**APPENDIX METHODS: RD Application, Methods, and Results**

***Description of Application and Dataset***

In this appendix, we utilize the steps outlined in our manuscript to walk through an application of RD design. We specifically apply RD design to a teaching version of the Framingham Heart Study (FHS) dataset, which contains redacted data and is therefore intended for teaching purposes only (and not for interpreting actual results). The dataset is available upon request at BioLINCC (<https://biolincc.nhlbi.nih.gov/teaching/>). The FHS is a longitudinal cohort study which began in 1948 and has since helped to uncover some of the determinants of cardiovascular disease.1 In our application of RD to the teaching version of the FHS dataset, we seek to estimate the causal effect of antihypertensive medications on cardiovascular disease incidence among individuals in this cohort. In our example, the cutoff (C) for treatment is a systolic blood pressure of 140 mmHg, and the treatment is the use of an anti-hypertensive medication. **Web Table 1** provides a brief data dictionary of relevant variables and their purpose in our walkthrough.

**Web Table 1. Data Dictionary of Relevant Variables in the RD Application**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of variable** |
| CVD | Cardiovascular Disease in follow-up period (0 = No, 1 = Yes) | Outcome variable (Y) |
| SYBP | Systolic Blood Pressure (mmHg) | Running variable (X) |
| BPMEDS | Use of Anti-hypertensive medication at exam (0 = No, 1 = Yes) | Treatment status variable (D) |
| AGE | Age at exam (years) | Variable at baseline used to assess covariate balance |

***Description of Statistical Analysis***

Next, we describe the 7 steps of our RD analysis on the FHS teaching dataset. We delineate the steps as follows: we specify the main components of our RD model (Step 1), plot our RD data (Step 2), set up the polynomial model (Step 3), estimate the RD effect (Step 4), estimate the confidence interval of the RD effect (Step 5), perform robustness checks (Step 6), and perform RD design accounting for imperfect compliance to treatment assignment (Step 7). Within our analysis, we utilize both sharp and fuzzy RD designs. In addition, we run RD models using a variety of weighting functions and polynomial orders to demonstrate how these may affect the results. We also provide sample statistical R code (available on GitHub)2 that can help to guide investigators interested in utilizing RD Design for their research. Required R packages include ***rdrobust****3,4* and ***rddensity***5(available at <https://rdpackages.github.io/rdrobust/>).

***Walkthrough of the Code with Results***

First, we set up our R environment by attaching the necessary packages, and by loading and summarizing our dataset. In **Step 1**, we define the variables we would like to use for our RD analysis, comprised of the outcome value (Y, cardiovascular disease), the score/running variable (X, systolic blood pressure), the actual treatment (D, anti-hypertensive medication use), the cutoff value (C, systolic BP of 140 mmHg), and the assigned treatment (T = whether somebody was above the cutoff value, 1/0).

rm(list=ls())

### Load R packages

library(rdrobust)

library(rddensity)

### Load dataset  
setwd("Insert Directory Here") # Put location of your working directory here   
frm <-read.csv("frmgham2.csv")

###########################################################################

##### Step 1. Identify main components of the RD model ####################

###########################################################################

Y = frm$CVD# outcome variable

X = frm$SYSBP# running variable

D = frm$BPMEDS# actual treatment take-up indicator

C = 140# cutoff value

T = 1\*(X>=C)# treatment assignment indicator

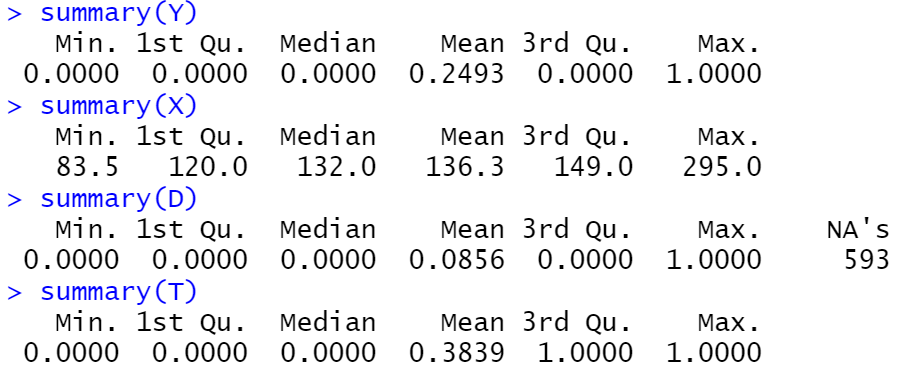
summary(Y)

summary(X)

summary(D)

summary(T)

