HW6:Regression Discontinuity Simulation

Yongchao Zhao

Objective

The goal of this exercise is to simulate and analyze data that might have arisen from a policy where eligibility was determined based on one observed measure. Data from this type of setting are often consistent with the assumptions of the regression discontinuity designs we discussed in class.

Setting

This assignment simulates hypothetical data collected on women who gave birth at any one of several hospitals in disadvantaged neighborhoods in New York City in 2010. We are envisioning a government policy that makes available pre- and post-natal (through 2 years post-birth) health care for pregnant women, new mothers and their children. This program is only available for women in households with income below $20,000 at the time they gave birth. The general question of interest is whether this program increases a measure of child health at age 3. You will generate data for a sample of 1000 individuals.

Clean regression discontinuity design. For this assignment we will make the unrealistic assumption that everyone who is eligible for the program participates and no one participates who is not eligible.

Question 1. God role: simulate income.

Simulate the "assignment variable" (sometimes referred to as the "running variable", "forcing variable", or "rating"), income, in units of thousands of dollars. Call the variable "income". Try to create a distribution that mimics the key features of the data displayed in income_hist.pdf.

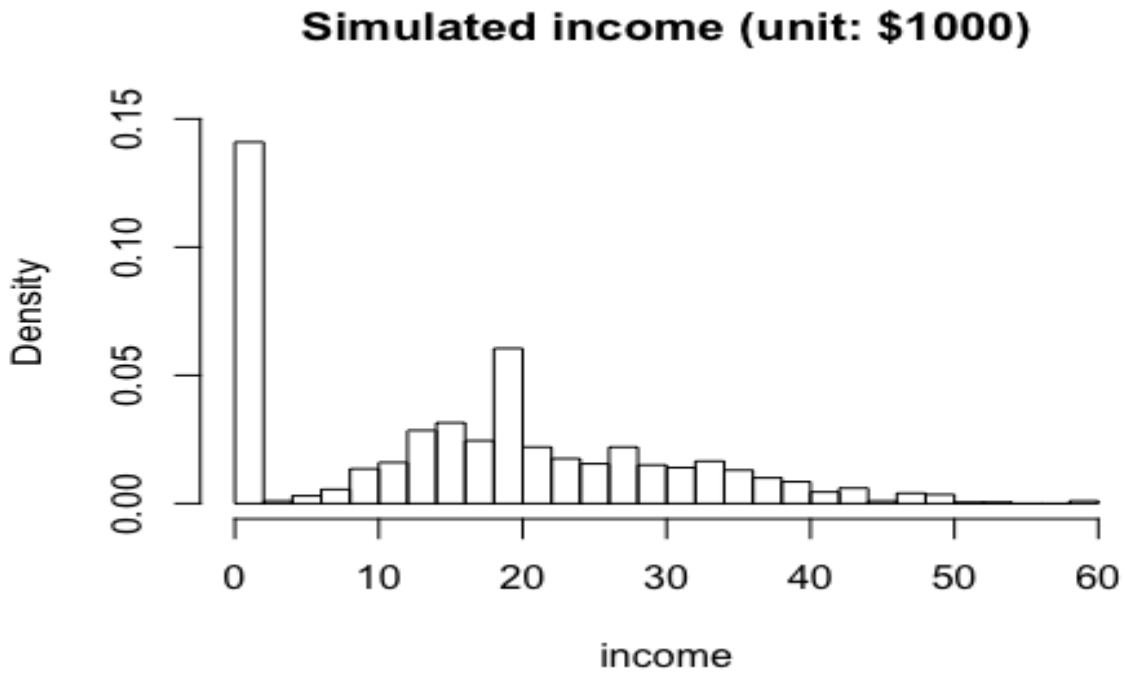
```
set.seed(123)
income <- c(
  runif(1000*0.28, 0, 2),
  runif(1000*0.07, 18, 20),
  rnorm(1000*0.30, mean=15, sd=5),
  rnorm(1000*0.30, mean=28, sd=8),
  rnorm(1000*0.05, mean=40, sd=10)
)

negs <- length(income[income<0])
income[income<0] <- runif(negs,0,60) #negative income into 0
income[income>60] <- 60 # maximal is 60

summary(income)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00125  1.78389 16.67363 16.53868 25.42599 60.00000
```

```r
hist(income, xlim = c(0, 60), ylim = c(0, 0.15),breaks = 30, freq = F, main =
 "Simulated income (unit: $1000)")
```



Question 2. Policy maker role: Assign eligibility indicator.

