790:676: Experimental Methods

Due: April 21st at 5pm on Canvas

Select Experimental Concepts and Analyses

Instructions

You will submit, on Canvas, two documents (or a single compiled R Markdown or KnitR document):

- An R script (.R) file with your code. Follow the best practices by titling your script and using # comments to explain your steps. This code should be clean. I should be able to run your code to verify that the code produces the answers you write down. (If you have decided not to use R, please provide an equivalent log of the steps you took in SPSS or Stata.)
- A word document or pdf with answers to the written questions. This document should also have a title with your name on it.

You are allowed to use any and all resources to help you answer the questions with the exception of your colleagues' solutions. You may work together on the problem set, but every individual should write up his or her problem set completely independently.

Choose Your Own Adventure

You can complete problem 1, problem 2, or problems 1 and 2.

1. **Encouragement Design Analysis**: We will replicate a portion of the analyses in Gwyneth McClendon's 2014 *AJPS* article "Social Esteem and Participation in Contentious Politics: A Field Experiment at an LGBT Pride Rally." You can download the .dta file at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/21992`. I recommend you choose the Stata format.

   The researcher randomly assigned subjects who were on the listserv of an LGBT advocacy organization to one of three groups, all of which received an email invitation to participate in a rally to recognize the repeal of "Don't Ask Don't Tell" and to demonstrate support for marriage equality in the state of New Jersey. In the *information-only* condition, individuals were only informed about the goals of the event and its location. We will treat this as a control condition. In the *newsletter* condition, they received the same information and were additionally told that participants are worthy of admiration and that their names would be listed in the monthly newsletter if they attended so that their attendance could be "celebrated" by other members of the organization. In the *Facebook* condition, they were given the same information but invited to post photos from the event to the group's Facebook page so that other group members could "like" the photos, rather than having their names published in a newsletter (this was meant to make sure that any treatment effect was due to social esteem more generally and not specific to newsletters). Because both of these conditions have to do with social esteem, we will eventually pool them into a "social esteem" condition.

Because people often ignore mass emails, we will consider being sent an email as an encouragement for the successful communication of the treatment. We can observe whether the individual read the email carefully through online tracking of who clicked on a link to display images within the email. We will consider opening this link <u>in the esteem email</u> to be the receipt of treatment. (Note: opening any email is not receipt of treatment. Only opening the esteem email represents receipt of treatment.) The dependent variable in this analysis is participation in the rally.

The dataset contains one observation per respondent, with key variables as follows:

- `facebook, newsletter, information`: 1 if in the experimental condition, 0 otherwise
- `attended`: observed outcome where 1 if attended rally, 0 otherwise
- `opened`: 1 if respondent closely read the email, 0 otherwise

(a) Load the data. Create a new variable `socesteem` in which you pool the treatment conditions such that there is a single treatment indicator where 1 = in the facebook or newsletter condition, and 0= information condition. What proportion of respondents were in an esteem condition (facebook or newsletter)?

Solution:

```
library(foreign)
lgb <- read.dta("lgbtrally_anondata.dta")

lgb$socesteem <- ifelse(lgb$newsletter == 1 | lgb$facebook == 1, 1, 0)

mean(lgb$socesteem, na.rm = T)

## [1] 0.6663011
```

About two-thirds in the social esteem condition.

(b) Create a variable that indicates if respondents in the social esteem condition closely read their email to receive the esteem treatment.

Solution:

```
## Alternative 1
lgb$treatreceived <- lgb$opened * lgb$esteem

## Alternative 2
lgb$treatreceived <- ifelse(lgb$socesteem == 1 & lgb$opened == 1, 1, 0)
```

(c) In encouragement designs, sometimes we are interested in the following equation, where $Y_i(d)$ is the potential outcome where d_i = receipt of treatment (0 or 1) for unit i, and we have $d_i(z)$ where z_i refers to experimental assignment (0 or 1). Please describe what quantity of interest this expression refers to: $E[\{Y_i(1) - Y_i(0)\}|d_i(1) - d_i(0) = 1]$ conceptually and in this particular application of evaluating the effect of promising social esteem on rally attendance. Assume monotonicity.

$$E[\{Y_i(1) - Y_i(0)\}|d_i(1) - d_i(0) = 1] = \frac{E[\overbrace{Y_i(1, d_i(1))}^{\text{encouraged}} - \overbrace{Y_i(0, d_i(0))}^{\text{not encouraged}}]}{E[d_i(1) - d_i(0)]}$$

Solution:

The quantity describe is the CACE. It is the average treatment effect among compliers. In this case, it is the effect of promising esteem on attending the rally among those who opened the email to view the esteem language if and only if they were in the esteem experimental condition.