**Output:**



In **Step 2**, we proceed by plotting the global relationship between the running variable (systolic blood pressure) which is depicted on the x-axis, and the outcome (cardiovascular disease) which is depicted on the y-axis. This relationship is modelled separately at each side of the cutoff (see **Web Figure 1**). As shown, there is an overall positive association between systolic blood pressure and cardiovascular disease. In addition, there appears to be a discontinuity at the cutoff, whereby individuals just above the cutoff have a lower risk of the outcome than individuals just below the cutoff.

##########################################################################

Step 2. RD Plots: Exploratory plot to illustrate the nature

of the relationship

##########################################################################

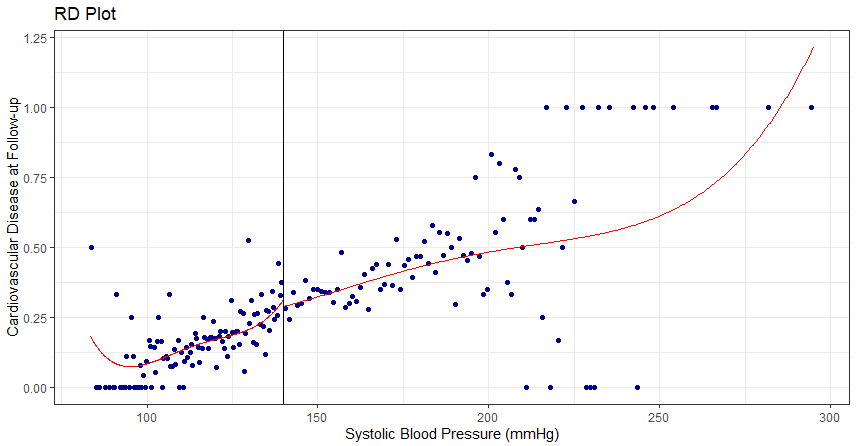
rdplot1 = rdplot(y = Y, x = X, c = C,

x.label = "Systolic Blood Pressure (mmHg)",

y.label = "Cardiovascular Disease at Follow-up")

summary(rdplot1)

Web Figure 1. Global Relationship between Running Variable (Systolic Blood Pressure) and Outcome (Cardiovascular Disease)



In **Step 3**, we set up the local polynomial model by specifying a polynomial order (i.e., P = 1, corresponding to a linear fit), a weighting function (i.e., kernel), and a selection procedure for computing the bandwidth (i.e., MSE optimal).6

####################################################################

##### Step 3. Set up the local polynomial model ####################

####################################################################

P = 1 # local linear approximation

K = 'triangular' # triangular kernel

BW = 'mserd' # MSE optimal bandwidth

In **Step 4,** we estimate the RD effect by running ***rdrobust*** using our choices for the tuning parameters.3 From the function output, we can recover the number of observations on each side of the cutoff, the bandwidth (h), and the number of observations within the bandwidth (i.e., Effective Number of Observations). In addition, the coefficient value on the row labeled “Conventional” corresponds to the RD causal estimate, the causal effect of the treatment on the outcome among subjects within the selected bandwidth. In this case, we obtain a reduction on the probability of a CVD event of 5.9 percentage points.

