MY457: Problem Set 5 - Regression Discontinuity Designs

Pedro Torres-Lopez, Michael Ganslmeier, Daniel de Kadt

Mon/03/Jun

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 5pm on Tue/11/Jun. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a setting with $i \in [1, ..., N]$ units. Treatment D_i is assigned based on values of a variable X_i, such that $D = \mathbf{1}[X_i > c]$. We are interested in the effect of D_i on Y_i.

1.1 What is the key identifying assumption in this setting, from the local randomization perspective?

The key identifying assumption in this setting, from the local randomization perspective, is that Within some small window around c, all units are as-if randomly assigned a value of X_i, and thus D_i.

1.2 What is the key assumption from the continuity perspective?

The key identifying assumption in this setting, from the continuity perspective, is that there is no discontinuity in potential outcomes at the threshold c.

$$\lim_{x \to c^-} E[Y_i(d)|X_i = x] = \lim_{x \to c^+} E[Y_i(d)|X_i = x]$$

1.3 Explain the difference between sharp regression discontinuity (RD) designs and fuzzy RD designs. Make sure you explain the difference in the design, assumptions, and estimands targeting by these two approaches.

In the sharp RDD treatment is assigned as a deterministic function of the running variable. In the fuzzy dseign, treatment is not deterministic, but the probability of receiving treatment changes discontinuously at the threshold.

Following the above, in a sharp design we assume perfect compliance while we allow for imperfect compliance in a fuzzy design.

Therefore, the estimated in a sharp design is the Local Average Treatment Effect (LATE), while the one for the fuzzy design is the LATE for compliers.

1.4 How is fuzzy RD related to instrumental variables?

In a fuzzy RD design, the running variable X_i serves as an instrumental variable. The estimation in a fuzzy RD context can be framed using a two-stage least squares (2SLS) approach, where in the first stage we regress D_i on X_i using the threshold as an indicator. On the second stage we regress Y_i on the fitted values of the first stage.

2 Simulations

In this question we will use simulated data to test some of our intuitions about instrumental variables. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Explain the code below.

```r
n <- 1000

set.seed(123)
X <- rnorm(n, mean = 0, sd = 15)
c <- 0

Y0 <- 100 + 0.45*X - 0.15*(X^2) + 0.001*(X^3) + rnorm(n, mean = 0, sd = 25)
Y1 <- 175 + 0.25*X + 0.15*(X^2) - 0.001*(X^3) + rnorm(n, mean = 0, sd = 25)

D <- ifelse(X > c, 1, 0)
Y <- ifelse(D == 1, Y1, Y0)

dat <- tibble(
  X,
  D,
  Y,
  Y0,
  Y1
)
```

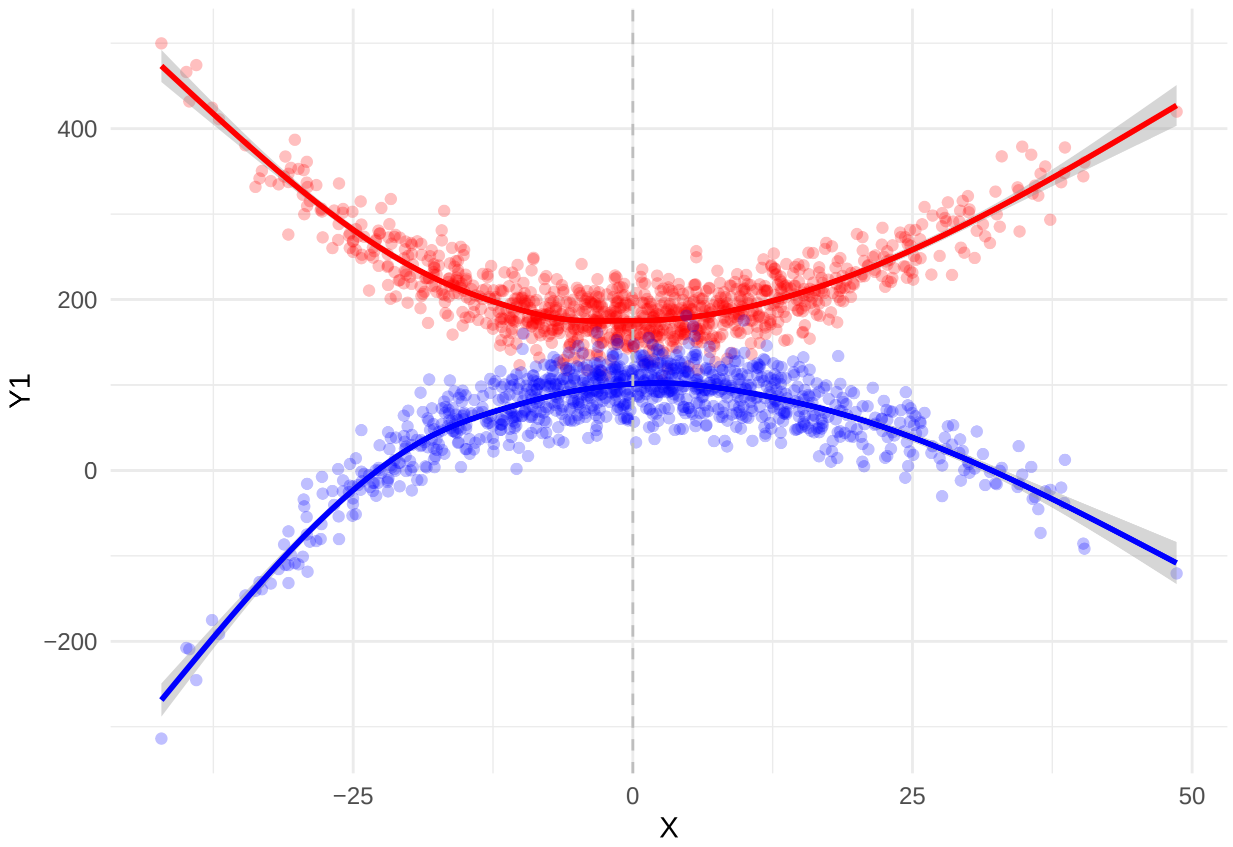
2.2 In this simulation, what is the true ATE and, what is the true LATE (at the threshold)? What do you conclude?

```r
mean(dat$Y1 - dat$Y0)
```

```
## [1] 138.9717
```