(d) The following is proposed as an estimator for the quantity of interest ($\hat{E}[\{Y_i(1)-Y_i(0)\}|d_i(1)-d_i(0) = 1]$) where N is the sample size and Y_i is the observed outcome. Use this estimator (or an equivalent estimator) in this application to estimate and interpret the quantity of interest. For now, you do not need to include uncertainty.

$$\hat{E}[\{Y_i(1) - Y_i(0)\}|d_i(1) - d_i(0) = 1] = \frac{\frac{\sum_{i=1}^N Y_i z_i}{\sum_{i=1}^N z_i} - \frac{\sum_{i=1}^N Y_i (1-z_i)}{\sum_{i=1}^N (1-z_i)}}{\frac{\sum_{i=1}^N d_i z_i}{\sum_{i=1}^N z_i} - \frac{\sum_{i=1}^N d_i (1-z_i)}{\sum_{i=1}^N (1-z_i)}}$$

Solution:

```
Di <- lgb$treatreceived
Zi <- lgb$esteem
Yi <- lgb$attended

## Three alternatives
ITT1 <- sum(Yi * Zi, na.rm = T)/sum(Zi, na.rm = T) - sum(Yi * (1 - Zi), na.rm = T)/sum(1
    Zi, na.rm = T)
ITT2 <- mean(lgb$attended[lgb$esteem == 1], na.rm = T) - mean(lgb$attended[lgb$esteem ==
    0], na.rm = T)
ITT3 <- lm(Yi ~ Zi)
coef(ITT3)[2]

##         Zi
## 0.01278561

ITTd1 <- sum(Di * Zi, na.rm = T)/sum(Zi, na.rm = T) - sum(Di * (1 - Zi), na.rm = T)/sum(1
    Zi, na.rm = T)
ITTd2 <- mean(lgb$treatreceived[lgb$esteem == 1], na.rm = T) - mean(lgb$treatreceived[lgb
    0], na.rm = T)
ITTd3 <- lm(Di ~ Zi)
coef(ITTd3)[2]

##         Zi
## 0.07860082

## Take ratio to get CACE estimate
ITT1/ITTd1

## [1] 0.162665
```

The estimate of the CACE is 0.16.

(e) Estimate the $\widehat{ITT}$ for this application. In addition, carry out a two-stage least squares procedure to estimate the CACE. Report the CACE, standard error, and provide a conclusion about the significance of the effect.

Solution:

```
## ITT was the numerator above
ITT1

## [1] 0.01278561

## CACE should be equivalent to the ratio calculated above
library(AER)

## Loading required package:  car
## Loading required package:  carData
## 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:  sandwich
## Loading required package:  survival

cace <- ivreg(attended ~ treatreceived | esteem, data = lgb)
summary(cace)

##
## Call:
## ivreg(formula = attended ~ treatreceived | esteem, data = lgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17992 -0.01726 -0.01726 -0.01726  0.98274
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.017256   0.004292   4.021 5.92e-05 ***
## treatreceived 0.162665   0.066890   2.432   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1497 on 3645 degrees of freedom
## Multiple R-Squared: 0.1078,Adjusted R-squared: 0.1076
## Wald test: 5.914 on 1 and 3645 DF,  p-value: 0.01507

## optional
coeftest(cace, vcovHC(cace))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   0.0172555  0.0037359  4.6188 3.992e-06 ***
## treatreceived 0.1626650  0.0628014  2.5901  0.009632 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The CACE is significant at the $p < 0.05$ level. This suggests we can reject the null hypothesis of no difference among compliers, and that promises of esteem among the LGBT in-group may influence attendance at pride rallies.

(f) Gerber and Green note that one way to get approximate standard errors for the $\widehat{CACE}$, is using equation 5.29: $SE(\widehat{CACE}) = \frac{SE(\widehat{ITT})}{ITT_d}$. Estimate the standard error of the CACE using this approach.

Solution:

```
## Can use standard error from regression equation above for ITT
summary(ITT3)$coefficient[2, 2]/ITTd2

## [1] 0.07076568
```

(g) Provide one example of something that might violate the required assumptions for identifying the ITT or identifying the CACE in this application.

Solution: If respondents forward the emails they receive it might violate the non-interference

assumption by treating and affecting the potential outcomes of respondents in the Control condition. The authors attempt to address this concern on pg. 283.

(h) An alternative way to estimate the complier average causal effect might be to treat anyone who opened a social esteem email as receiving the treatment and anyone who opened a non-esteem email as receiving a “placebo" treatment. We could estimate the complier average causal effect by subsetting our data to those who opened emails and then calculating the average treatment effect among this subgroup. Please calculate the average treatment effect among compliers using this approach and report your conclusion about whether the effect is significant. In addition, include one critique of this approach (e.g., What do you have to assume in order to be able to subset the data in this way?).

Solution:

```
openers <- subset(lgb, opened == 1)
caceplacebo <- lm(attended ~ socesteem, data =openers)
summary(caceplacebo)

##
## Call:
## lm(formula = attended ~ socesteem, data = openers)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.26178 -0.26178 -0.09474 -0.09474  0.90526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09474    0.04087   2.318 0.021148 *
## socesteem    0.16704    0.05001   3.340 0.000949 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3983 on 284 degrees of freedom
## Multiple R-squared:  0.0378,Adjusted R-squared:  0.03442
## F-statistic: 11.16 on 1 and 284 DF,  p-value: 0.0009486
```

We might think that the types of people who open email images differ depending on if it is an information or social esteem email.

2. **Let's get jumpy.** We will use a regression discontinuity design to estimate whether female state legislative incumbents have more trouble fundraising than do male legislative incumbents. Observations consist of information on all U.S. state legislative races in which a male and female candidate ran against each other. This is based on the article:

   Barber, Michael, Daniel Butler, and Jessica Preece. 2016. "Gender Inequalities in Campaign Finance.`http://dx.doi.org/10.1561/100.00015126` Quarterly Journal of Political Science 11(2): 219-248.

   The independent variables are the winning (incumbent) candidate's gender and the vote share the female candidate received in the election.

   The outcome variable is the amount the winning candidate in a given election (which could be male or female) raised in their *next* election cycle. We will analyze their data to help determine if male winning candidates raise more money than female winning candidates in their subsequent election cycles when they are incumbents.

   The data file you will use is 'fundsub.csv', a CSV data file. The file contains the following variables, which we will make use of in this exercise:

Name	Description
`male.winner`	Coded 1= a man won the election and 0= a woman won the election in the most recent election in a legislative district (time t)
`female.margin`	The two-candidate vote share that the woman candidate won in the most recent election in a legislative district in time t (if above 50 a woman won, if a below 50, a man won)
`total.candidate.raised`	Amount raised by the winning candidate (i.e., the incumbent) in the subsequent election cycle (time t+1)
`log.total.candidate.raised`	The log of the total amount raised by the winning candidate (i.e., the incumbent) in the subsequent election cycle (time t +1)
`log.total.dist.last`	The log of the total spending in the district in the previous election cycle (time t - 1)

   (a) Load the data and assign it to an object `fund`. Report how many rows are in the dataframe. These represent the number of elections.
   Solution:

```
fund <- read.csv("fundsub.csv")
nrow(fund)