Create an indicator for program eligibility for this sample. Call this variable "eligible".

```r
eligible <- ifelse(income < 20, 1, 0)
table(eligible, useNA = "ifany")
```

```
## eligible
##   0   1
## 350 650
```

Question 3: God role.

For question 3 you will simulate a health measure with a minimum possible score on *observed data* of 0 and maximum possible score of *30*. You will simulate data from two possible worlds that vary with regard to the relationships between health and income.

Question 3a

(a) God role. Simulate potential outcomes for World A.

i) Generate the potential outcomes for health assuming linear models for both E[Y(0) | X] and E[Y(1) | X]. This health measure should have a minimum possible score of 0 and maximum possible score of 30. The *expected* treatment effect for everyone should be 4 (in other words, E[Y(1) - Y(0) | X] should be 4 at all levels of X). The residual standard deviation of each potential outcome should be *2*.

ii) Save two datasets: (1) fullA should have the forcing variable and both potential outcomes and (2) obsA should have the forcing variable, the eligibility variable, and the observed outcome.

```
# Q3a: linear models for both y1 and y0
fullA <- data.frame(income, y0 = NA, y1 = NA)
set.seed(31)

fullA$y0 <-  5.5 + 0.25*income + rnorm(1000, mean = 0, sd = 2)
summary(fullA$y0)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6514  6.9762  9.5958  9.6976 12.2562 22.4783

fullA$y1 <- fullA$y0 + rnorm(1000, mean = 4, sd = 2)
summary(fullA$y1)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.191  10.579  13.438  13.634  16.399  28.261

mean(fullA$y1 - fullA$y0)

## [1] 3.93674

library(dplyr)

obsA <- fullA %>% cbind(eligible) %>%
        mutate(y = ifelse(eligible == 1, y1, y0))%>%
        select(income, eligible, y)
```

Question 3b

(b) Simulate potential outcomes for World B.

i) Generate the potential outcomes for health assuming a linear model for E[Y(0) | X] and a quadratic model for E[Y(1) | X]. The treatment effect at the threshold (the level of X that determines eligibility) should be 4. The residual standard deviation of each potential outcome should be *2*. Creating this DGP may be facilitated by using a transformed version of your income variable that subtracts out the threshold value.

ii) Save two datasets: (1) fullB should have the forcing variable and both potential outcomes and (2) obsB should have the forcing variable, the eligibility variable, and the observed outcome.

```
fullB <- data.frame(income, y0 = NA, y1 = NA)
income_new <- income - 20

set.seed(32)
fullB$y0 <- 14 + 0.4*income_new + rnorm(1000, 0, 2)

summary(fullB$y0)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.631   8.201  12.375  12.618  16.185  29.688
```

```
for (i in 1:1000){
if(abs(income_new[i]) < 2) {
  fullB$y1[i] = fullB$y0[i] + rnorm(1, 4, 2)
} else {
  fullB$y1[i] = 1/125*income_new[i]^2 + rnorm(1, 14, 2)
}
}
summary(fullB$y1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.994  14.102  15.812  15.902  17.567  29.886
```

```
fullB_sub <- subset(fullB, abs(fullB$income-20) < 2)
mean(fullB_sub$y1-fullB_sub$y0)
```

```
## [1] 3.720192
```

