Problem Set 6: Regression Discontinuity

Claire Duquennois

Name: Xiang Li

1. Empirical Analysis using Data from Manacorda, Miguel, & Vigorito (2011, American Economic Journal: Applied Economics)

This exercise uses data from Manacorda, Miguel, & Vigorito's paper, "Government Transfers and Political Support," published in the *American Economic Journal: Applied Economics* in 2011. This paper studies how receipt of a government anti-poverty cash transfer changes how beneficiary households support and view the government.

2. Finding the data

The data can be found on Edward Miguel's faculty website. Download and extract the contents from the `Government_Transfers_replication.zip` file.

3. Set up and constructing the data

The original data used in the paper is confidential. The authors instead provide the `reg_panes.dta` data file which is anonymized and created from the original data.

3.1 Question: Loading the Packages

Load any R packages you will be using: **Code:**

```
library("haven")
library("dplyr")

##
## 载入程辑包：'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library("stargazer")
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summ
ary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargaz
er

library("lfe")

## 载入需要的程辑包：Matrix

## Warning: 程辑包'Matrix'是用R版本4.3.2 来建造的

#install.packages("statar")
library("statar")

## Warning: 程辑包'statar'是用R版本4.3.2 来建造的

library("ggplot2")

## Warning: 程辑包'ggplot2'是用R版本4.3.2 来建造的
```

3.2 Question: Open the `reg_panes.dta` file. To complete this problem set you will need the following variables from this data file:

Name	Description
aprobado	Ever received PANES 2005-2007
untracked07	Untracked in 2007
h_89	Supports current government 2007 [1 to 3]
hv34	Supports current government 2008 [1 to 3]
ind_reest	Predicted Income
newtreat	PANES eligibility
geo	Geographic locality
bl_medad	Mean age
lnbl_ytoth_pc	Log per capita income
bl_hhsize	Mean household size
bl_meduc	Mean education
missbl_medad	Missing mean age
misslnbl_ytoth_pc	Missing log per capita income
missbl_hhsize	Missing mean household size
missbl_meduc	Missing mean education
sexo	Respondent is female
edad	Respondent age

Name	Description
aniosed07	Education in 2007
misssexo	Missing gender
missedad	Missing age
missaniosed	Missing education

Drop all other variables. If needed, give the variables you are keeping more intuitive names.

Code:

```
reg_panes <- read_dta("reg_panes.dta")

panes <- select(reg_panes, aprobado, untracked07, h_89, hv34, ind_reest,
 newtreat, geo, bl_medad, lnbl_ytoth_pc, bl_hhsize, bl_meduc, missbl_me
dad, misslnbl_ytoth_pc, missbl_hhsize, missbl_meduc, sexo, edad, aniose
d07, misssexo, missedad, missaniosed)
```

3.3 Question: The data as downloaded will require that you clean the variables of interest and construct a new dataset to generate the graphs. Start by generating the following cleaned variable:

-An indicator for receiving PANES that is NA if a respondent is untracked in 2007

Code:

```
panes$aprobado <- ifelse(panes$untracked07 == 1, NA, panes$aprobado)
```

3.4 Question: We are going to re-scale the variables that indicate support for the current government so that responses range from 0 to 1. To do this, tabulate the current variable to see how it is distributed and then generate a variable that will be NA if it is currently coded as 9, 0 if currently 2, 0.5 if currently 1 and 1 if currently 3. Do this for both the 2007 and 2008 variable.

Note: This is how the authors modify this variable in their code. It seems counter intuitive and does not correspond to the description of how this variable is coded in the survey questionnaire as reported in their appendix though it does correspond to their discussion in footnote 12. My guess is the transcription/translation of the survey question is incorrect.