In addition, in **Step 5** we compute the 95% CI, corresponding to a plausible range of RD estimates consistent with our data, and the associated p-value. Here, the appropriate 95% CI can be recovered from the row labeled “Robust”, [-xxx, xxx], which accounts for the nonparametric features of our estimation approach and should be preferred over the “Conventional” one. The p-value associated with this 95% CI is 0.072.7

####################################################################

##### Steps 4-5. RD Estimation and Inference #######################

####################################################################

rd\_manual = rdrobust(y = Y, x = X, c = C,

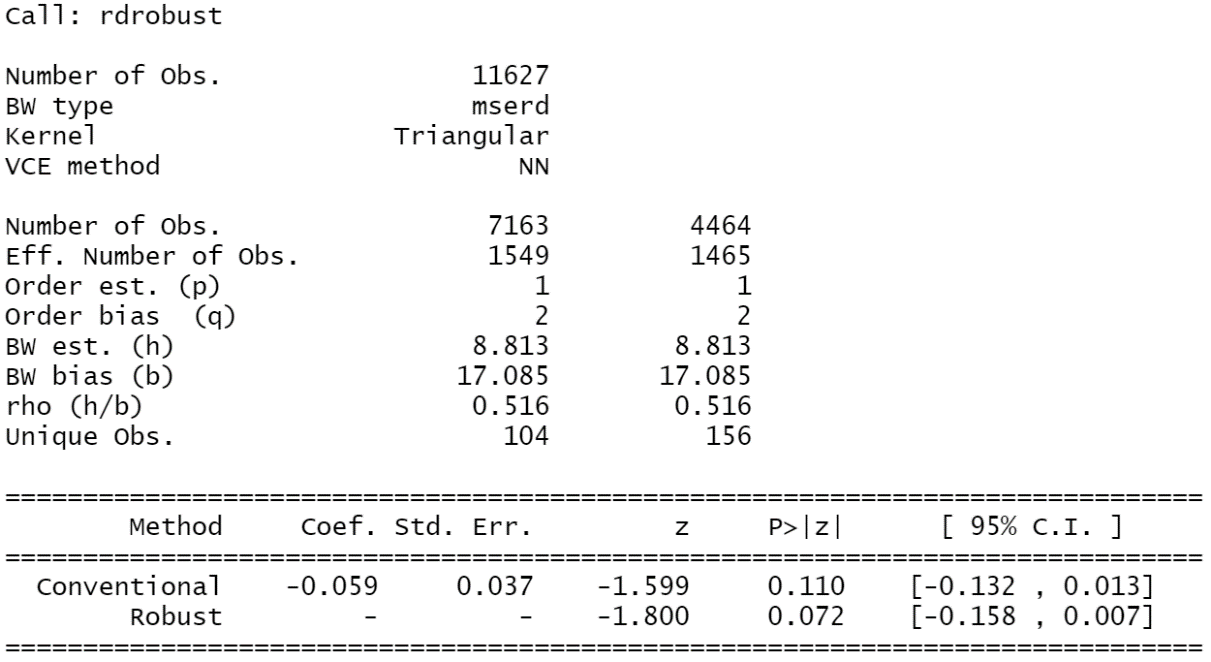
kernel = K, p = P, bwselect = BW)

summary(rd\_manual)

# You can also run RD using default choices, as shown here:

rd\_default = rdrobust(y = Y, x = X, c = C)

summary(rd\_default)

**Output**:

The results of our Sharp RD model (**Steps 3-5**) can be found in the first row of **Web Table 2**. Here, the coefficient value (-5.9%) corresponds to the RD causal estimate, the causal effect of the treatment on the outcome among subjects within the selected bandwidth. That is, antihypertensive treatment as indicated by the cutoff of 140 mmHg (assuming perfect compliance) is associated with an absolute risk reduction of cardiovascular disease of 5.9 percentage points. The 95% CI for this estimate ranges from a reduction of 15.8 percentage points to an increase of 0.7 percentage points.