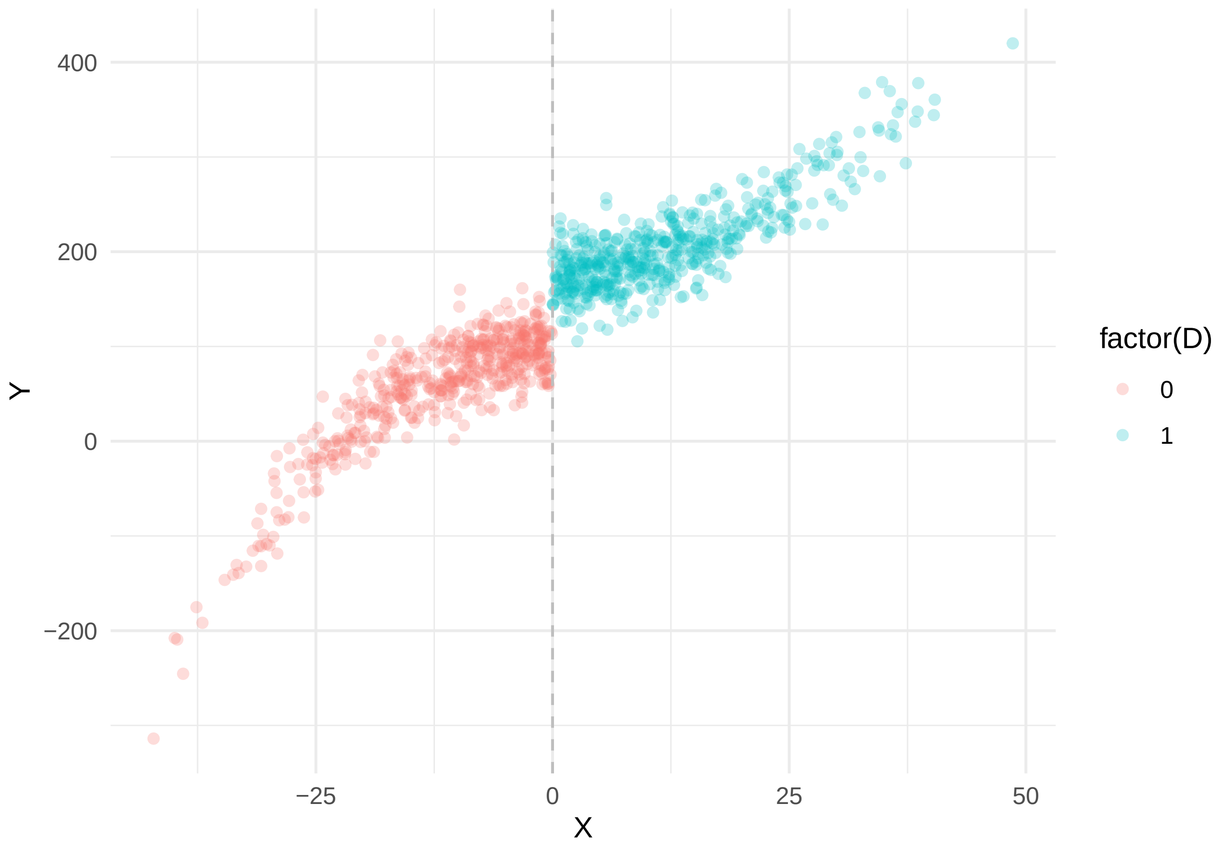
2.3 In a clear, well formatted, and compelling figure, visualise *both* potential outcomes and how they relate to the cutpoint.

```r
ggplot(dat) + aes(x = X, y = Y1) +
  geom_point(color = "red", alpha = 0.25) +
  geom_smooth(color = "red") +
  geom_point(aes(x=X, y=Y0), color = "blue", alpha = 0.25) +
  geom_smooth(aes(x=X, y=Y0), color = "blue") +
  geom_vline(xintercept = c, linetype = "dashed", color = "gray") +
  theme_minimal()
```