Code:

```
panes$"2007rescale" <- ifelse(panes$h_89 == 9, NA,
                                  ifelse(panes$h_89 == 2, 0,
                                          ifelse(panes$h_89 == 1, 0.5,
                                              ifelse(panes$h_89 == 3, 1,
 NA))))

panes$"2008rescale" <- ifelse(panes$hv34 == 9, NA,
                                ifelse(panes$hv34 == 2, 0,
                                       ifelse(panes$hv34 == 1, 0.5,
                                              ifelse(panes$hv34 == 3, 1,
 NA))))
```

3.5 Question: Generate a variable that is the square of predicted income.

Code:

```
panes$predincome_sq <- panes$ind_reest^2
```

4. We start by reproducing the main figures (2,3,and 4) of the paper as good figures are key to any regression discontinuity paper.

4.1 Question: The data consists of over 3000 observations. How many points are plotted on these figures? How should we interpret the y axis? What does each point below the threshold represent? What does each point above the threshold represent?

Answer: Each cell contains approximatelythe same number of observations (43 households). The y-axis shows either: 1. The proportion receiving PANES benefits (Figure 2) 2. The level of political support for the government (Figures 3 and 4) For points below the eligibility threshold (score < 0)(left): Each point represents the average for households within that score bin who were eligible for PANES For points above the threshold (score ≥ 0)(right): Each point represents the average for households within that score bin who were ineligible for PANES

4.2 Question: Why is the number of points above the threshold different from the number below?

Answer: “Since there are approximately twice as many households to the left of the eligibility threshold (i.e., the PANES eligible households) as to the right, we present twice as many cells for eligible households (30) as for ineligible ones (15), such that each cell contains approximately the same number of observations (43 households).” So they chose to have 30 bins intervals below the threshold and 15 bins above the threshold. This makes it so each bin has around 43 households. If they had used the same number of bins on both sides, the bins would be very unevenly sized.

4.3 Question: Replicating these figures will require restructuring our data and calculating the values that are plotted. Generate a variable that will indicate the percentile group the observation is in. Note the difference in the number of percentile groups above and below the threshold.

Note: you may find the `xtile` function in R useful.

Code:

```
panes$group1 <- ifelse(panes$ind_reest< 0, panes$ind_reest, NA)
panes$group2 <- ifelse(panes$ind_reest >= 0, panes$ind_reest, NA)

panes$percentile1 <- xtile(panes$group1, n = 30)
panes$percentile2 <- xtile(panes$group2, n = 15)

panes$percentile2 <- panes$percentile2 + 30

panes$group <- rowSums(panes[, c("percentile1","percentile2")], na.rm =
 TRUE)
```

4.4 Question: For each of the percentile groups, calculate the mean of each of the variables we will use for plotting: predicted income, receipt of PANES, support for the government in 2007, and support for the government in 2008.

Code:

```
panes_mean <- panes %>%
  group_by(group) %>%
  dplyr::summarize(
    mean_predincome = mean(ind_reest, na.rm = TRUE),
    mean_receipt = mean(aprobado, na.rm = TRUE),
    mean_support_2007 = mean(h_89, na.rm = TRUE),
    mean_support_2008 = mean(hv34, na.rm = TRUE))
```

4.5 Question: Replicate figure 2. Make the figure as clear and informative as possible. You may want to create an indicator variable for percentiles above and below the threshold.

Code:

```
#panes_mean$group = as.numeric(as.character(panes_mean$group))

#figure2 <- ggplot(panes_mean, aes(x = mean_predincome, y = mean_receip
t) +
                  #geom_point() +
                  #geom_vline(xintercept = 0))
```

```
##My code keep showing error that "You've supplied a <NULL> object", an
d I don't understand why
```