**Web Table 2.** Results from RD Model for the treatment effect of antihypertensive medication on incident cardiovascular disease

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Type of RD Model** | **Original Number of observations** | **RD Effect** | **Standard Error** | **Robust 95% CI** | **Number of Observations selected within bandwidth – left of cutoff** | **Number of Observations selected within bandwidth – right of cutoff** | **Bandwidth (h)** |
| Sharp RD | 11627 | -0.059 | 0.037 | -0.158; 0.007 | 1549 | 1465 | 8.813 |
| Fuzzy RD | 11034 | -0.230 | 1.060 | -3.184; 1.727 | 3713 | 2448 | 18.679 |

We can also plot our estimated RD model using the ***rdplot*** function, similar to the one in Web Figure 1, but now reflecting our choices for the estimation of the RD model (i.e., limiting the functions on either side of the cutoff to the individuals selected within the bandwidth chosen by the RD model) instead of a global fit.

# Plotting the RD model

bw = rdrobust(y = Y, x = X, c = C,

kernel = 'triangular', p = 1, bwselect = 'mserd')$bws[1]

RD\_plot = rdplot(y = Y, x = X, c = C,

p = 1, kernel = 'triangular', h = bw,

x.label = "Systolic Blood Pressure (mmHg)",

y.label = "Cardiovascular Disease at Follow-up",

subset = (X >= (C - bw) & X <= (C + bw)), x.lim = c(C - bw –

1,+ C + bw + 1), y.lim = c(0,+ 1))

**Web Figure 2.** Estimation of the RD effect within the selected bandwidth.

Chart, scatter chart

Description automatically generated

Web Figure 2 legend: The distance between the lines at the cutoff gives you the ATE, which corresponds to the parameter estimate provided in Web Table 2.

In **Web Figure 2**, we can see a discontinuity at the cutoff. Individuals just above the cutoff (eligible for antihypertensive treatment) observed a decreased incidence in cardiovascular disease, as compared to individuals just below the cutoff. The distance between the lines at the cutoff gives you the ATE, which corresponds to the parameter estimate provided in the summary function (-0.059).

In **Step 6**, we can run a set of robustness checks and validation tests to assess the validity of our RD model. First, we explore the robustness of our RD effect estimates to the implementation choices of our local polynomial estimator. In particular, we look at alternative kernels, polynomial orders, and bandwidth choices.

###################################################################

##### Step 6. Robustness checks and validation ####################

###################################################################

# Change to uniform kernel

rd\_2 <- rdrobust(y = Y, x = X, c = C,

kernel = 'uniform', p = 1, bwselect = 'mserd')

summary(rd\_2)

# Change to local quadratic regression

rd\_3 <- rdrobust(y = Y, x = X, c = C,

kernel = 'triangular', p = 2, bwselect = 'mserd')

summary(rd\_3)

# Change to different bandwidths on each side:

rd\_4 <- rdrobust(y = Y, x = X, c = C,

kernel = 'triangular', p = 1, bwselect = 'msetwo')

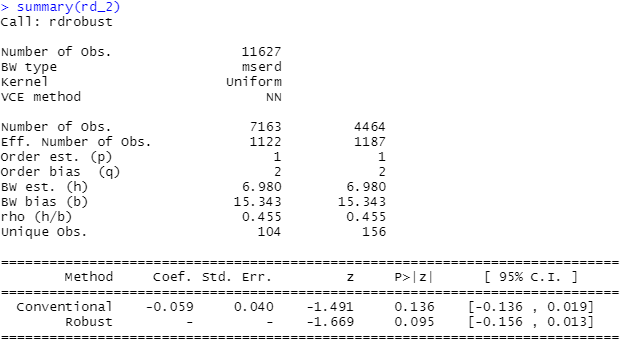
summary(rd\_4)

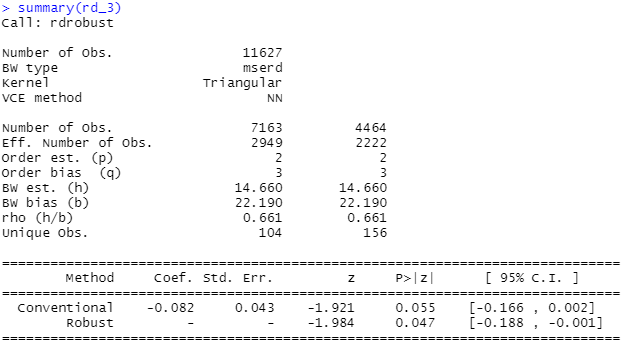
# Other bandwidth choices:

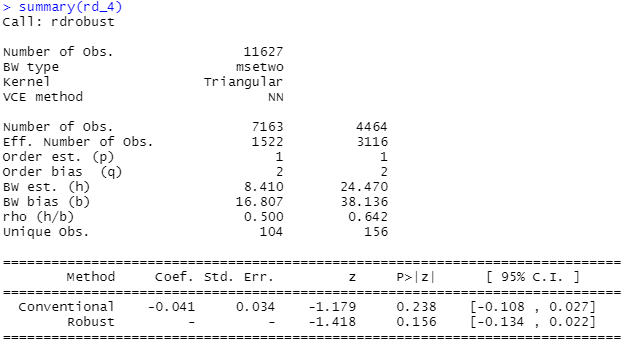
rd\_bws <- rdbwselect(y = Y, x = X, c = C, all = TRUE)

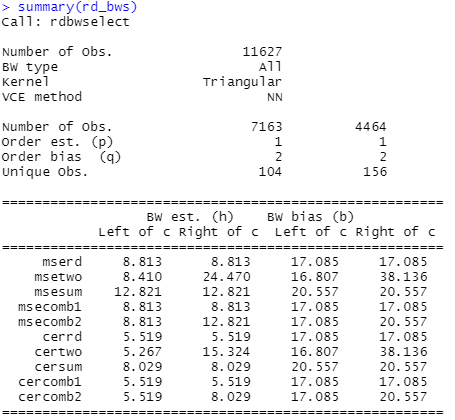
summary(rd\_bws)

**Output:**









From our robustness checks, we see that the main results remain (i.e., a protective effect is observed between the indicated anti-hypertensive medication and cardiovascular disease). Next, we conduct a manipulation test, looking for potential violation of the assumed continuity of observations around the threshold. If this assumption holds, we should expect to see a relatively consistent frequency of observations at each side of the cutoff. This test can be easily computed using the ***rddensity*** package, which provides the results of a statistical test together with a density plot illustrating the findings.5 We display our density plot for testing the continuity assumption (of observations around the cutoff) in **Web Figure 3**.

# Test for manipulation of the running variable

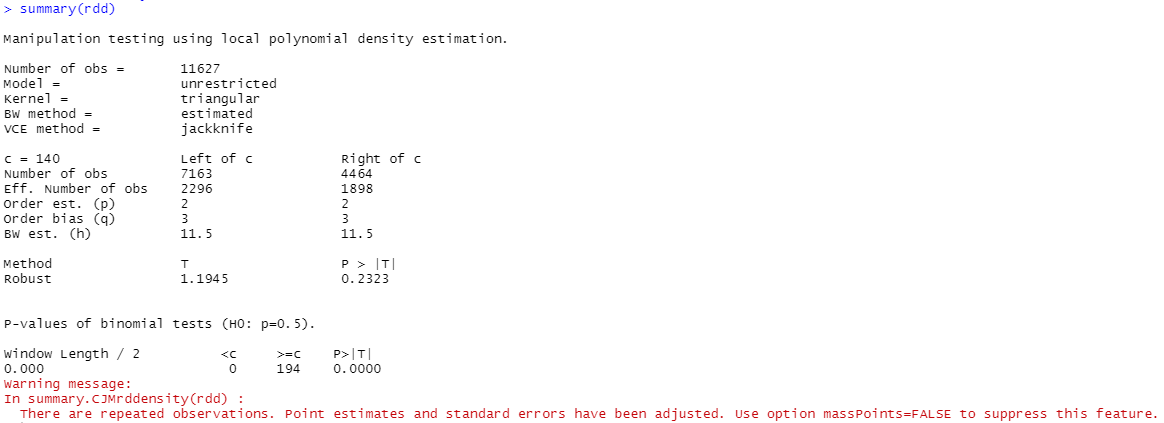
rdd <- rddensity(X = X, c = C)

summary(rdd)