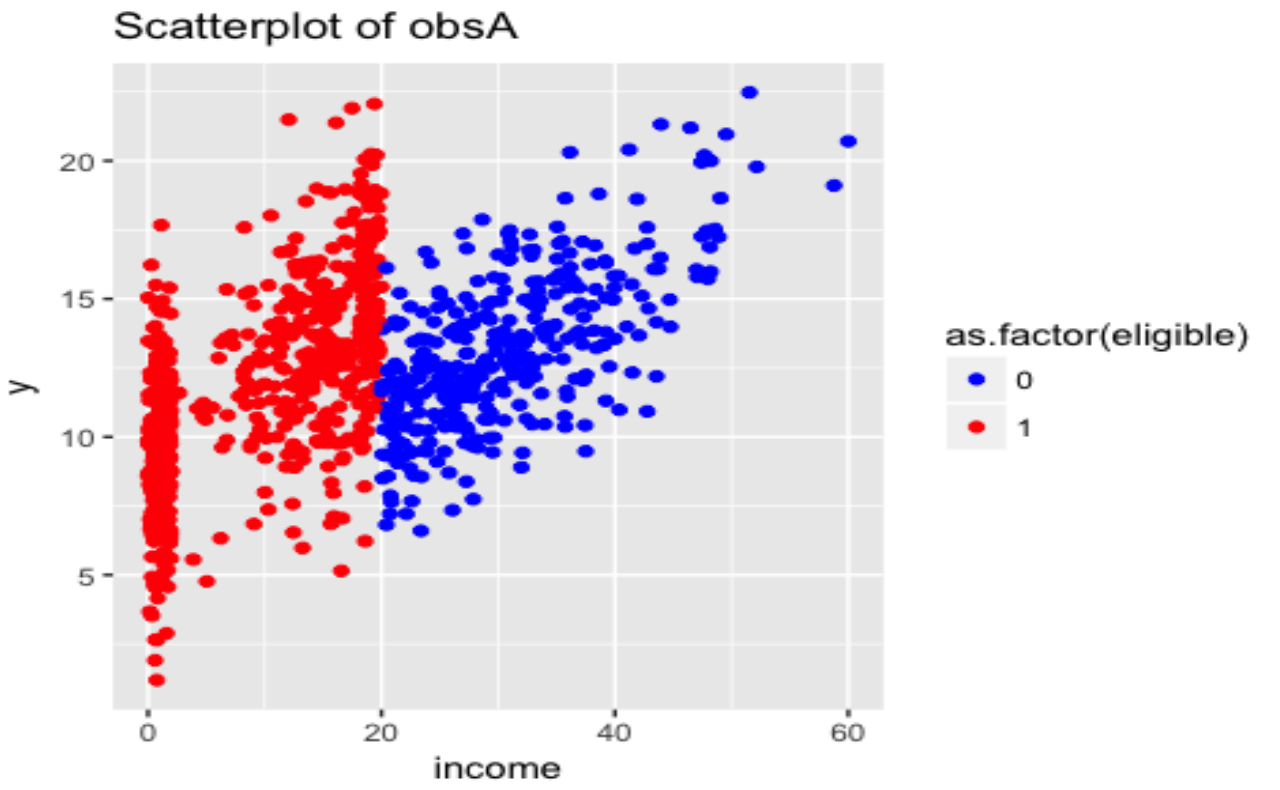
```
obsB <- fullB %>% cbind(eligible) %>%
        mutate(y = ifelse(eligible == 1, y1, y0))%>%
        select(income, eligible, y)
```

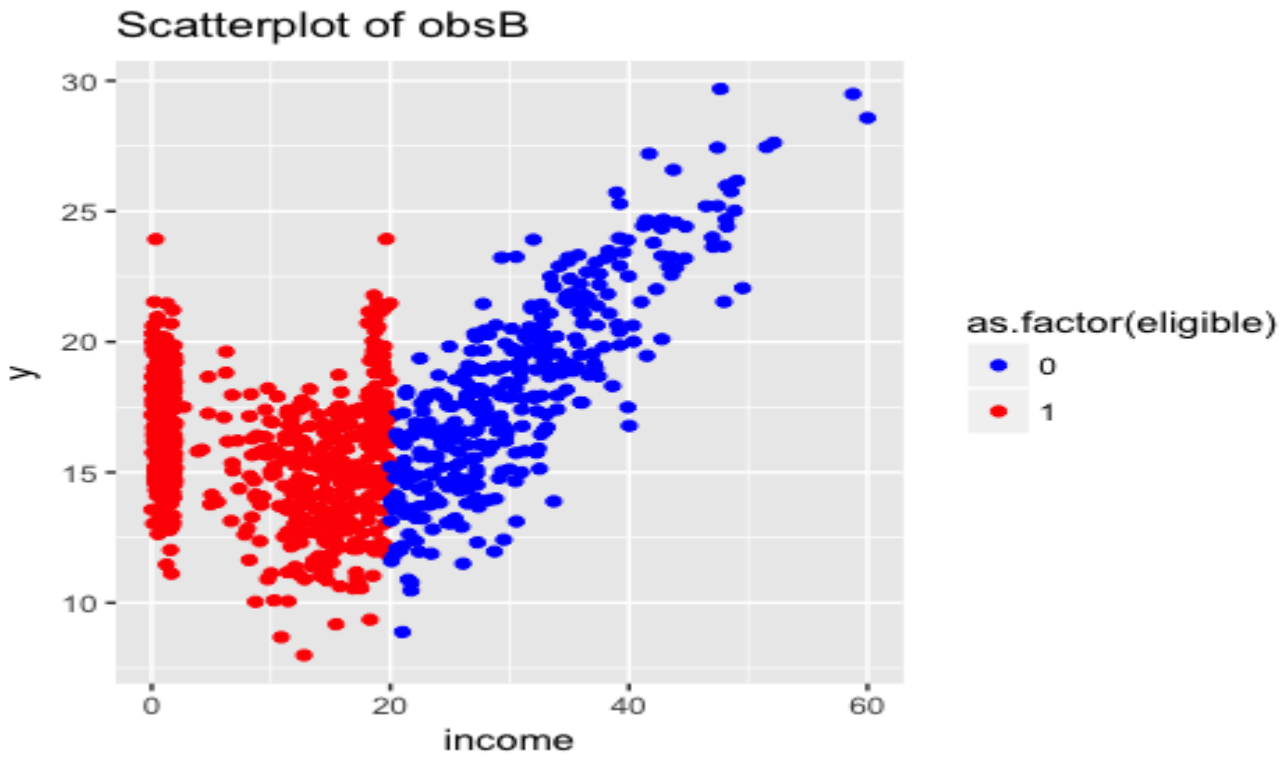
Question 4. Researcher role. Plot your data!