4.6 Question: What is the purpose of this figure and what should we take away from it?

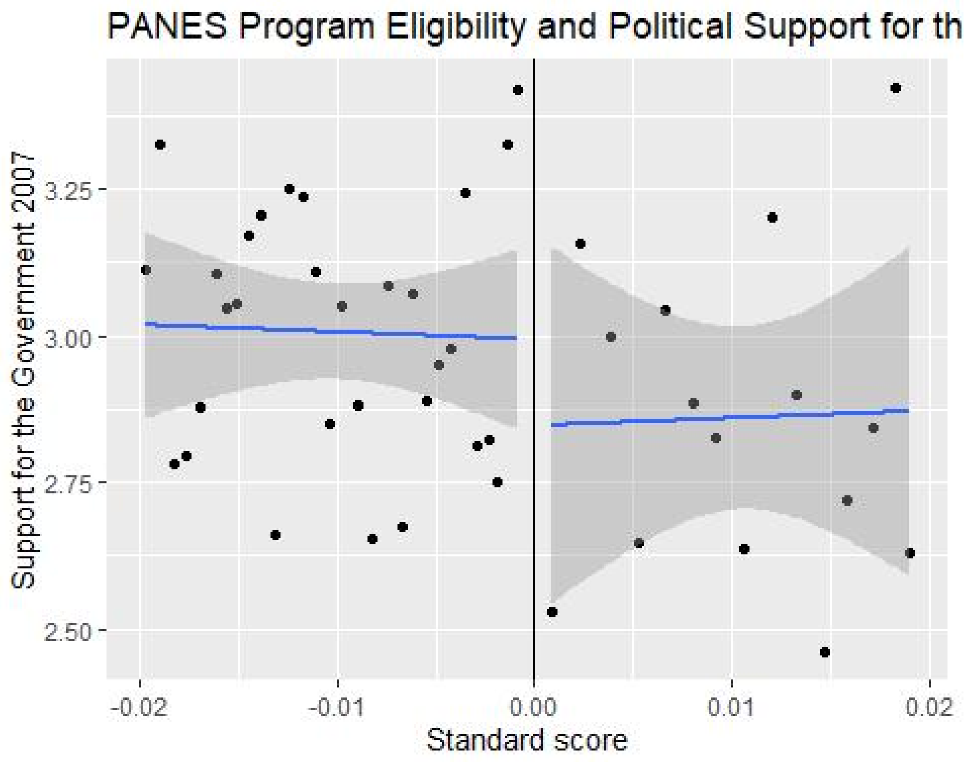
Answer: The figure provides graphical evidence that the program assignment rule was adhered to closely in practice, allowing the researchers to credibly estimate program impacts using the regression discontinuity design that relies on the sharp implementation of this rule.

4.7 Question: (2 pages) Replicate figures 3 and 4. Make these figures as clear and informative as possible.

Code:

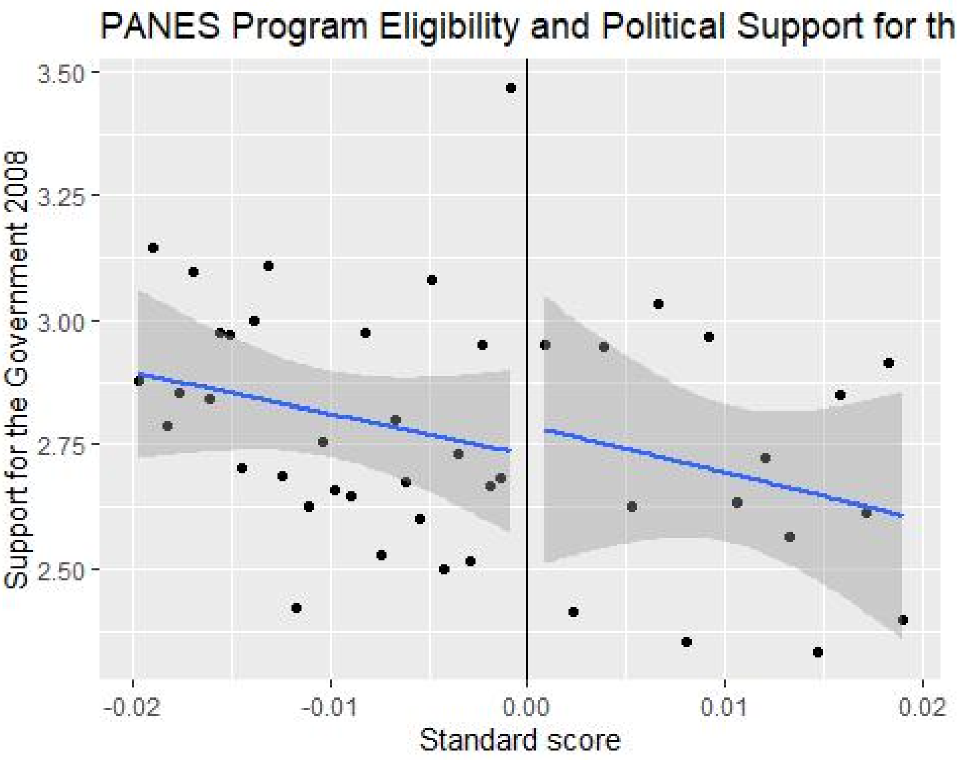
```
figure3 <- ggplot(panes_mean, aes(x = mean_predincome, y = mean_support
_2007))+
  geom_point()+
  geom_vline(xintercept = 0)+
  geom_smooth(aes(group=ifelse(mean_predincome < 0,1,2)), method = "lm",
 se = TRUE)+
  labs(title = "PANES Program Eligibility and Political Support for the
 Government, 2007 Follow-up Survey Round",
       x="Standard score",
       y="Support for the Government 2007")
figure3

## `geom_smooth()` using formula = 'y ~ x'
```



```
figure4 <- ggplot(panes_mean, aes(x = mean_predincome, y = mean_support
_2008))+
  geom_point()+
  geom_vline(xintercept = 0)+
  geom_smooth(aes(group=ifelse(mean_predincome < 0,1,2)), method = "lm",
 se = TRUE)+
  labs(title = "PANES Program Eligibility and Political Support for the
 Government, 2008 Follow-up Survey Round",
       x="Standard score",
       y="Support for the Government 2008")
figure4

## `geom_smooth()` using formula = 'y ~ x'
```