2.4 In a clear, well formatted, and compelling figure, visualise only the observed `Y`.

```r
ggplot(dat) + aes(x = X, y = Y, color = factor(D)) +
  geom_point(alpha = 0.25) +
  geom_vline(xintercept = c, linetype = "dashed", color = "gray") +
  theme_minimal()
```



2.5 Use a global (do not focus on a narrow window) linear regression with common slopes to estimate the LATE. What do you find? How does this estimator perform?

```
lm(Y ~ X*D, data = dat) %>% summary()
```

```
##
## Call:
## lm(formula = Y ~ X * D, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -184.617  -17.934    0.287   19.100   91.626
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 127.9907     2.2011  58.148 <0.0000000000000002 ***
## X             6.1044     0.1503  40.602 <0.0000000000000002 ***
## D            27.9271     3.0736   9.086 <0.0000000000000002 ***
## X:D          -1.9409     0.2069  -9.380 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.63 on 996 degrees of freedom
## Multiple R-squared:  0.8983, Adjusted R-squared:  0.898
## F-statistic:  2933 on 3 and 996 DF,  p-value: < 0.00000000000000022
```

2.6 Use a global (do not focus on a narrow window) polynomial regression with varying slopes to estimate the LATE. You may pick the polynomial order. What do you find? How does this estimator perform? How might you expect this estimator to perform outside of a simulated setting?

```
lm(Y~X*D + D*I(X^2), data = dat) %>% summary()
```

```
## 
## Call:
## lm(formula = Y ~ X * D + D * I(X^2), data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -78.277 -16.858  -0.881  16.733  77.588 
## 
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)    
## (Intercept) 94.83580    2.49445  38.019 < 0.0000000000000002 ***
## X           -0.97251    0.38491  -2.527             0.011673 *  
## D           76.73926    3.49135  21.980 < 0.0000000000000002 ***
## I(X^2)      -0.23019    0.01185 -19.426 < 0.0000000000000002 ***
## X:D          2.01517    0.52359   3.849             0.000126 ***
## D:I(X^2)     0.32458    0.01559  20.820 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 24.49 on 994 degrees of freedom
## Multiple R-squared:  0.9307, Adjusted R-squared:  0.9303 
## F-statistic:  2669 on 5 and 994 DF,  p-value: < 0.00000000000000022
```

```
lm(Y~X*D + D*I(X^2) + D*I(X^3), data = dat) %>% summary()
```

```
## 
## Call:
## lm(formula = Y ~ X * D + D * I(X^2) + D * I(X^3), data = dat)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -76.651 -16.665  -0.593  16.417  79.263 
## 
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)    
## (Intercept) 98.167578   3.201299  30.665 < 0.0000000000000002 ***
## X            0.270875   0.843112   0.321               0.7481    
## D           76.020685   4.429801  17.161 < 0.0000000000000002 ***
## I(X^2)      -0.138483   0.056599  -2.447               0.0146 *  
## I(X^3)       0.001730   0.001044   1.657               0.0978 .  
## X:D         -0.147387   1.122546  -0.131               0.8956    
## D:I(X^2)     0.297156   0.073335   4.052            0.0000547 ***
## D:I(X^3)    -0.002866   0.001318  -2.175               0.0299 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 24.45 on 992 degrees of freedom
## Multiple R-squared:  0.931,  Adjusted R-squared:  0.9305 
## F-statistic:  1913 on 7 and 992 DF,  p-value: < 0.00000000000000022
```

