Causality and Program evalustion_Assignment

Sunyoung Ji, ID: 229979 (TU Dortmund)

2022-07-24

Used libraries are given below:

0. Set up

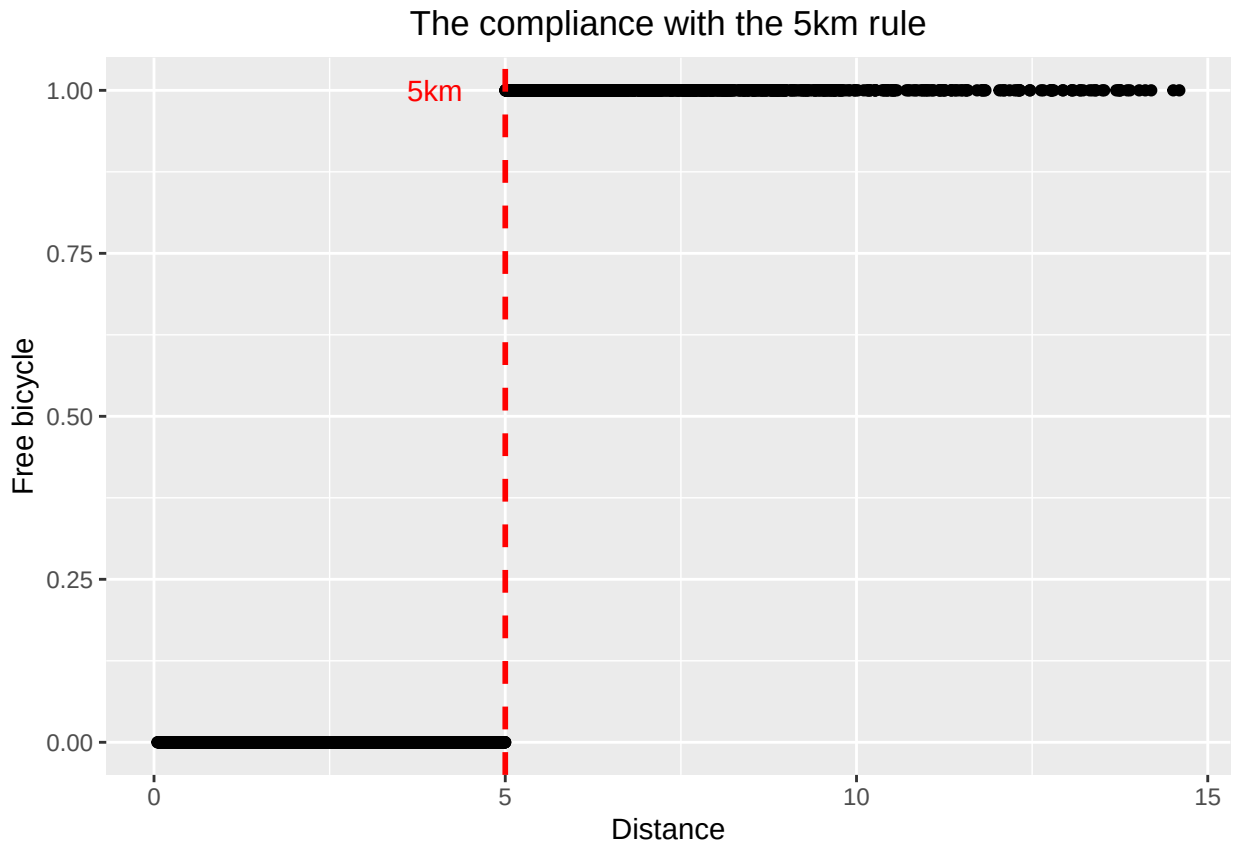
```
ebike <- read_csv("dataset1.csv")
```

```
## Rows: 4900 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): student_id, neighborhood, age, distance, free_bicycle, income_hh, s...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ebike <- as_tibble(ebike)
view(ebike)
```

1. Task(a)

```
threshold <- ggplot(ebike,
                    aes(x = distance, y = free_bicycle)) +
  geom_point() +
  labs(x = "Distance", y = "Free bicycle",
       title = ("The compliance with the 5km rule")) +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_vline(xintercept = 5, colour = "red",
             size = 1, linetype = "dashed") +
  annotate("text", x = 4, y = 1, label = "5km",
           size=4, color = "red")
threshold
```



This graph illustrates that "dataset1" satisfies the rule of distance.