4.8 Question: Interpret these figures. What should we take away from them?

Answer: receiving PANES is higher than not receiving PANES

4.9 Question: Replicate the results of the three regressions estimated in the first column of table 1. Present your results in a table. Interpret the coefficients.

Code:

```
reg4.9.1 <- felm(aprobado ~ newtreat|0|0|ind_reest, panes)
reg4.9.2 <- felm(h_89 ~ newtreat|0|0|ind_reest, panes)
reg4.9.3 <- felm(hv34 ~ newtreat|0|0|ind_reest, panes)

stargazer(reg4.9.1, reg4.9.2, reg4.9.3, type="text", nobs=FALSE, mean.s
d=TRUE, median=TRUE, iqr=TRUE, se=list(reg4.9.1$rse, reg4.9.2$rse, reg4.
9.3$rse))

##
## ====================================================================
=====
##                                      Dependent variable:

##                     ------------------------------------------------
-----
```

```
##                                 aprobado          h_89             hv34

##                                   (1)              (2)              (3)

## ----------------------------------------------------------------------
-----
## newtreat                       0.993***          0.151**          0.130*

##                                  (0.002)          (0.075)          (0.076)

##

## Constant                         0.004**          2.853***         2.685***

##                                  (0.002)          (0.061)          (0.063)

##

## ----------------------------------------------------------------------
-----
## Observations                      2,232            2,219            2,051

## R2                                0.987            0.002            0.002

## Adjusted R2                       0.987            0.002            0.001

## Residual Std. Error 0.056 (df = 2230) 1.695 (df = 2217) 1.660 (df = 
2049)
## ======================================================================
=====
## Note:                                         *p<0.1; **p<0.05; ***p
<0.01
```

Answer: PANES received, support in 2007, support in 2008 are all significant and are positive related tp PANES eligibility. PANES received is the most statistical significant.

4.10 Question: Write down the specifications used in row 2 of columns 1,2 and 3 of table 1.

Answer:

$$ Support2007_i=\beta_0+\beta_1newtreat+\epsilon.\\
Support2007_i=\beta_0+\beta_1NewTreat*PredictedIncome+\epsilon.\\
Support2007_i=\beta_0+\beta_1NewTreat*PredictedIncome*PredictedIncome^2+\epsilon.\\ $$

4.11 Question: (2 pages) Replicate all of the results reported in row 2 of Table 1. Explain the difference between these specifications and interpret their coefficients.