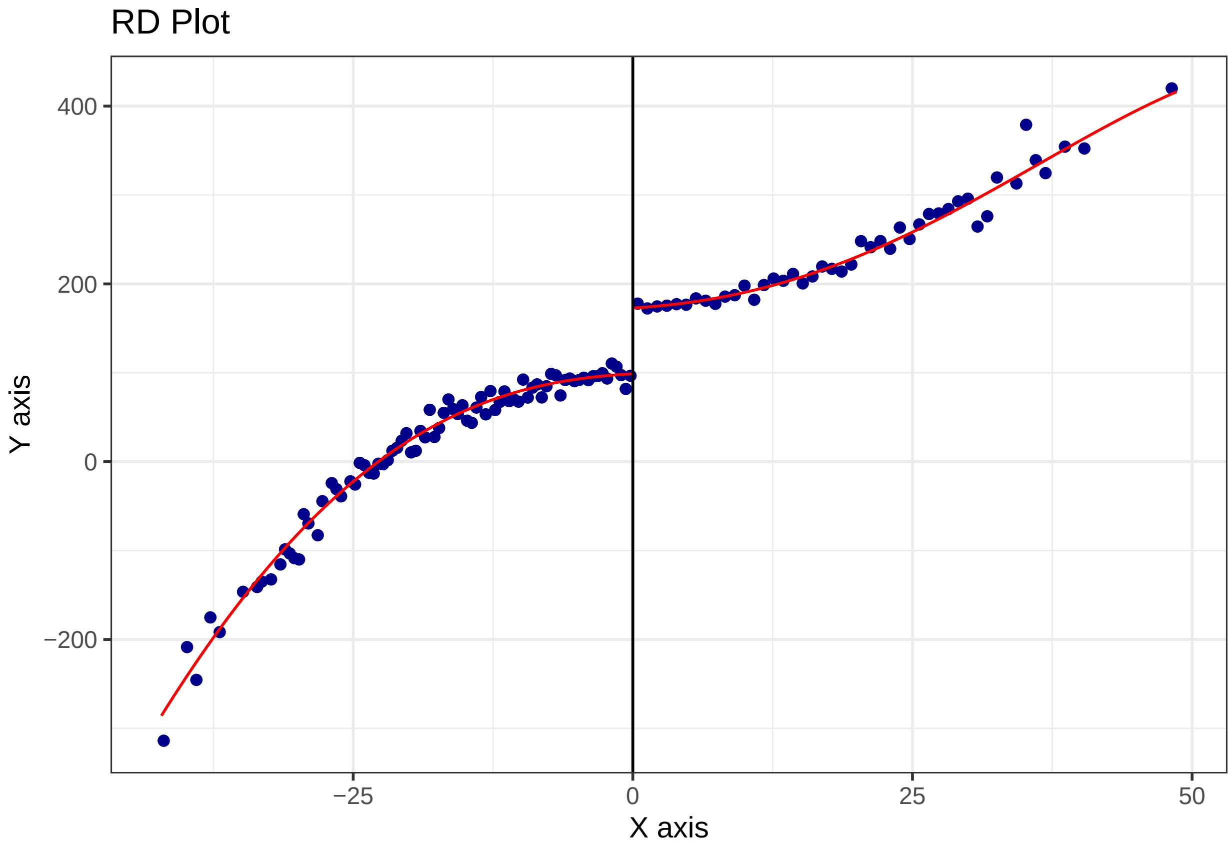
2.7 Let's now analyse our data with the `rdrobust` package. Estimate the LATE with local polynomial approximation. You may leave the settings at their defaults, or change settings if you so choose. What do you find?

```
rdrobust::rdrobust(y = Y, x = X, c = c) %>% summary()
```

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  495          505
## Eff. Number of Obs.             204          205
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   7.744        7.744
## BW bias (b)                  17.043       17.043
## rho (h/b)                     0.454        0.454
## Unique Obs.                     495          505
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    74.838     4.976    15.040     0.000    [65.085 , 84.590]
##         Robust         -          -    14.040     0.000    [66.348 , 87.878]
## =============================================================================
```

2.8 Use the `rdplot` function within the package to create a regression discontinuity plot.

```
rdrobust::rdplot(y = Y, x = X, c = c)
```



3 Replication