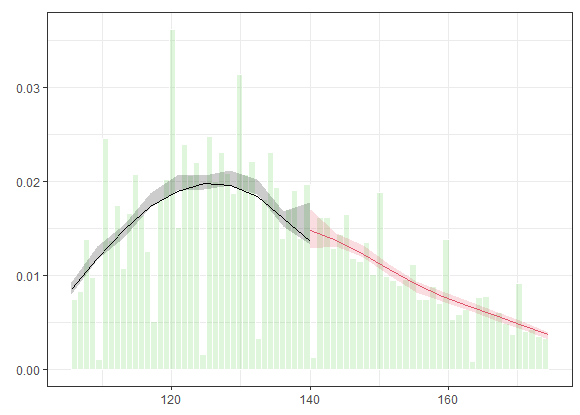
# Figure

plot <- rdplotdensity(rdd, X)

**Output:**



**Web Figure 3.** Density plot for test of manipulation of running variable.



From both the density plot and the results of the test, we can conclude that there is no evidence of a jump in the density of the running variable at the cutoff, which is consistent with a lack of manipulation.

As an additional check, we assess exchangeability around the cutoff by looking at covariate balance; this involves taking a given covariate (e.g., age) and treating it as an outcome in an RD model. One can subsequently view the RD effect and visualize a plot to see whether there is a significant difference in the level of the covariate of interest, comparing observations below and above the cutoff. If the exchangeability assumption holds, then we should expect the levels of this covariate to be comparable below and above the cutoff (i.e., the RD effect here should be close to 0). **Web Table 3** displays the results from this RD model, and **Web Figure 4** provides the corresponding plot.

## Covariate balance assessment (Age)

rd\_covs = rdrobust(y = frm$AGE, x = X, c = C)

summary(rd\_covs)

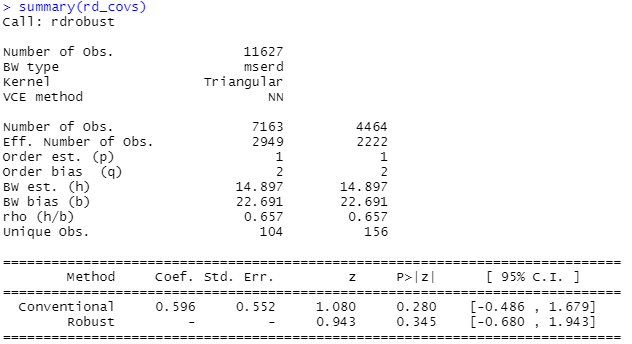
rdplot(y = frm$AGE, x = X, c = C,

title = "Covariate Balance - RD Plot",

x.label = "Systolic Blood Pressure (mmHg)",

y.label = "Age")

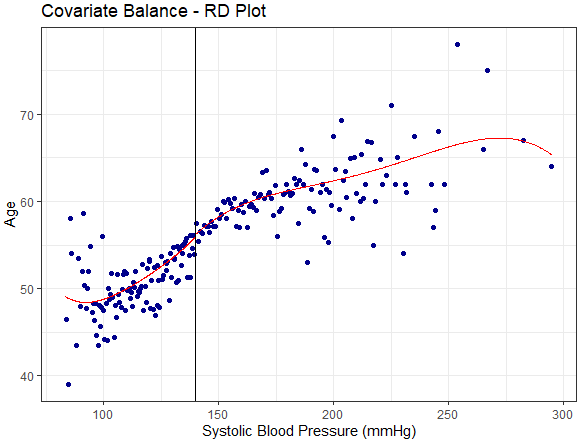
**Output:**



**Web Table 3.** Results from RD Model checking for Covariate Balance (i.e. exchangeability) around the cutoff. We show the example of age as the covariate of interest.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Original Number of observations** | **RD Effect** | **Standard Error** | **Robust 95% CI** | **Number of Observations selected within bandwidth – left of cutoff** | **Number of Observations selected within bandwidth – right of cutoff** | **Bandwidth (h)** |
| 11627 | 0.596 | 0.552 | -0.680; 1.943 | 2249 | 2222 | 14.897 |

**Web Figure 4.** Plot of RD Model checking for Covariate Balance (i.e. exchangeability) around the cutoff. We show the example of age as the covariate of interest.



