0.003                0.007                0.008
## 9                0.129                0.129                0.133
## 10               0.118                0.501                0.365
## 11               0.007                0.027                0.028
## 12               0.317                0.493                0.475
## 13               0.304                1.658                0.986
## 14               0.029                0.165                0.124
## 15               1.390                2.969                2.111
## 16               2.808                1.183                1.345
## 17               0.057                0.118                0.115
## 18               2.775                2.122                1.949
## 19               0.527                1.949                1.275
## 20               0.032                0.166                0.131
## 21               1.550                2.986                2.224
## 22               2.943                1.375                1.529
## 23               0.058                0.120                0.119
## 24               2.805                2.157                2.022
## 25            2354.000              322.000              288.000
##    Parametric fit - winner Parametric fit - loser
## 1                    0.531                  0.454
## 2                    0.008                  0.008
## 3                       NA                     NA
## 4                    0.611                  0.253
## 5                    0.039                  0.035
## 6                       NA                     NA
## 7                    0.477                  0.481
## 8                    0.009                  0.010
## 9                       NA                     NA
## 10                   0.419                  0.416
## 11                   0.038                  0.039
## 12                      NA                     NA
## 13                   1.219                  1.183
## 14                   0.229                  0.145
## 15                      NA                     NA
## 16                   1.424                  1.293
## 17                   0.131                  0.170
## 18                      NA                     NA
## 19                   1.485                  1.470
## 20                   0.230                  0.151
## 21                      NA                     NA
## 22                   1.624                  1.502
## 23                   0.132                  0.174
## 24                      NA                     NA
## 25                3818.000               2740.000
```

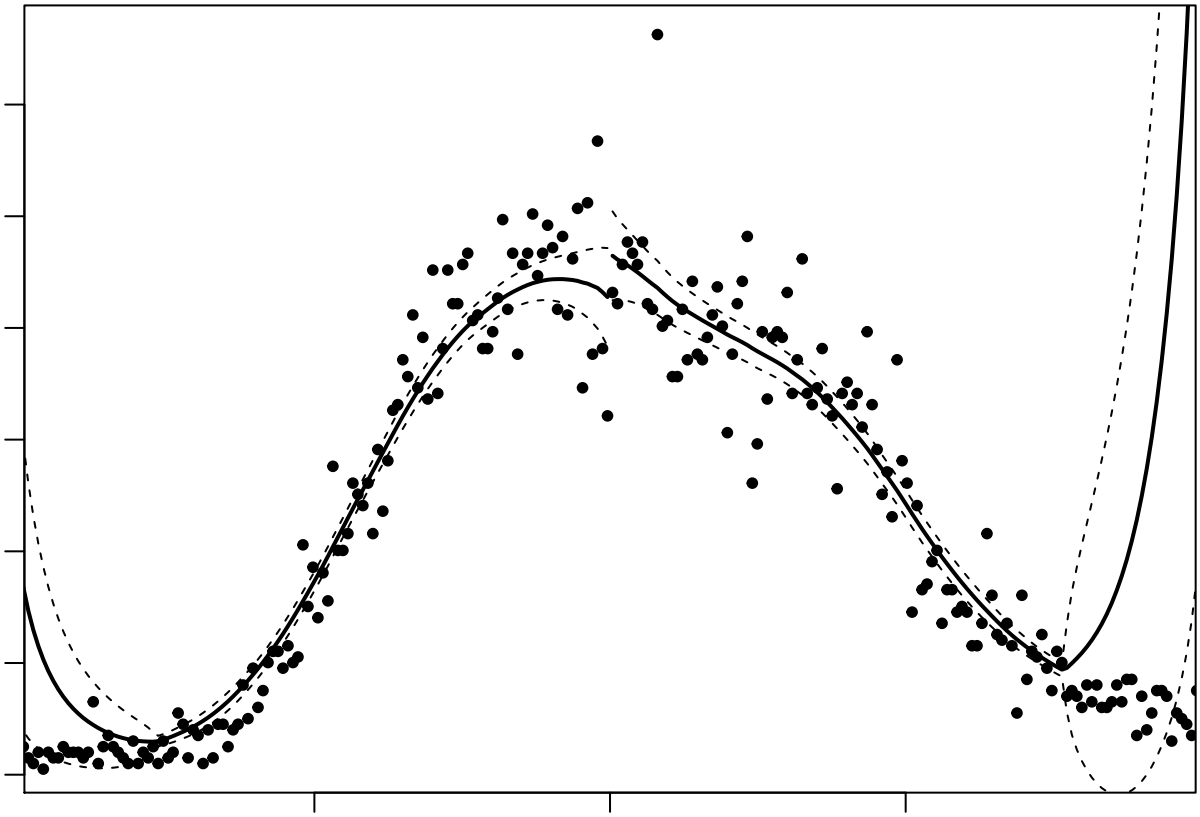
Lee uses this table to show that except for the first two variables, which are the outcome variables, the averages of all others become very similar to one another as the margin of victory becomes smaller. This

suggests that those candidates near the cutoff point are comparable to one another.

McCrary (2006) develops another test that observes the density of the running variable at the cutoff point. This test is based on the intuition that if there is no manipulation or sorting around the cutoff point, then the density should be continuous around the cutoff point. That is, the probability of observing an observation just above or below the cutoff should be around the same. Conversely, if observations are able to sort into the treatment (or control group), the density around the cutoff should be discontinuous.

```
par(mar=c(1,1,1,1))

DCdensity(data$difdemshare, 0)
```



```
## [1] 0.1398982
```

Alternatively, it is possible to perform balance tests by substituting the outcome variable of interest with different baseline covariates. This formally tests whether the mean of the outcome is statistically different on each side of the cutoff point.

It is also possible to plot these baseline covariates along the y-axis and the running variable on the x-axis. If there is no sorting around the cutoff, then there should be no visual evidence of a discontinuity, unlike with the outcome variable. Like a randomized experiment, the distribution of these baseline covariates should not change discontinuously at the threshold.

Fuzzy RD

Often, the treatment may be determined *partly* by whether the assignment variable passes a known cutoff point. This this different from the RD design discussed so far, when the assignment variable is completely determined past the cutoff. Reasons for this include imperfect compliance by program participants. This

setting is referred to as a "fuzzy" RD design versus a "sharp" RD design. In the former case, the probability of treatment does not jump from 0 to 1.

When this occurs, the jump in the relationship between the assignment variable and the outcome variable at the cutoff can no longer be interpreted as an average treatment effect. Instead, the treatment effect can be recovered by dividing the jump in the relationship between the assignment variable and the outcome variable by the fraction induced to take-up the treatment at the threshold.

This is akin to an instrumental variable approach with a dichotomous instrument or Wald estimator. In this case, all of the assumptions and interpretability of instrumental variable is applied to the fuzzy RD, including monotonicity and excludability.

Checklist for RD Design Implementation

This checklist is from Lee and Lemieux (2008).

1. To assess the possibility of manipulation of the assignment variable, show its distribution
2. Present the main RD graph using binned local averages
3. Graph a benchmark polynomial specification
4. Explore the sensitivity of the results to a range of bandwidths, and a range of order to the polynomial
5. Conduct a parallel RD analysis on the baseline covariates
6. Explore the sensitivity of the results to the inclusion of baseline covariates