Make two scatter plots of income (x-axis) versus observed health (y-axis), one corresponding to each world. In each, plot eligible participants in red and non-eligible participants in blue.

```
library(ggplot2)
ggplot(data = obsA, aes(income, y, color = as.factor(eligible))) + geom_point
() + scale_color_manual(values = c("blue","red")) + ggtitle("Scatterplot of o
bsA")

ggplot(data = obsB, aes(income, y, color = as.factor(eligible))) + geom_point
() + scale_color_manual(values = c("blue","red")) + ggtitle("Scatterplot of o
bsB")
```


Scatterplot of obsA
y
income
0
20
40
60
5
10
15
20
as.factor(eligible)
0
1


Scatterplot of obsB
y
income
0
20
40
60
10
15
20
25
30
as.factor(eligible)
0
1

Question 5. Researcher role. Estimate the treatment effect for World A and World B using all the data.

Now we will estimate effects in a number of different ways. Each model should include reported income and eligible as predictors. In each case use the model fit to report the estimate of the effect of the program at the threshold level of income. All models in Question 5 will be fit to all the data.

Question 5a: Researcher role. Estimates for World A using all the data.

(a) Using all the data from World A, perform the following analyses.

(i) Fit a linear model to the full dataset. Do not include an interaction.

(ii) Fit a linear model to the full dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between treatment and control groups.

```
(for.estimate <- data.frame(income = c(20, 20), eligible = c(1, 0)))
```

```
##   income eligible
## 1     20        1
## 2     20        0
```

```
fitA.1 <- lm(y ~ income + eligible, data = obsA)
pred1 <- predict(fitA.1, for.estimate, type = "response")
pred1[[1]]-pred1[[2]]
```

```
## [1] 4.396793
```

```
fitA.2 <- lm(y ~ eligible * income, data = obsA)
pred2 <- predict(fitA.2, for.estimate, type = "response")
pred2[[1]]-pred2[[2]]
```

```
## [1] 4.249367
```

```
fitA.3 <- lm(y ~ I(income^2) * eligible * income, data = obsA)
pred3 <- predict(fitA.3, for.estimate, type = "response")
pred3[[1]]-pred3[[2]]
```

```
## [1] 4.739339
```

Question 5b: Researcher role. Estimates for World B using all the data.

(b) Using all the data from World B, perform the following analyses.

(i) Fit a linear model to the full dataset. Do not include an interaction.

(ii) Fit a linear model to the full dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between the treatment and control groups).

```r
fitB.1 <- lm(y ~ income + eligible, data = obsB)
pred4 <- predict(fitB.1, for.estimate, type = "response")
pred4[[1]]-pred4[[2]]
```

```
## [1] -0.2332295
```

```r
fitB.2 <- lm(y ~ eligible * income, data = obsB)
pred5 <- predict(fitB.2, for.estimate, type = "response")
pred5[[1]]-pred5[[2]]
```

```
## [1] 1.08561
```

```r
fitB.3 <- lm(y ~ I(income^2) * eligible * income, data = obsB)
pred6 <- predict(fitB.3, for.estimate, type = "response")
pred6[[1]]-pred6[[2]]
```

```
## [1] 3.518541
```

Question 6. Researcher role. Estimate the treatment effect for World A and World B using data close to the threshold.

We will again estimate effects in a number of different ways. Each model should include "income" and "eligible" as predictors. In each case use the model fit to report the estimate of the effect of the program at the threshold level of income. All models in Question 6 will be fit only to women with incomes ranging fro $18,000 to $22,000.

Question 6a: Researcher role. Estimates for World A using the restricted data.

(a) Using the restricted data (for participants with incomes between $18K and $22K) from World A, perform the following analyses.

(i) Fit a linear model to the restricted dataset. Do not include an interaction.