In this section, we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *Rural Roads and Local Economic Development*. Physical connectivity through paved roads may be a catapult for economic development. The authors of this paper study the economic impact of a $40 billion national programme designed to build roads across rural India.

3.1 Explain the authors' research design, and how they approach the identification strategy.

The authors use a research design based on a fuzzy regression discontinuity design (RDD) to evaluate the impacts of India's national rural road construction program. This approach leverages specific population thresholds set by the program for road construction, which provides a quasi-random assignment of roads to villages.

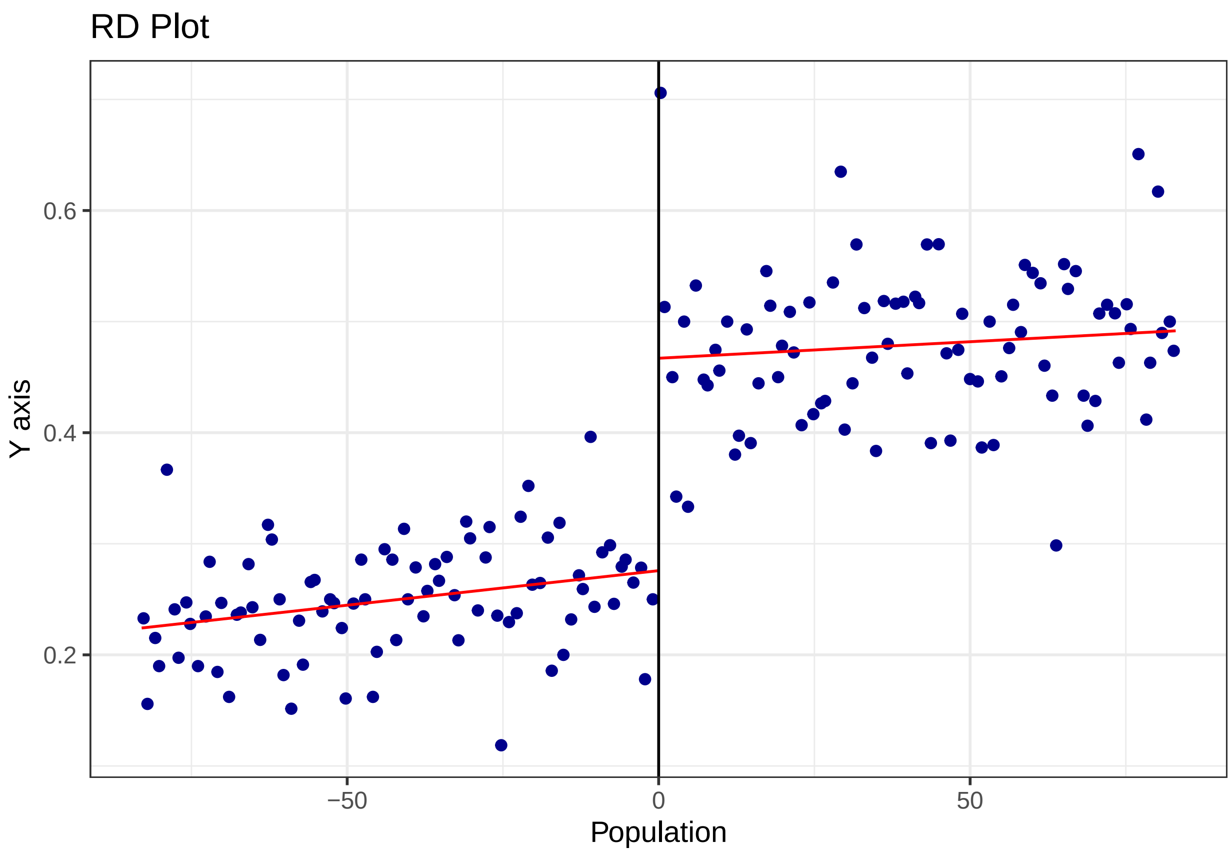
The key identification strategy is based on the quasi-random assignment of roads due to the population-based eligibility criteria. They employ a Fuzzy design becasue the assignment to treatment (road construction) is not perfectly defined at the threshold.

3.2 Read the data into `R` and using the `rdrobust` package, visualise any discontinuity in the variable `r2012` as a function of a cutoff or threshold in population (`v_pop`). What is the point of this analysis, and what do you find?