## [1] 2869
```

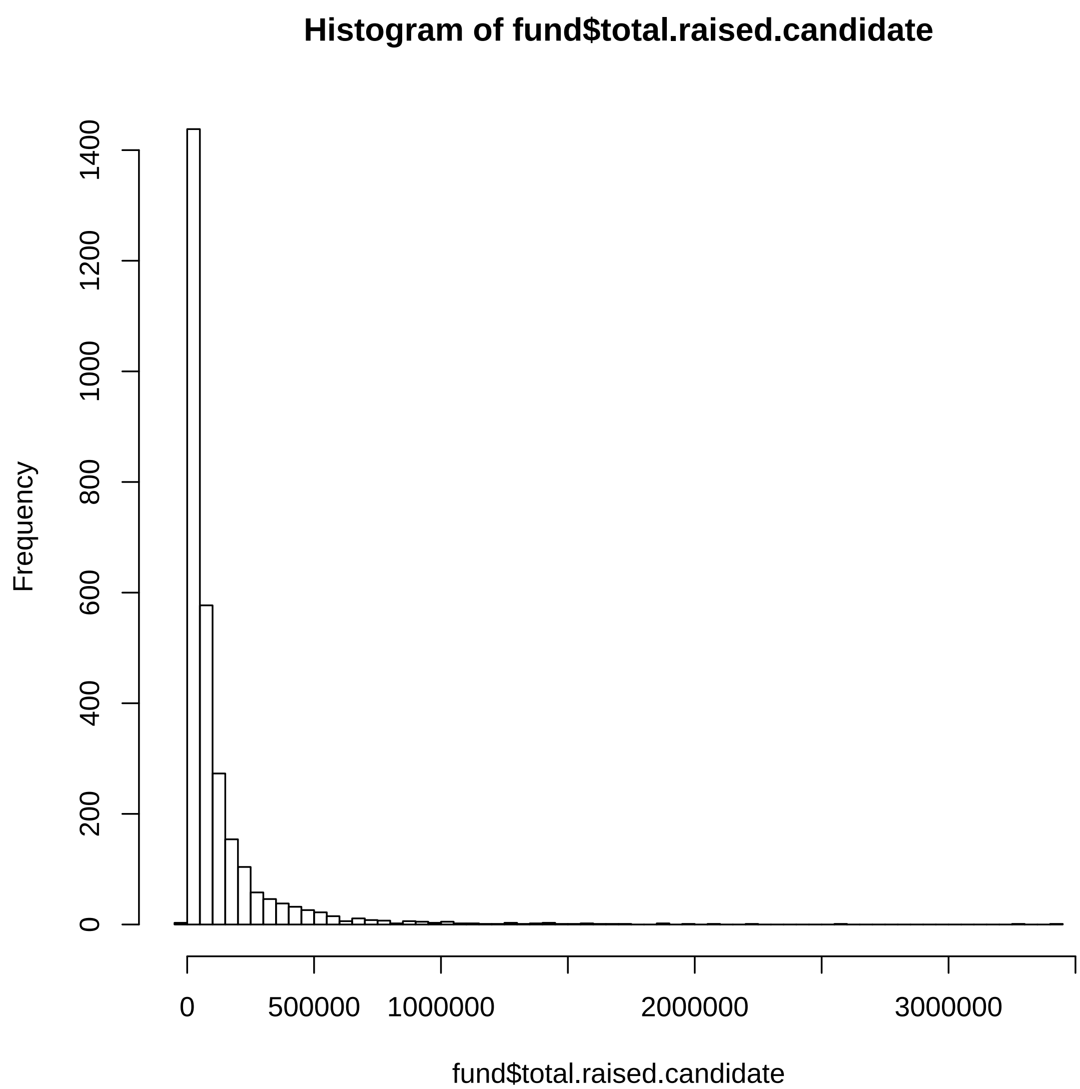
   (b) How many winners were male, and how many were female in our data?
   Solution:

```
table(fund$male.winner)

##
##    0    1
## 1445 1424
```

(c) Fundraising data are often skewed, whereby some candidates raise very little or zero money at the legislative level, and some candidates raise A LOT of money. Let's visualize this skew in the total amount of money raised using a histogram in R with 100 breaks. The syntax is: `hist(data$variable, breaks=100)`.

Solution:

```
hist(fund$total.raised.candidate, breaks = 100)
```



(d) Because the data is heavily skewed, researchers used the log of the variable instead of the raw variable. This makes our data look more "normal" because it compresses the scale. You may have seen versions of logarithmic scales in the news recently in the reporting of COVID-19 cases. Example: `https://datausa.io/coronavirus`

To see how a log compresses our scale, let's look at our data. For example, the first two observations in our data are:

```
fund$total.raised.candidate[1]

## [1] 71947

fund$total.raised.candidate[2]

## [1] 10784
```

By taking the log, we get

```
# base of the log is e (exp) here, sometimes learned as ln
log(fund$total.raised.candidate[1])

## [1] 11.18369

log(fund$total.raised.candidate[2])