(ii) Fit a linear model to the restricted dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between the treatment and control groups).

```r
selected <- (income <= 22 & income >= 18)

rd.fitA.1 <- lm(y ~ income + eligible, data = obsA, subset = selected)
pred7 <- predict(rd.fitA.1, for.estimate, type = "response")
pred7[[1]]-pred7[[2]]
```

```
## [1] 5.260308
```

```r
rd.fitA.2 <- lm(y ~ eligible * income, data = obsA, subset = selected)
pred8 <- predict(rd.fitA.2, for.estimate, type = "response")
pred8[[1]]-pred8[[2]]
```

```
## [1] 5.2965
```

```
rd.fitA.3 <- lm(y ~ I(income^2) * eligible * income, data = obsA, subset = se
lected )
pred9 <- predict(rd.fitA.3, for.estimate, type = "response")
pred9[[1]]-pred9[[2]]
```

```
## [1] 1.790781
```

Question 6b: Researcher role. Estimates for World B using the restricted data.

(b) Using the restricted data (for participants with incomes between $18K and $22K) from World B, perform the following analyses.

(i) Fit a linear model to the restricted dataset. Do not include an interaction.

(ii) Fit a linear model to the restricted dataset, include an interaction between income and eligible.

(iii) Fit a model that is quadratic in income and includes an interaction between both income terms and eligible (that is – allow the shape of the relationship to vary between treatment and control groups.

```
rd.fitB.1 <- lm(y ~ income + eligible, data = obsB, subset = selected)
pred10 <- predict(rd.fitB.1, for.estimate, type = "response")
pred10[[1]]-pred10[[2]]
```

```
## [1] 3.40085
```

```
rd.fitB.2 <- lm(y ~ eligible * income, data = obsB, subset = selected)
pred11 <- predict(rd.fitB.2, for.estimate, type = "response")
pred11[[1]]-pred11[[2]]
```

```
## [1] 3.237317
```

```
rd.fitB.3 <- lm(y ~ I(income^2) * eligible * income, data = obsB, subset = se
lected )
pred12 <- predict(rd.fitB.3, for.estimate, type = "response")
pred12[[1]]-pred12[[2]]
```

```
## [1] 4.122016
```

Question 7. Researcher role. Displaying your estimates.

Present your estimates from questions 5 and 6 into one or two tables or figures, clearly noting which world the data are from, which models the estimates are from, and which analysis sample was used.

```
# for model estimation
estimate_df <- as.data.frame(rbind(pred1, pred2, pred3, pred4, pred5, pred6,
pred7, pred8, pred9, pred10, pred11, pred12))

names(estimate_df) <- c("eligible", "not_eligible")
estimate_df$estimation <- estimate_df$eligible - estimate_df$not_eligible

print(estimate_df)
```

```
##         eligible not_eligible estimation
## pred1  14.75399     10.35719  4.3967927
## pred2  14.96608     10.71672  4.2493668
## pred3  15.46073     10.72139  4.7393392
## pred4  16.97979     17.21302 -0.2332295
## pred5  15.08241     13.99680  1.0856100
## pred6  17.57830     14.05976  3.5185405
## pred7  15.68270     10.42240  5.2603076
## pred8  15.66346     10.36696  5.2964999
## pred9  13.91216     12.12138  1.7907811
## pred10 17.29033     13.88948  3.4008498
## pred11 17.37730     14.13998  3.2373173
## pred12 17.62895     13.50694  4.1220164
```

Question	Data	Subset	Model	Estimation
Q5a(i)	obsA	full data	y ~ income + eligible	**4.397**
Q5a(ii)	obsA	full data	y ~ income * eligible	**4.249**
Q5a(iii)	obsA	full data	y ~ I(income^2) * income * eligible	**4.739**
Q5b(i)	obsB	full data	y ~ income + eligible	**-0.233**
Q5b(ii)	obsB	full data	y ~ income * eligible	**1.086**
Q5b(iii)	obsB	full data	y ~ I(income^2) * income * eligible	**3.519**
Q6a(i)	obsA	income between 18k to 22k	y ~ income + eligible	**5.260**
Q6a(ii)	obsA	income between 18k to 22k	y ~ income * eligible	**5.296**
Q6a(iii)	obsA	income between 18k to 22k	y ~ I(income^2) * income * eligible	**1.791**
Q6b(i)	obsB	income between 18k to 22k	y ~ income + eligible	**3.401**
Q6b(iii)	obsB	income between 18k to 22k	y ~ income * eligible	**3.237**
Q6b(iii)	obsB	income between 18k to 22k	y ~ I(income^2) * income * eligible	**4.122**

Question 8. Researcher role. Thinking about the data.