```
df <- read_dta("pmgsy_working_aer_mainsample_reduced.dta")
```

```
rdrobust::rdplot(y = df$r2012, x = df$v_pop, x.label = "Population", p = 1) %>% summary()
```

```
## [1] "Mass points detected in the running variable."
```



```
## Call: rdplot
##
## Number of Obs.                 11432
## Kernel                       Uniform
##
## Number of Obs.                  6018             5414
## Eff. Number of Obs.             6018             5414
## Order poly. fit (p)                1                1
## BW poly. fit (h)              83.000           83.000
## Number of bins scale               1                1
##
## Bins Selected                    133              132
## Average Bin Length             0.624            0.629
## Median Bin Length              0.624            0.629
##
## IMSE-optimal bins                  5                7
## Mimicking Variance bins          133              132
##
## Relative to IMSE-optimal:
## Implied scale                 26.600           18.857
## WIMSE variance weight          0.000            0.000
## WIMSE bias weight              1.000            1.000
```

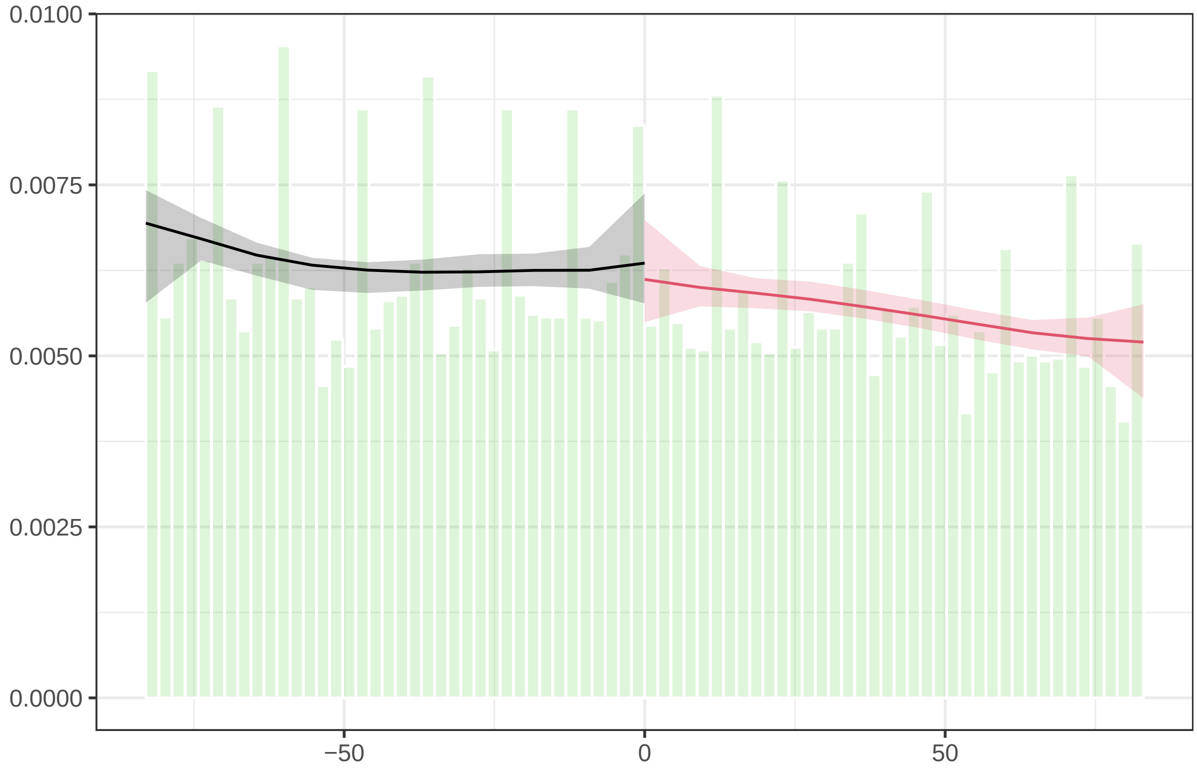
By plotting the probability of assignment with the running variable, we can see if we find a discontinuity in the assignment to treatment that is related to the cutoff point.

We do find a discontinuity, nevertheless it does not seem to be perfect, therefore a fuzzy RDD

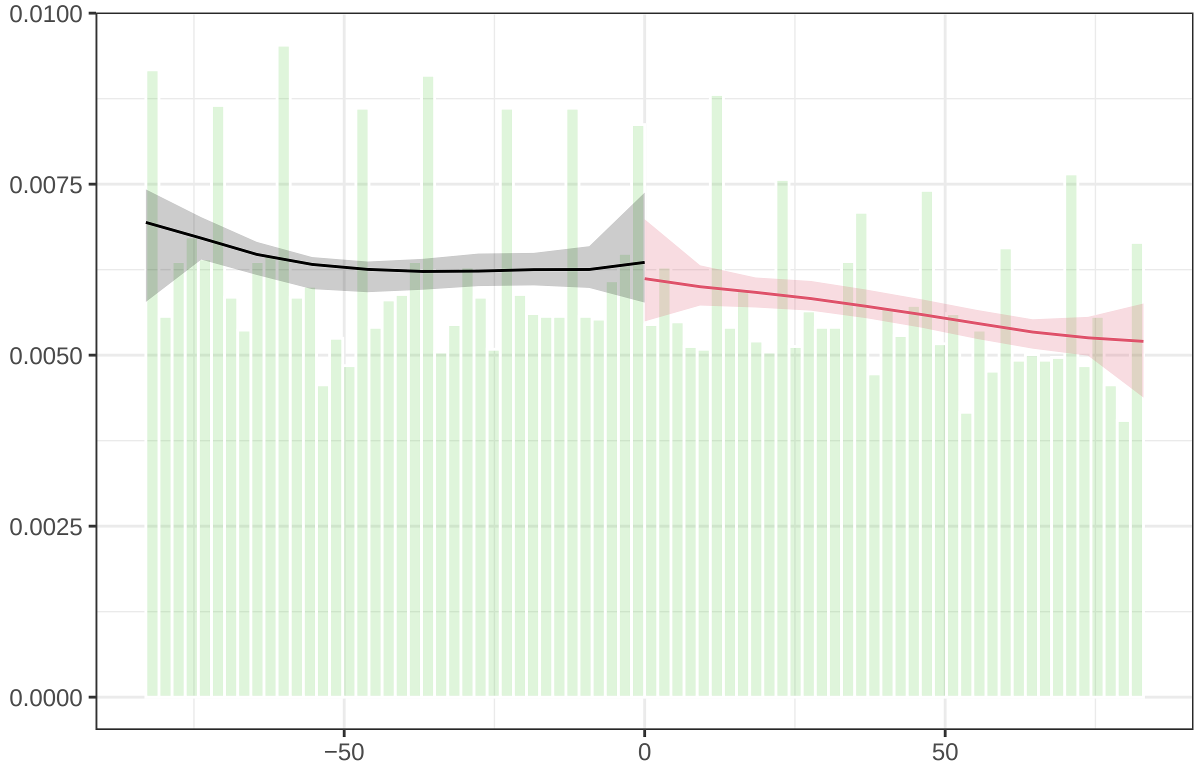
is a good approach.

3.3 A common concern in RD designs using population thresholds is manipulation at the threshold (see Eggers et al.). Using the `rddensity` package, test for sorting in around the cutoff. What do you find?