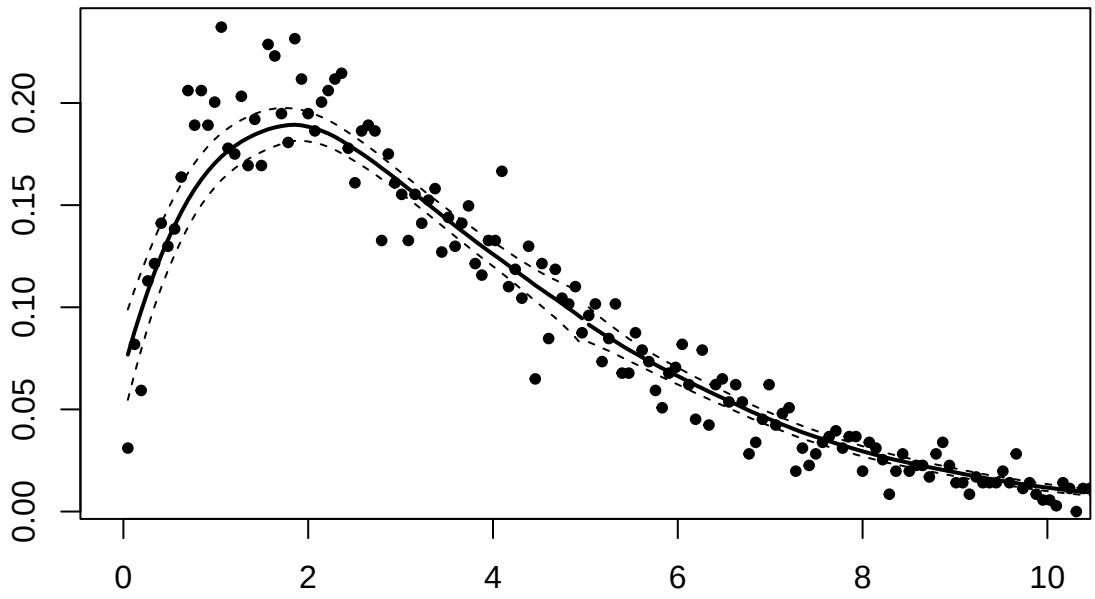
Task(b)

From (a), we can see that the treatment(`free_bicycle`) is assigned by the threshold(5km Rule). Therefore, the impact of free ebike policy can be estimated by Sharp Regression Discontinuity Designs. It should hold the following assumptions: 1) There is a discontinuity in the probability of treatment at the threshold 2) Individuals' value of the treatment variable was not manipulated 3) Continuity of potential outcomes around threshold

Task(c)

McCrary Sorting Test: We can check that the score would not manipulated in the McCrary test. The density of distance almost continuous.

```
DCdensity(ebike$distance, cutpoint = 5, ext.out = FALSE, plot = TRUE)
```




```
## [1] 0.94085
```

Task (d)

Placebo test with age and number of roommates as dependent variables:

```
plac_age <-lm(age ~ distance, data = ebike)
plac_room <-lm(number_roomates ~ distance, data = ebike)

plac_linear <- cbind(plac_age$coefficients[2], plac_room$coefficients[2])
colnames(plac_linear) <- c("age", "roommates")
rownames(plac_linear) <- "plac_linear"

plac_linear <- round(plac_linear, 4)
```

Through the placebo regression, we can know that these dependent variables have almost zero linear dependency with other variables.