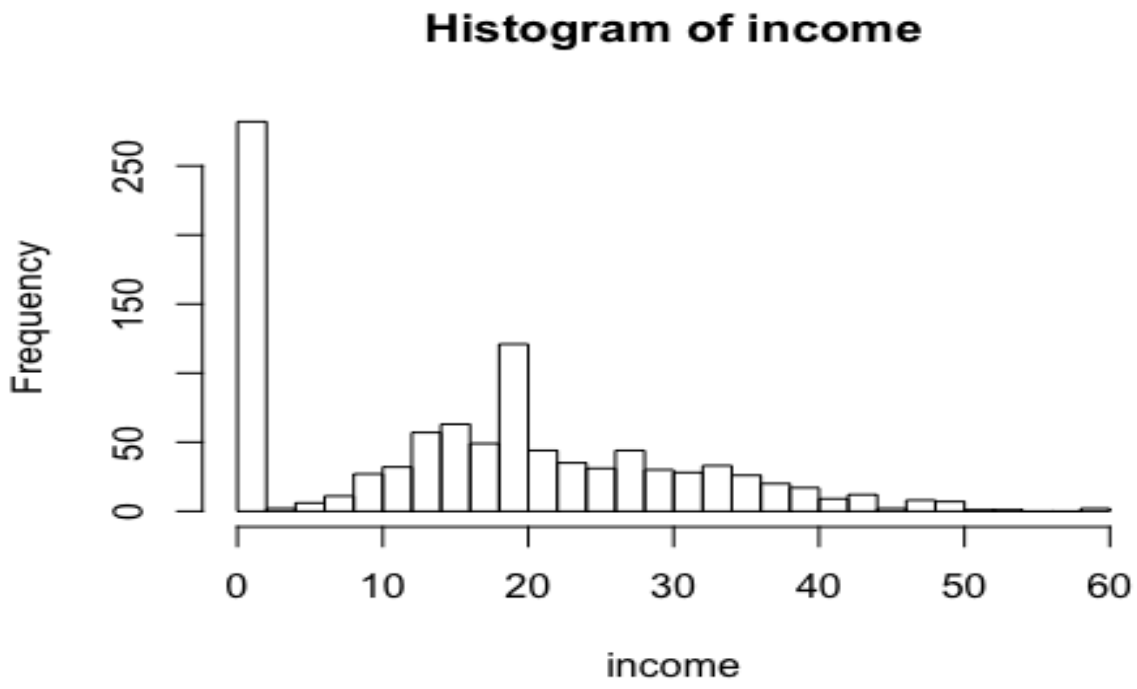
(a) A colleague now points out to you that some women may have incentives in these settings to misreport their actual income. Plot a histogram of reported income (using the default settings which should give you 33 bins) and look for anything that might support such a claim. What assumption is called into question if women are truly misreporting in this manner (choose the best answer below)?

Solutions.

According to the histogram of income below, quite a fraction of women in this sample have reported their income close to 0.

Because the reported income would be used as the forcing variable by selecting a cut point, if participants misreport their income, the value of the forcing variable is then manipulated, thus, it cannot be considered as a sharp/clean RD design anymore. Accordingly, the assumption of ignorability of treatment assignment given the forcing variable around the cutoff could be questionable.

```
hist(income, breaks = 33)
```



(b) Another colleague points out to you that several other government programs (including food stamps and Headstart) have the same income threshold for eligibility. How might this knowledge impact your interpretation of your results?

Solutions.

In order for a RD design to be valid, it is important to determine the cut-point of the forcing variable. As for the current study, the same income eligible threshold adopted by other government programs is supportive and a cross-validation for our threshold selection. Thus, the internal validity of current RDD study is enhanced, and the interpretation of our results is more credible.

Question 9. Researcher role. Thinking about the assumptions?

What are the three most important assumptions for causal estimates in questions 5 and 6?

Solutions.

The three most important assumptions are:

1) **A clear discontinuity in the probability of receiving treatment must exist at the cut-point of the forcing variable;**
2) **Each participant's value of the forcing variable was not manipulated;**
3) **The treatment assignment is ignorable given the forcing variable within a narrow interval around the threshold.**

Question 10.

Provide a causal interpretation of your estimate in Question 6biii.

Solutions.