```
rddensity::rdplotdensity(rddensity::rddensity(df$v_pop), df$v_pop)
```



```
## $Estl
## Call: lpdensity
##
## Sample size                                      6086
## Polynomial order for point estimation    (p=)    2
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    3
## Kernel function                                  triangular
## Scaling factor                                   0.532324381069023
## Bandwidth method                                 user provided
##
## Use summary(...) to show estimates.
##
## $Estr
## Call: lpdensity
##
## Sample size                                      5414
## Polynomial order for point estimation    (p=)    2
## Order of derivative estimated            (v=)    1
## Polynomial order for confidence interval (q=)    3
## Kernel function                                   triangular
## Scaling factor                                    0.473536873414399
## Bandwidth method                                  user provided
##
## Use summary(...) to show estimates.
##
## $Estplot
```



The plot shows no discontinuity in the density of villages around the cutoffpoint. We find no evidence of sorting.

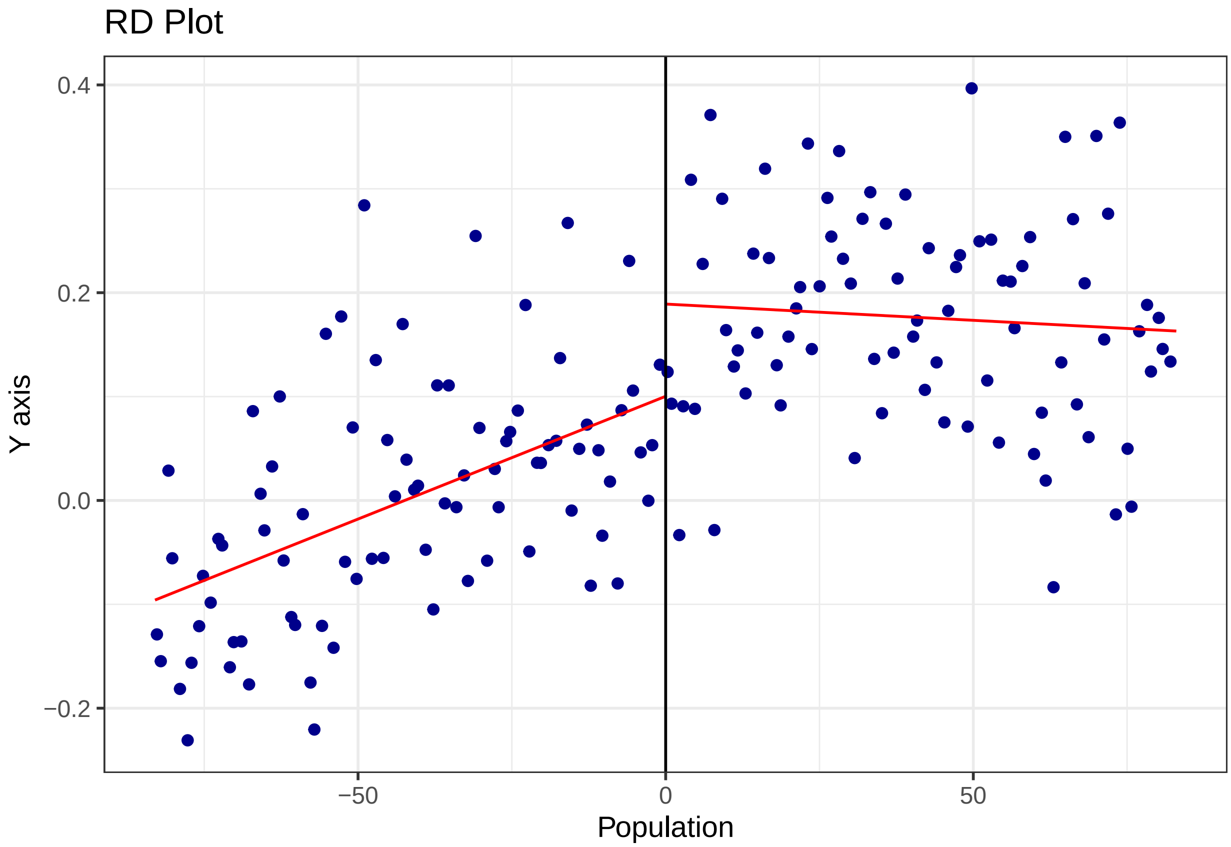
3.4 Select any two outcomes of interest and estimate the Fuzzy RDD. For each dependent variable, generate a single regression discontinuity plot, and overlay on the plot the key statistical results (point estimate of interest and 95% confidence interval). Use a polynomial of order 1 (`p = 1`), a triangular kernel, and the MSE-optimal bandwidth, as is done in the paper. Explain carefully what the quantity of interest you are estimating is, and in what ways it is `local`. What do you find?

```
rdrobust::rdrobust(y = df$transport_index_andrsn, x = df$v_pop, fuzzy = df$r2012) %>% summary()
```

```
## Fuzzy RD estimates using local polynomial regression.
##
## Number of Obs.                 11432
## BW type                        mserd
## Kernel                    Triangular
## VCE method                        NN
##
## Number of Obs.                  6018         5414
## Eff. Number of Obs.             2232         2169
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                  31.803       31.803
## BW bias (b)                  50.267       50.267
## rho (h/b)                     0.633        0.633
## Unique Obs.                      83           84
##
## First-stage estimates.
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.230     0.030     7.645     0.000     [0.171 , 0.289]
##         Robust         -         -     6.915     0.000     [0.176 , 0.316]
## =============================================================================
##
## Treatment effect estimates.
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.280     0.294     0.952     0.341    [-0.297 , 0.857]
##         Robust         -         -     0.576     0.565    [-0.480 , 0.880]
## =============================================================================
```

```
rdrobust::rdplot(y = df$transport_index_andrsn, x = df$v_pop, p = 1, x.label = "Population")
```