## [1] 9.285819
```

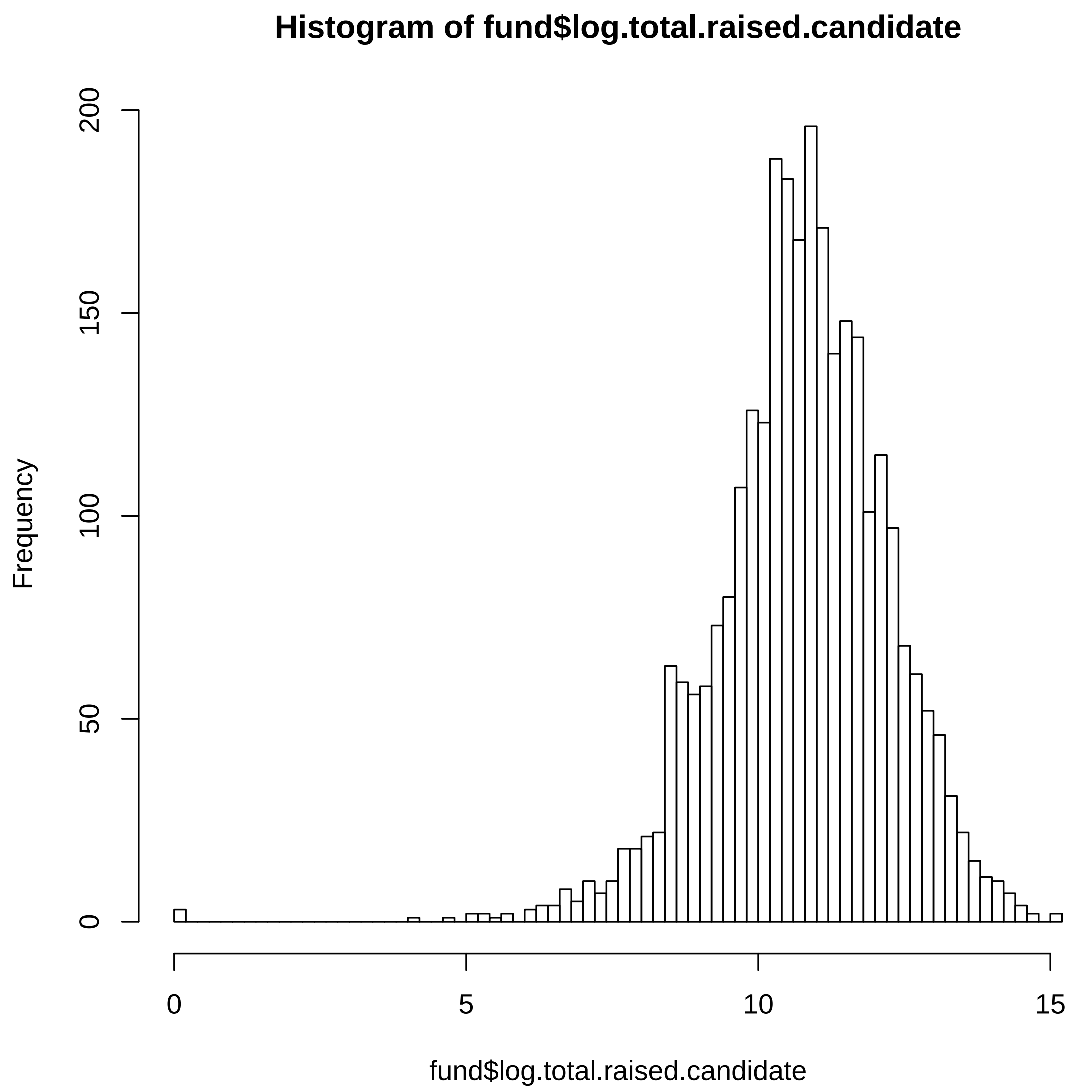
You can always revert back to the raw money scale by "exponentiating" a log. For example:

```
exp(fund$log.total.raised.candidate[1])

## [1] 71947

exp(fund$log.total.raised.candidate[2])

## [1] 10784
```

Repeat the previous exercise and create a histogram using `log.total.raised.candidate`. Note any visual differences in the figure.

Solution:

```
hist(fund$log.total.raised.candidate, breaks = 100)
```



We see the log makes the distribution appear more normal and less skewed.

(e) We are going to compare the amount of money raised among incumbents who are male vs. incumbents who are female using a regression discontinuity design. Let's replicate the first column of Table 1 in their paper. Following the researchers, we are going to restrict our analysis to compare candidates who won races that were competitive: within +/- 2 percentage points in vote share. We also want to remove ties, cases where the female margin is 50.

- To make our regression easier to interpret, we should also center our forcing variable on 50, such that when this variable is 0, that is our threshold of winning. Create a new variable called `forcing.variable` that does this. In addition, to match the authors, we actually want the variable to be in a direction where positive values mean that the male candidates were winners (and negative, female winners.
- Create a subset of your data that includes only elections where the female candidate margin was > 48 and < 52 and not equal to 50.
- Regress `log.total.raised.candidate` on our indicator variable for whether a male candidate wins, our new centered forcing variable, and the interaction between these variables. (Hint: this should match Table 1, column 1 in the paper)

- What is the RDD estimate? Interpret this is 1-2 sentences. Why do we say this is a local average treatment effect?

(f) To make the output easier to interpret, let's exponentiate the log. To do this, we can use the predict command in R to estimate the expected dollar amount male candidates would raise at the point of discontinuity and compare it to the amount of dollars female candidates are expected to raise at the same point.

In a regression model, you can estimate the output at any point by supplying the desired independent variable values to the predict command. For example, if I wanted to predict the amount female winners would raise if they won by 1, I would type:

- `predict(fit, data.frame(male.winner = 0, forcing.variable = -1))`.
- The first input of the predict command is the name of the regression model object. Then, inside the `data.frame`, you supply values for each of your independent variables.
- E.g., when male.winner is 0, a female has won.

Once you get that output, you can use `exp()` to exponentiate it into dollars and answer the following questions:

- What is the estimated difference between the amount of dollars male and female winners are expected to raise at the point of discontinuity?
- What is the percent increase in fundraising that this represents? $\frac{\text{(Male - female expected fundraising)}}{\text{Expected female fundraising}} * 100$

Solution:

```
fund$forcing.variable <- -1 * (fund$female.margin - 50)

rsub <- subset(fund, female.margin > 48 & female.margin < 52 & female.margin != 50)

fit <- lm(log.total.raised.candidate ~ male.winner * forcing.variable, data = rsub)

summary(fit)

##
## Call:
## lm(formula = log.total.raised.candidate ~ male.winner * forcing.variable,
##     data = rsub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2401 -0.9995  0.1598  0.9429  3.9878
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  11.06441    0.24359  45.421   <2e-16 ***
## male.winner                   0.81377    0.35173   2.314   0.0213 *
## forcing.variable              0.09145    0.22043   0.415   0.6785
## male.winner:forcing.variable -0.66338    0.31187  -2.127   0.0342 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.522 on 307 degrees of freedom
```

```
## Multiple R-squared:  0.03337,Adjusted R-squared:  0.02392
## F-statistic: 3.533 on 3 and 307 DF,  p-value: 0.0152

##
f.raised <- exp(predict(fit, data.frame(male.winner = 0, forcing.variable = 0)))
m.raised <- exp(predict(fit, data.frame(male.winner = 1, forcing.variable = 0)))