Hint: the variables listed in the table above after newtreat are the controls you will want to include.

Code:

```
reg4.11.1 <- felm(h_89~newtreat|0|0|ind_reest, panes)
reg4.11.2 <- felm(h_89~newtreat*ind_reest|0|0|ind_reest, panes)
reg4.11.3 <- felm(h_89~newtreat*ind_reest*predincome_sq|0|0|ind_reest,
panes)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is
 either
## rank-deficient or not positive definite

reg4.11.4 <- felm(h_89~newtreat+bl_medad+lnbl_ytoth_pc+bl_hhsize+bl_med
uc+missbl_medad+misslnbl_ytoth_pc+missbl_hhsize+missbl_meduc+factor(sex
o)+misssexo+missedad+missaniosed|geo+edad+aniosed07|0|ind_reest, panes)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is
 either
## rank-deficient or not positive definite

reg4.11.5 <- felm(h_89~newtreat*ind_reest+bl_medad+lnbl_ytoth_pc+bl_hhs
ize+bl_meduc+missbl_medad+misslnbl_ytoth_pc+missbl_hhsize+missbl_meduc+
factor(sexo)+misssexo+missedad+missaniosed|geo+edad+aniosed07|0|ind_ree
st, panes)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is
 either
## rank-deficient or not positive definite

reg4.11.6 <- felm(h_89~newtreat*ind_reest*predincome_sq+bl_medad+lnbl_y
toth_pc+bl_hhsize+bl_meduc+missbl_medad+misslnbl_ytoth_pc+missbl_hhsize
+missbl_meduc+factor(sexo)+misssexo+missedad+missaniosed|geo+edad+anios
ed07|0|ind_reest, panes)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is
 either
## rank-deficient or not positive definite

stargazer(reg4.11.1, reg4.11.2, reg4.11.3, reg4.11.4, reg4.11.5, reg4.1
1.6, type="text", nobs=FALSE, mean.sd=TRUE, median=TRUE, iqr=TRUE, se=l
ist(reg4.11.1$rse, reg4.11.2$rse, reg4.11.3$rse, reg4.11.4$rse, reg4.11.
5$rse, reg4.11.6$rse))

##
## ====================================================================
=======================================================================
=
##
         Dependent variable:

##                                  -----------------------------------
-----------------------------------------------------------------------
-
##
                h_89

##                                         (1)               (2)
        (3)               (4)               (5)               (6)

## --------------------------------------------------------------------
-----------------------------------------------------------------------
-
## newtreat                              0.151**            0.112
       0.226            0.131*             0.135             0.226

##                                       (0.075)           (0.162)
      (0.255)           (0.079)           (0.166)           (0.256)

##


## ind_reest                                               -1.227
       5.364                               2.547            23.133

##                                                        (10.863)
     (43.609)                            (11.179)          (43.591)

##


## predincome_sq
    -323.957                                             -1,011.564

##
    (2,065.900)                                          (2,087.574)

##


## newtreat:ind_reest                                      -1.399
      18.454                              -4.783            -19.855

##                                                        (13.294)
     (56.121)                            (13.658)          (56.856)
```

```
##
## newtreat:predincome_sq
     1,628.184                                                 1,281.773
##
    (2,644.778)                                               (2,679.678)
##
## ind_reest:predincome_sq
##
      (0.000)                                                   (0.000)
##
## newtreat:ind_reest:predincome_sq
##
      (0.000)                                                   (0.000)
##
## bl_medad
                        -0.008*           -0.008*           -0.008*
##
                        (0.004)           (0.004)           (0.004)
##
## lnbl_ytoth_pc
                        0.00000           0.00000           0.00000
##
                       (0.00000)         (0.00000)         (0.00000)
##
```

```
## bl_hhsize
                        -0.006            -0.006            -0.005

##
                        (0.026)           (0.026)           (0.026)

##

## bl_meduc
                        0.00000           0.00000           0.00000

##
                       (0.00000)         (0.00000)         (0.00000)

##

## missbl_medad

##
                        (0.000)           (0.000)           (0.000)

##

## misslnbl_ytoth_pc

##
                        (0.000)           (0.000)           (0.000)

##

## missbl_hhsize

##
                        (0.000)           (0.000)           (0.000)

##

## missbl_meduc
```

```
##
                        (0.000)          (0.000)          (0.000)
##

## factor(sexo)2
                         0.014            0.015            0.015
##
                        (0.077)          (0.078)          (0.078)
##

## factor(sexo)999999
                        0.755**          0.765**          0.731**
##
                        (0.296)          (0.299)          (0.307)
##

## misssexo

##
                        (0.000)          (0.000)          (0.000)
##

## missedad

##
                        (0.000)          (0.000)          (0.000)
##

## missaniosed

##
                        (0.000)          (0.000)          (0.000)
```

```
##

## Constant                             2.853***          2.865***
     2.842***

##                                       (0.061)           (0.134)
      (0.201)

##

## --------------------------------------------------------------------
-----------------------------------------------------------------------
-
## Observations                           2,219             2,219
       2,219             2,219             2,219             2,219

## R2                                     0.002             0.002
       0.002             0.071             0.071             0.071

## Adjusted R2                            0.002             0.001
      0.00003            0.005             0.004             0.003

## Residual Std. Error              1.695 (df = 2217) 1.696 (df = 2215)
 1.696 (df = 2213) 1.692 (df = 2071) 1.693 (df = 2069) 1.694 (df = 2067)
## ====================================================================
=======================================================================
=
## Note:
                                             *p<0.1; **p<0.05; ***p<0.0
1
```