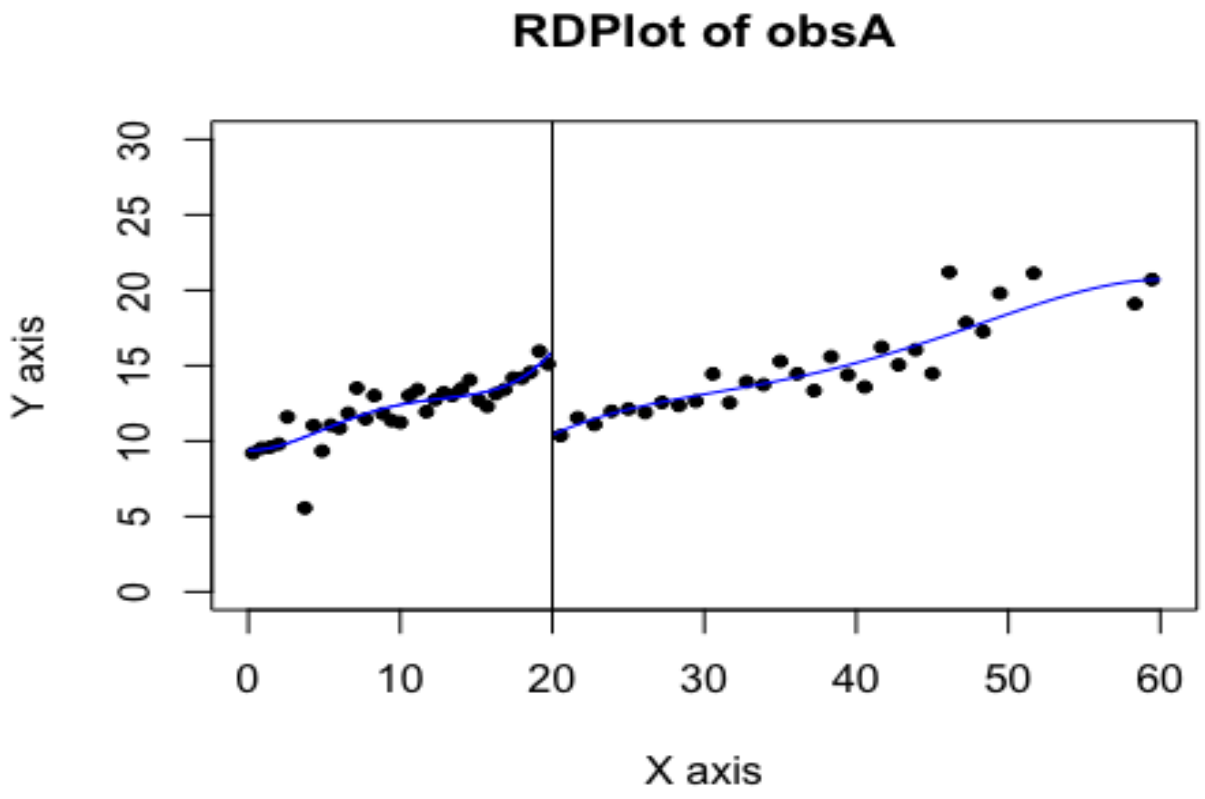
In terms of the estimation from 6biii, the causal interpretation is:

For new mothers in households with income equals to $20, 000 among the analysis sample and received the pre- and post-natal (through 2 years post birth) health care, the health measure of their children at age 3 were about 4.1 points higher than had they not attended the pre- and post-natal health care program.