However, when these dependent variables are applied for the same model in (e), the placebo effects increase:

```
plac_age_rd <- rdd_data(y = ebike$age, x = ebike$distance,
                        cutpoint = 5)
reg_para_age <- rdd_reg_lm(rdd_object=plac_age_rd)

plac_room_rd <- rdd_data(y = ebike$number_roomates, x = ebike$distance,
                        cutpoint = 5)
reg_para_room <- rdd_reg_lm(rdd_object=plac_room_rd)

plac_rd <- cbind(reg_para_age$coefficients, reg_para_room$coefficients)
colnames(plac_rd) <- c("age", "roommates")
plac_rd <- round(plac_rd[2,], 4)
```

We can comparison two results from two models:

```
rbind(plac_linear, plac_rd)
```

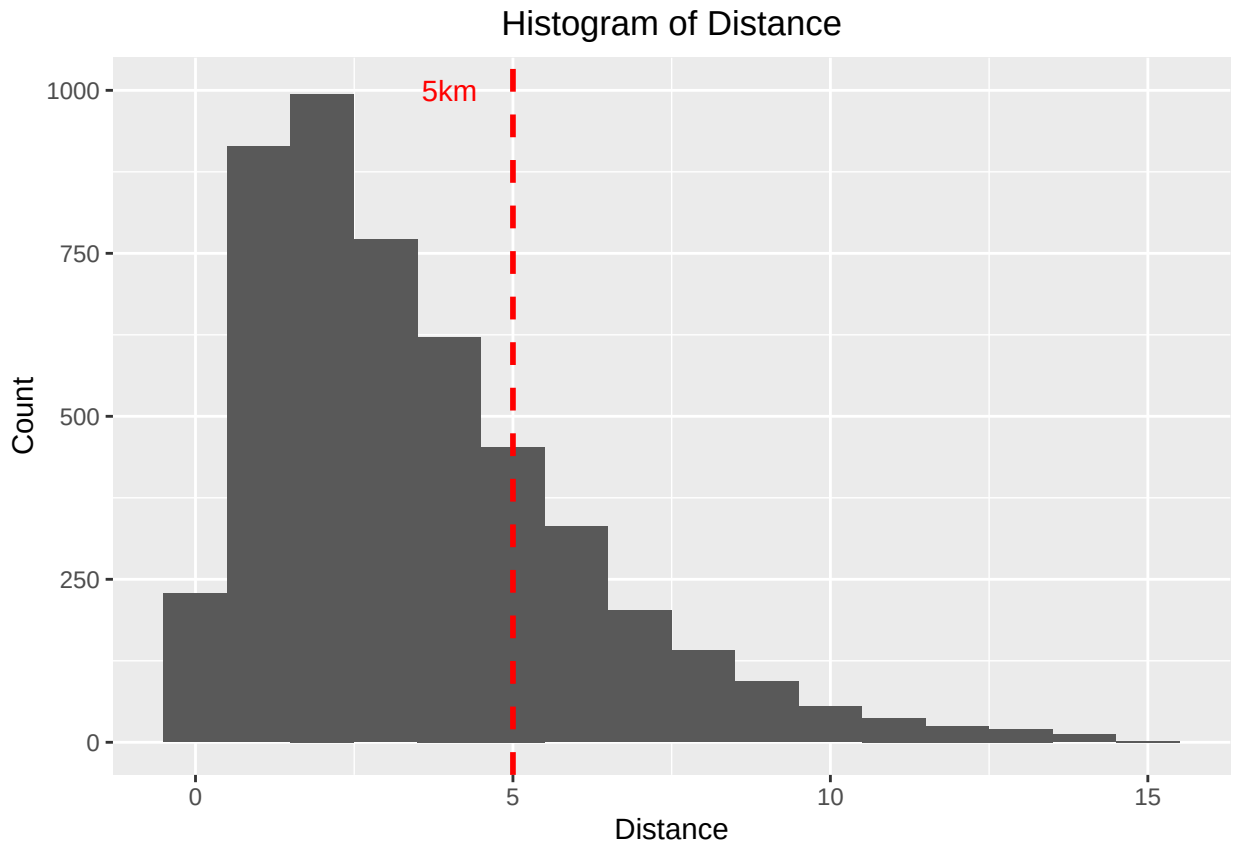
```
##                 age roommates
## plac_linear  0.0083   -0.0022
## plac_rd     -0.0475    0.0119
```

Task (e)

```
distance_hist <- ggplot(ebike, aes(distance)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Distance", y = "Count",
       title = ("Histogram of Distance")) +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_vline(xintercept = 5, colour = "red",
             size = 1, linetype = "dashed") +

  annotate("text", x = 4, y = 1000, label = "5km",
           size=4, color = "red")

distance_hist
```



As histogram of distance shows, we have observations that are distributed very differently depending on the distance. Including more observations farther away from threshold would generate excessive bias. Thus, I choose a local linear regression model and bandwidth 1.