```
## [1] "Mass points detected in the running variable."
```

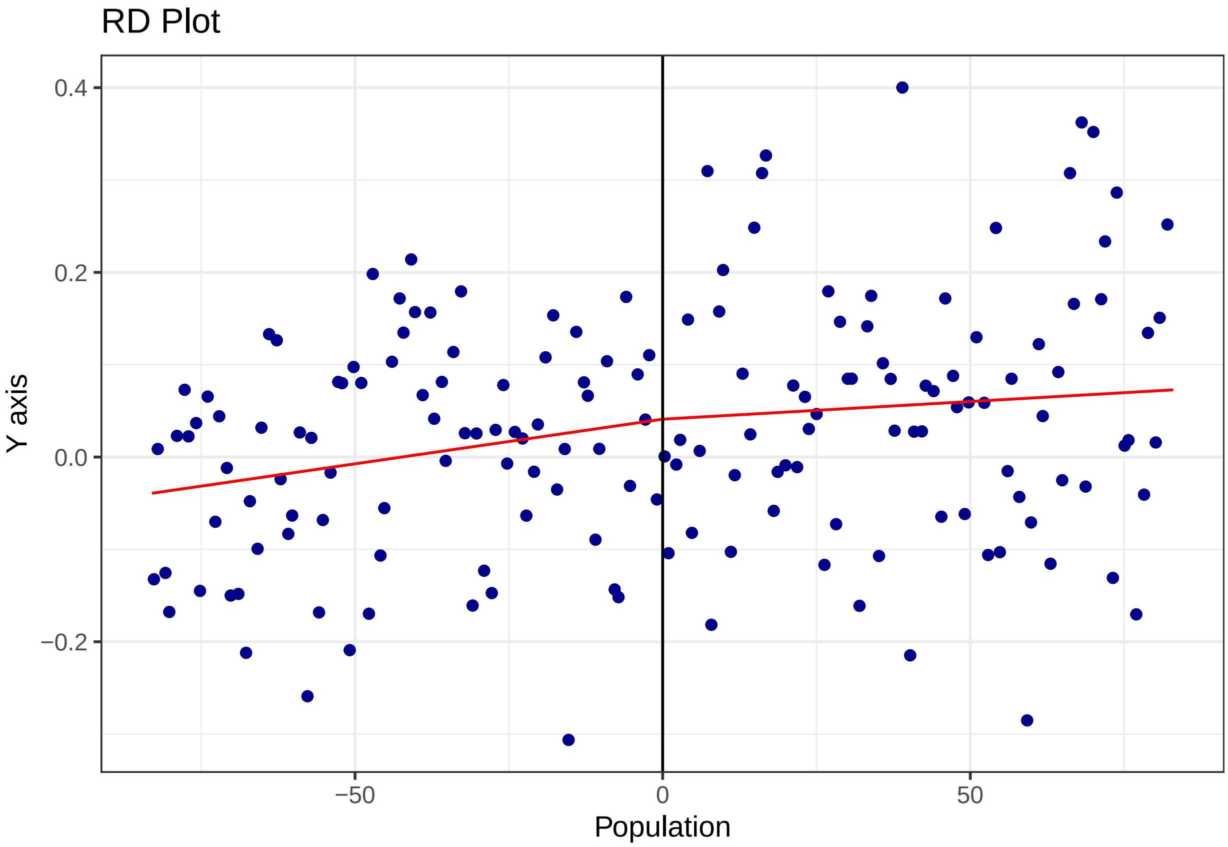


```
rdrobust::rdrobust(y = df$consumption_index_andrsn, x = df$v_pop, fuzzy = df$r2012) %>% summary()
```

```
## Fuzzy RD estimates using local polynomial regression.
##
## Number of Obs.                 11432
## BW type                        mserd
## Kernel                    Triangular
## VCE method                        NN
##
## Number of Obs.                  6018         5414
## Eff. Number of Obs.             1512         1513
## Order est. (p)                     1            1
## Order bias  (q)                    2            2
## BW est. (h)                   21.422       21.422
## BW bias (b)                   33.738       33.738
## rho (h/b)                      0.635        0.635
## Unique Obs.                       83           84
##
## First-stage estimates.
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional     0.259     0.036     7.151     0.000     [0.188 , 0.330]
##         Robust         -         -     6.576     0.000     [0.198 , 0.366]
## =============================================================================
##
## Treatment effect estimates.
##
## =============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## =============================================================================
##   Conventional    -0.195     0.305    -0.639     0.523    [-0.793 , 0.403]
##         Robust         -         -    -0.719     0.472    [-0.972 , 0.450]
## =============================================================================
```

```
rdrobust::rdplot(y = df$consumption_index_andrsn, x = df$v_pop, p = 1, x.label = "Population")
```

```
## [1] "Mass points detected in the running variable."
```



Choosing transport (transport__index__andrsn) and consumption (consumption__index__andrsn), we find evidence for the first stage to be significant. This means that the village size does affect receiving treatment, however, we do not find evidence of a discontinuity in these outcomes (second stage).