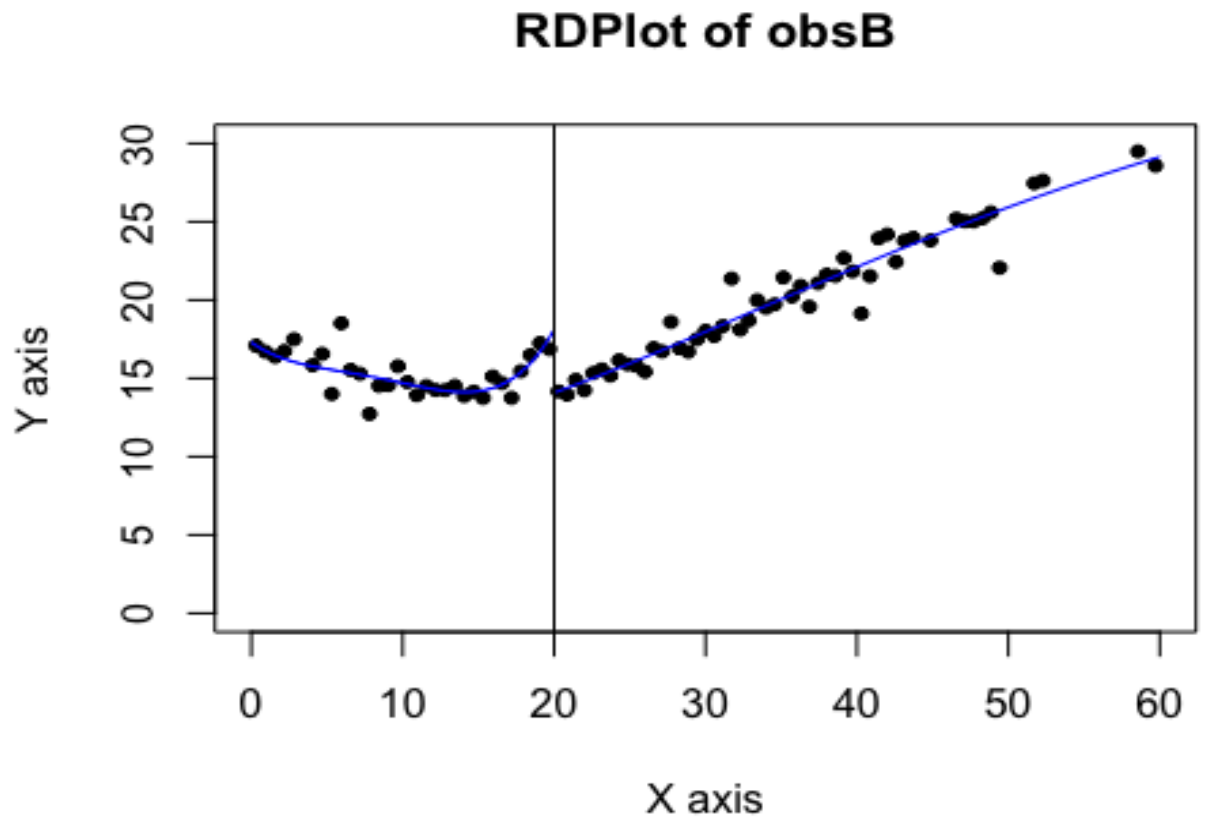
Challenge Problem.

Use the rdrobust package in R to choose an optimal bandwidth for the data in World B (can use any of the approaches they support). Two points for each one you try for a maximum of 6 points.

```
library(rdrobust)
rdplot(obsA$y, obsA$income, c = 20, binselect = "esmv", x.lim = c(0, 60), y.l
im = c(0, 30), title = "RDPlot of obsA")
```



```
rdplot(obsB$y, obsB$income, c = 20, binselect = "esmv", x.lim = c(0, 60), y.l
im = c(0, 30), title = "RDPlot of obsB")
```




```
# Way 1:  MSE-optimal method (same bandwith on both sides of the cutoff)
rd.bw1 <- rdbwselect(obsB$y, obsB$income, kernel = "triangular", c = 20, p =
1, bwselect = "mserd")
summary(rd.bw1)

## Call: rdbwselect
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 650         350
## Order est. (p)                   1           1
## Order bias  (q)                  2           2
##
## ===================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## ===================================================
##      mserd     4.648      4.648      8.271      8.271
## ===================================================
```

```r
# Way 2: MSE(Mean Square Error)-optimal method - different bandwidth on each
side of the cutoff
rd.bw2 <- rdbwselect(obsB$y, obsB$income, kernel = "triangular", c = 20, p =
1, bwselect = "msetwo")
summary(rd.bw2)
```

```
## Call: rdbwselect
##
## Number of Obs.                 1000
## BW type                      msetwo
## Kernel                    Triangular
## VCE method                        NN
##
## Number of Obs.                  650         350
## Order est. (p)                    1           1
## Order bias  (q)                   2           2
##
## ===========================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## ===========================================================
##     msetwo     4.138     11.423      7.520     18.441
## ===========================================================
```

```r
# Way 3: CER(Coverage Error Rate)-optimal-different bandwidth on each side of
 the cutoff
rd.bw3 <- rdbwselect(obsB$y, obsB$income, kernel = "triangular", c = 20, p =
1, bwselect = "cercomb2")
summary(rd.bw3)
```

```
## Call: rdbwselect
##
## Number of Obs.                 1000
## BW type                    cercomb2
## Kernel                    Triangular
## VCE method                        NN
##
## Number of Obs.                  650         350
## Order est. (p)                    1           1
## Order bias  (q)                   2           2
##
## ===========================================================
##                   BW est. (h)    BW bias (b)
##             Left of c Right of c  Left of c Right of c
## ===========================================================
##   cercomb2     3.290      3.377      8.271      8.676
## ===========================================================
```