The output from this model shows that there is no significant difference in age comparing individuals below and above the cutoff (i.e., no discontinuity is observed at the cutoff); this provides some evidence for the exchangeability assumption required for RD design.

Finally, in **Step 7**, we account for imperfect compliance by running a fuzzy RD model. To accomplish this, we can first assess the level of compliance by computing the RD effect using the observed treatment, BPMEDS, as the outcome. A plot of this relationship, between Systolic Blood Pressure (x-axis) and the observed treatment of anti-hypertensive medication (y-axis), can be found in **Web Figure 5**.

#############################################################

#### Step 7: Fuzzy RD: average effect for compliers #########

#############################################################

# First, assess compliance by computing RD effect using D as the outcome

RD\_comp = rdrobust(y = D, x = X, c = C)

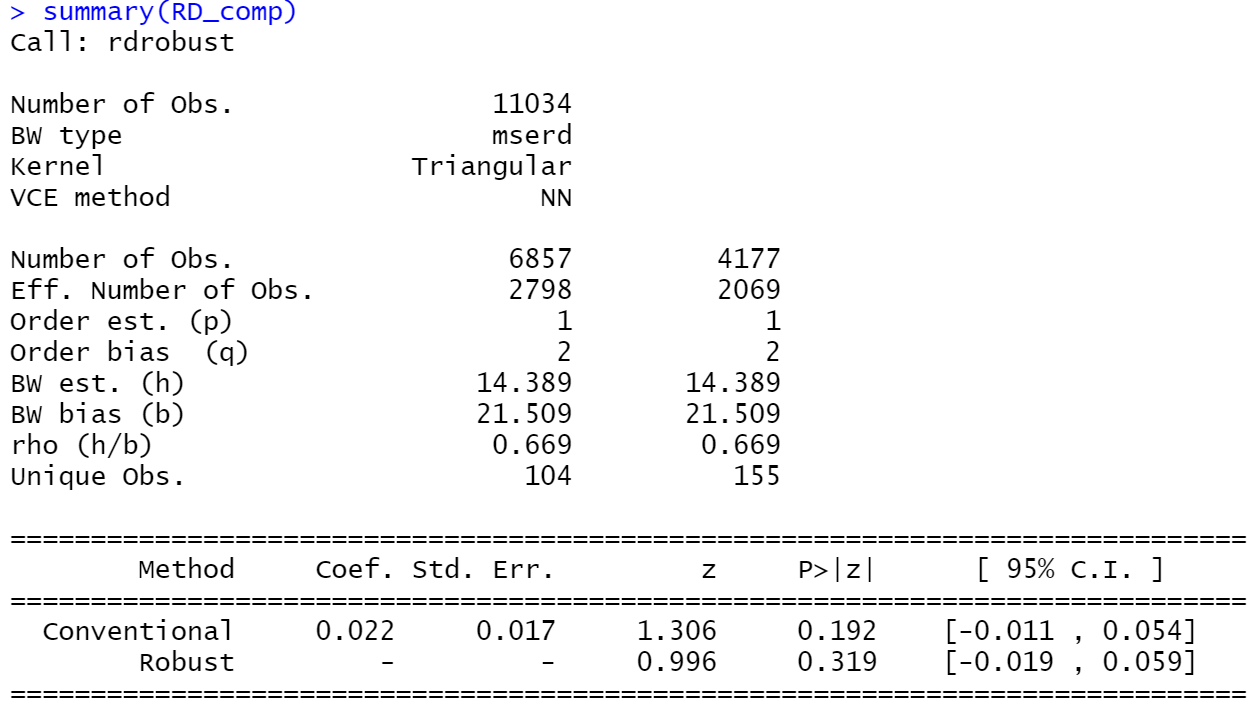
summary(RD\_comp)