f.raised - m.raised

##         1
## -80230.65

(m.raised - f.raised)/(f.raised) * 100

##        1
## 125.6404
```

We see a 125% increase in fundraising, going from a female to male winner. This effect is significant and represents the LATE as it is an average treatment effect specifically at the point of discontinuity.

(g) Let's make a plot.

- Repeat the previous regression RDD estimate, but this time with +/- 4 (also removing ties). No need to exponentiate, just note the coefficient of interest.
- Next: Use the `rdrobust()` function to estimate the rdd effect. This time, again, subset the data to get rid of ties (where female.margin=50), but instead of subsetting our data to a specific bandwidth, allow the function to choose the optimal bandwidth using Mean-Squared Error optimal bandwidth (`mserd`). Also specify a "uniform" weight.
  Note: you can get a summary of the result of rdrobust by using `summary(fit.rdr)` where `fit.rdr` just represents whatever you called your rdrobust object. The summary will show you the bandwidth estimate (BW est.). `fit.rdr$coef` will return for you the RDD estimate, and `fit.rdr$ci` will supply you with the confidence intervals.
- Run one additional RDD specification of your choosing. You can adjust the bandwidth, the weighting (e.g., kernel), the polynomial, etc. If you want, you could explore `rdplot()` to help make this decision.
- Create a plot with the original RDD estimate from the paper and associated 95% confidence interval (Hint: you can use `confint(fit)` on the regression model object.), the estimate based on +/- 4, and the 2 new estimates based on the rdrobust function.
  You can just use the "conventional" rdrobust estimates for this application, but you can read more in the documentation and potentially opt to use the bias corrected estimates in your own research.
- Add text to the plot to label the estimates.
- Interpret the results. If you were the researcher, how would you decide the specification to use?

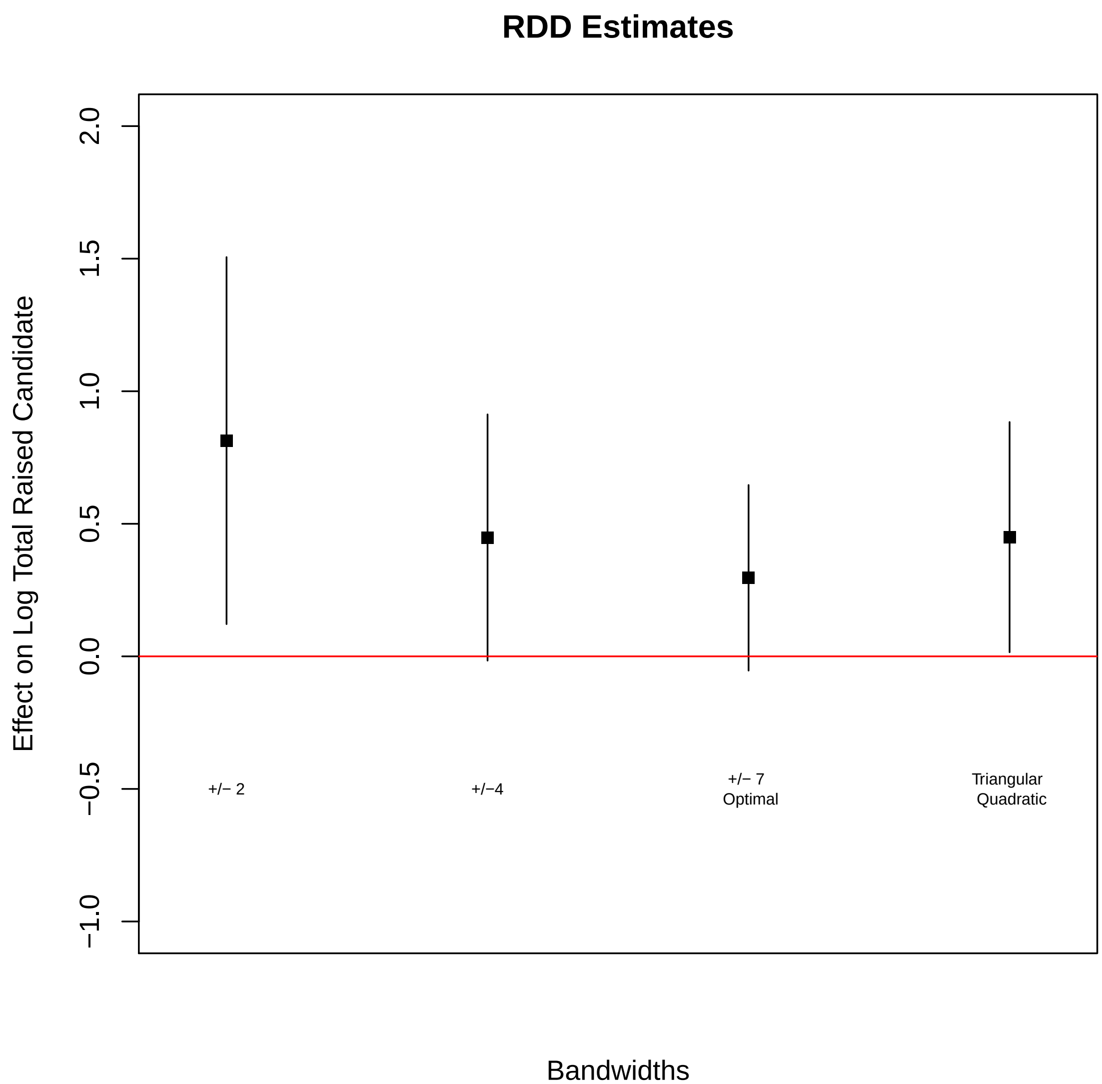
Solution:

```
rsub2 <- subset(fund, female.margin > 46 & female.margin < 54 & female.margin !=
    50)
```

```r
fit2 <- lm(log.total.raised.candidate ~ male.winner * forcing.variable, data = rsub2)

summary(fit2)
```

```
##
## Call:
## lm(formula = log.total.raised.candidate ~ male.winner * forcing.variable,
##     data = rsub2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7618 -0.9035  0.1302  0.9865  4.0788
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  11.12937    0.16291  68.314  < 2e-16 ***
## male.winner                   0.44827    0.23641   1.896  0.05841 .
## forcing.variable              0.09035    0.07622   1.185  0.23632
## male.winner:forcing.variable -0.29087    0.10546  -2.758  0.00599 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.465 on 606 degrees of freedom
## Multiple R-squared:  0.01922,Adjusted R-squared:  0.01436
## F-statistic: 3.958 on 3 and 606 DF,  p-value: 0.008223
```

```r
library(rdrobust)
rdprep <- subset(fund, female.margin != 50)
rdb <- rdrobust(y = rdprep$log.total.raised.candidate, x = rdprep$forcing.variable,
    kernel = "uniform", p = 1, bwselect = "mserd")

rdb2 <- rdrobust(y = rdprep$log.total.raised.candidate, x = rdprep$forcing.variable,
    kernel = "triangular", p = 2, bwselect = "mserd")


plot(x = 1:4, y = c(coef(fit)["male.winner"], coef(fit2)["male.winner"], rdb$coef[1],
    rdb2$coef[1]), ylim = c(-1, 2), xlim = c(0.8, 4.2), pch = 15, xaxt = "n", xlab = "Ban
    main = "RDD Estimates", ylab = "Effect on Log Total Raised Candidate")
lines(c(1, 1), confint(fit)[2, ])
lines(c(2, 2), confint(fit2)[2, ])
lines(c(3, 3), rdb$ci[1, ])
lines(c(4, 4), rdb2$ci[1, ])

abline(h = 0, col = "red")
text(1:4, -0.5, c("+/- 2", "+/-4", "+/- 7 \n Optimal", "Triangular \n Quadratic"),
    cex = 0.6)
```



We see a positive local average treatment effect of male winners on fundraising in each specification, but whether this is significant depends on the specification. The choice should consider tradesoffs in bias created by expanding the bandwidth vs. variance, which may be affected by changing the sample size.

(h) One assumption of regression discontinuity is that treatment assignment is "as-if" random around the cutpoint (discontinuity). This means that any pre-treatment covariates should be continuous at the cutpoint, just as pre-treatment covariates should be balanced across conditions in an experiment. Any evidence that there is a significant discontinuity effect on a pre-treatment covariate would be evidence that perhaps we do not have an as-if random design. Let's do one test of this. Let's repeat our original RDD model of a +/-2 bandwidth, but now let's have the outcome be the total spending in the district in a previous election in t-1. Run the model and interpret the results in light of the design assumptions.

Solution:

```r
fit.sorting <- lm(log.total.dist.last ~ male.winner * forcing.variable, data = rsub)
summary(fit.sorting)
```

```
##
## Call:
## lm(formula = log.total.dist.last ~ male.winner * forcing.variable,
##     data = rsub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5764 -0.7318  0.0079  0.9009  3.4478
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   11.13381    0.24941  44.640   <2e-16 ***
## male.winner                    0.20155    0.36826   0.547    0.585
## forcing.variable               0.08924    0.22699   0.393    0.695
## male.winner:forcing.variable -0.27047    0.32816  -0.824    0.411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.434 on 247 degrees of freedom
##   (60 observations deleted due to missingness)
## Multiple R-squared:  0.004391,Adjusted R-squared:  -0.007702
## F-statistic: 0.3631 on 3 and 247 DF,  p-value: 0.7797
```

We see a null effect here, which helps lend support that this "pre-treatment" covariate is balanced around the discontinuity. If, instead, it were significant, that might call into question that a male vs. female winner is "as-if" random in close elections.