1) The RDD model:

$$Y = \alpha + \tau * D + \beta_1(X - c) + \beta_2 * D(X - c) + \epsilon$$

2) Finding a bandwidth: Kernel selection:

```
rdd_data <- rdd_data(y = ebike$score, x = ebike$distance,
                     cutpoint = 5)

bandwidth <- rdd_bw_cct_estim(rdd_data)
bandwidth
```

```
## Call: rdbwselect
##
## Number of Obs.                 4900
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                 3776         1124
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## Unique Obs.                    3776         1124
```

3) RD simulation:

```
reg_para_bw1 <- rdd_reg_lm(rdd_object=rdd_data, bw = 1)
reg_para_bw1
```

```
## ### RDD regression: parametric ###
##  Polynomial order:  1
##  Slopes:  separate
##  Bandwidth:  1
##  Number of obs: 921 (left: 541, right: 380)
##
##  Coefficient:
##   Estimate Std. Error t value  Pr(>|t|)
## D  0.97854    0.24799   3.946 8.555e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

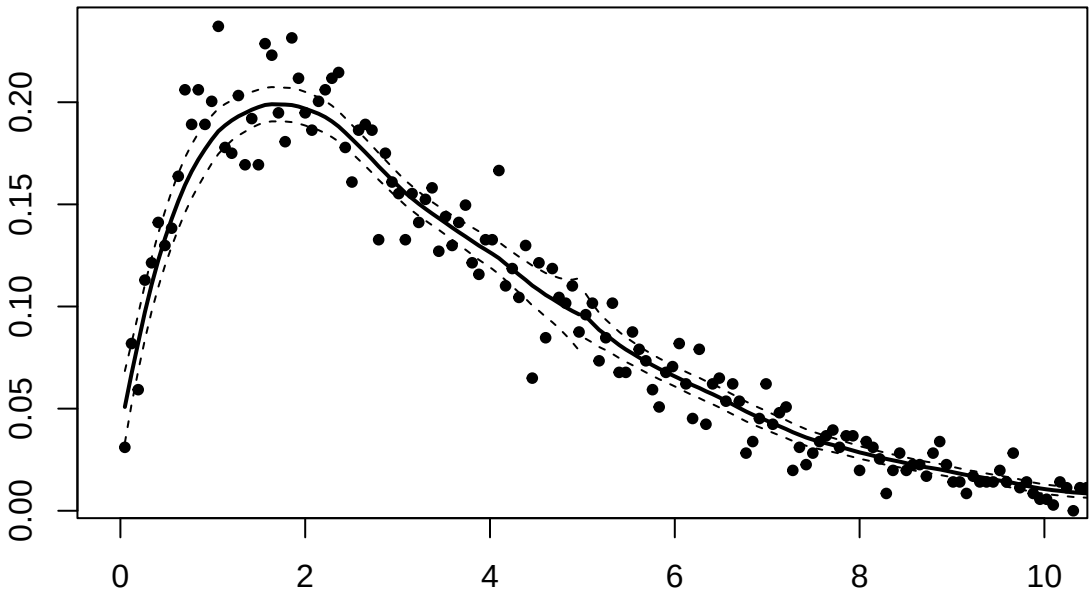
```
ebike %>%
  select(distance, score) %>%
  mutate(threshold = as.factor(ifelse(distance >= 5, 1, 0))) %>%
  ggplot(aes(x = distance, y = score, color = threshold)) +
  geom_point(size = 0.3) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_brewer(palette = "Accent") +
  guides(color = "none") +
  geom_vline(xintercept = 5, color = "black",
             size = 1, linetype = "dashed") +
  labs(y = "Exam Score",
       x = "Distance")
```

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

Task(f)

According to result from McCrary test, the external validity of the model in (e) is satisfied.

```
DCdensity(rdd_data$x, cutpoint = 5, ext.out = FALSE, plot = TRUE, bw = 1)
```




```
## [1] 0.9433047
```