RD\_comp\_plot = rdplot(y = D, x = X, c = C,

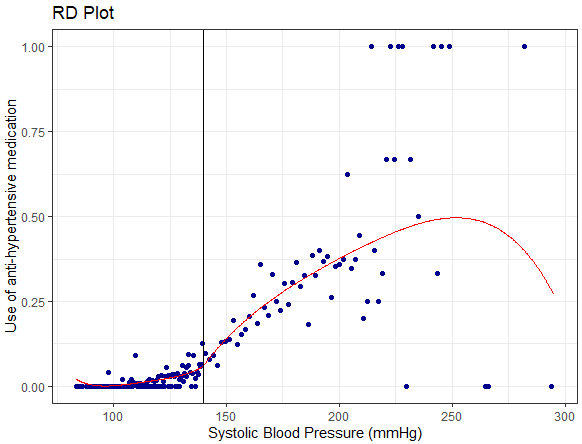
x.label = "Systolic Blood Pressure (mmHg)",

y.label = "Use of anti-hypertensive medication")

**Output:**



**Web Figure 5.** Plot of Uptake of Anti-Hypertensive Medication around the cutoff.



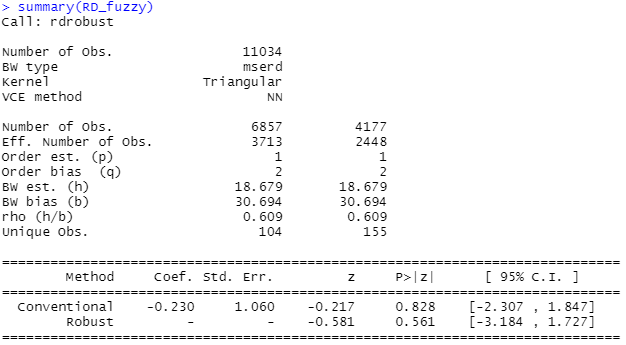
If the compliance is not perfect (i.e., the absolute value of the RD effect in the above output is not close to 1), as is evident here, we can proceed by subsequently following the same steps as above to run a fuzzy RD model where the potential endogenous treatment take-up (BPMEDS) is instrumented with the indicator for being above/below the cutoff, which is still exogenous.

# Fuzzy RD model:

RD\_fuzzy = rdrobust(y = Y, x = X, c = C, fuzzy = D)

summary(RD\_fuzzy)

**Output:**



From this final output (results also reported in **Web Table 2**), the coefficient value (-0.230) from the fuzzy RD model corresponds to the per-protocol treatment effect of anti-hypertensive medication on cardiovascular disease; that is, this is the complier-specific treatment effect, computed by dividing the sharp RD effect by the compliance rate (the percentage of participants who received the treatment that they were assigned to).

**References**

1. Tsao CW, Vasan RS. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology.* 2015;44(6):1800-1813.

2. Zeki Al Hazzouri A. Overview and Steps to Conduct a Regression Discontinuity Analysis. <https://github.com/AdinaZekiAlHazzouri/RDD_Paper>. Published 2022. Accessed.

3. Calonico S, Cattaneo MD, Titiunik R. rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *R Journal.* 2015;7:38-51.

4. Calonico S, Cattaneo MD, Farrell MH, Titiunik R. Rdrobust: Software for Regression-discontinuity Designs. *The Stata Journal.* 2017;17(2):372-404.

5. Cattaneo MD, Jansson M, Ma X. Manipulation Testing Based on Density Discontinuity. *Stata Journal.* 2018;18:234-261.

6. Calonico S, Cattaneo MD, Farrell M. Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal.* 2019.

7. Calonico S, Cattaneo M, Titiunik R. Robust Data-Driven Inference in the Regression-Discontinuity Design. *Stata J.* 2014;14.