Answer:

4.12 Question: What is the point of including all of these specifications?

Answer: Avoid bias

4.13 Question: Using the coefficients estimated above, write out the function you would use to predict the probability a household supports the current government based on their predicted income score:

a) If they are eligible for the transfer using the results from column 1.

b) If they are not eligible for the transfer using the results from column 1.

c) If they are eligible for the transfer using the results from column 2.

d) If they are not eligible for the transfer using the results from column 2.

e) If they are eligible for the transfer using the results from column 3.

f) If they are not eligible for the transfer using the results from column 3.

Answer:

$$ a: Support2007_i=2.853+0.151*newtreat\\ b: Support2007_i=2.853\\ c:
Support2007_i=2.865-1.339*newtreat*predicted_income\\ d:
Support2007_i=2.865\\ e:
Support2007_i=2.842+0.000*newtreat*predicted_income*predicted_income^2\\ f:
Support2007_i=2.842\\ $$

4.14 Question: How narrow is the "bandwidth" used by the authors. Why does this matter? Check that the results are robust to a narrower bandwidth.

Answer: Smaller bandwith could effect the standard errors.

4.15 Question: The authors attribute these effects to the causal effect of receiving the government transfers. What is the implied assumption behind this interpretation?

Answer: allows the authors to interpret these effects as causally due to receiving the government transfers is that assignment to the PANES program near the eligibility threshold was as good as random.

4.16 Question: What evidence do they provide to support this assumption?

Answer: There are no significant discontinuities at the cutoff in a wide range of pretreatment covariates measured in 2005, including demographics, income, political behaviors like voter turnout (Table 2 and Figure A1) The discontinuity in actual program receipt at the threshold was nearly 100 percentage points (Figure 2). This suggests minimal manipulation of the assignment variable or non-compliance.

4.17 Question: Was this threshold eligibility score specifically designed for this particular program? Why does this matter?

Answer: No, the threshold eligibility score was not specifically designed for the PANES program. As mentioned in the paper, the formula for the predicted income score was devised by outside researchers at the University of the Republic in Uruguay. It was estimated using national household survey data collected in 2004, before the PANES program even existed.

It avoids the score being manipulated to serve political ends, rather than just targeting based on need. This bolsters the claim that assignment near the cutoff was as-if random.

Submission instructions:

- Make sure the final version of your assignment is knit in pdf format and uploaded to gradescope. Make sure you have one question response per page (unless otherwise indicated) so that question positions align with the template in gradescope.The final PDF should be 25 pages long.