Regression Discontinuity Design

Saumya Seth

2023-12-07

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(rdrobust)
library(gt)

set.seed(123)
```

1. Scenario

I would use RDD to evaluate a policy which provides relief, say in the form of free child care, to people below a certain income threshold (say 96 dollars per week). Say the policy was intended to help increase the level of economic satisfaction for those below the threshold. For evaluation, I would want to know if the free child care caused them to have a higher/lower economic satisfaction. I would do this by comparing the economic satisfaction levels of those whose incomes were below 96 dollars and those whose were above.

Say within a certain income 'bandwidth' (which includes the threshold of 96 dollars), we can't accurately determine whether the values of economic satisfaction that we observe came from the group which got free child care or from the group which didn't, given the income. Thus, if we compare the economic satisfaction levels of these people, within this bandwidth, who ONLY differ in terms of their income, which is what the policy based its implementation of the intervention on, we would accurately be able to assess the impact the intervention had (given RDD assumptions hold).

2. DGP

World A

$$X \sim \mathrm{N}(100, 15^2)$$

$$X_{scaled} = X - 96$$

$$Z = \begin{cases} 1, & \text{if } X \leq 96 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{We have to find } \mathrm{E}[Y(1) - Y(0) | X = 96]$$

$$Y(0) = X_{scaled} + \mathrm{N}(90, 5^2)$$

$$Y(1) = X_{scaled} + \mathrm{N}(110, 5^2) + X_{scaled} * Z$$

$$Y = \begin{cases} Y(1), & \text{if } Z = 1 \\ Y(0), & \text{otherwise} \end{cases}$$

World B

$$X \sim \mathrm{N}(100, 15^2)$$

$$X_{scaled} = X - 96$$

$$Z = \begin{cases} 1, & \text{if } X \leq 96 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{We have to find } \mathrm{E}[Y(1) - Y(0) | X = 96]$$

$$Y(0) = X_{scaled} + \mathrm{N}(90, 5^2) + 10 * normalized((X_{scaled})^2)$$

$$Y(1) = X_{scaled} + \mathrm{N}(110, 5^2) + X_{scaled} * Z$$

$$Y = \begin{cases} Y(1), & \text{if } Z = 1 \\ Y(0), & \text{otherwise} \end{cases}$$

World C

$$X1 \sim \mathrm{N}(100, 15^2)$$

$$X2 \sim \mathrm{Unif}(95.9, 96)$$

$$X = c(X1, X2)$$

$$X_{scaled} = X - 96$$

$$Z = \begin{cases} 1, & \text{if } X \leq 96 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{We have to find } \mathrm{E}[Y(1) - Y(0) | X = 96]$$

$$Y(0) = X_{scaled} + \mathrm{N}(90, 5^2)$$

$$Y(1) = X_{scaled} + \mathrm{N}(110, 5^2) + X_{scaled} * Z$$

$$Y = \begin{cases} Y(1), & \text{if } Z = 1 \\ Y(0), & \text{otherwise} \end{cases}$$

For all worlds: - X = income - threshold = 96 - Z = eligibility (treatment assignment) - Y(1) and Y(0) = potential economic satisfaction (potential outcomes) - Y = observed economic satisfaction (observed outcome)

3. Methods and Estimands

a)

The RDD approach tries to estimate the causal effect of only one confounding covariate on an outcome. Ideally, the observations on one side of the threshold of this covariate are treated and the observations on the other side are not (a sharp RDD design). A problem with this approach is that there is no balance and overlap over this one covariate because it is this one covariate which is determining treatment receipt. So, what RDD does is that it restricts our data to a certain bandwidth of the covariate so that it is possible to compare observations on either side of the threshold while reasonably assuming that there are no other confounders (we should use RDD only when the observations on either side of the threshold are identical on average, apart from treatment receipt). Thus, RDD only helps us estimate the causal effect of observations at the threshold.

Ideally, the estimand that RDD is trying to target is the ATE at the cutoff. Following the example provided in the 'Scenario' section, we want to be able to determine if and by how much the free child care intervention increased/decreased economic satisfaction. However, the RDD approach cannot help estimate this; it can only help estimate the increase/decrease in economic satisfaction for those who earned 96 dollars per week (the threshold level). This can be done because the approach assumes that observations falling on either side of a narrow interval of the cut-off/threshold happens due to random chance. Thus, allowing for it to be reasonable to compare the groups on either side of the threshold to assess the impact treatment receipt has on the individuals (we have essentially 'created a randomized experiment' at the cutoff where the only confounder is the running variable).

b)

Using the correct linear regression model to estimate the results:

```
# defining the threshold
discontinuity_point <- 96
# choosing an arbitrary bandwidth around the threshold
bandwidth <- c(90, 102)
# restricting data to be within this threshold
data_restricted <- data %>% filter(x > bandwidth[1] & x < bandwidth[2])
# fitting the correct model on the restricted data
fit <- lm(observed_outcome ~ x_scaled * factor(eligible), data = data_restricted)
summary(fit)

# the estimated causal effect using linear regression with interaction
# (the correct model) at the threshold is:
summary(fit)$coef[3, 1]
```

Using the correct Rdrobust model to estimate the results:

```
# fitting the correct model
rdd_implementation <- rdrobust(observed_outcome, x,
  c = discontinuity_point,
  bwselect = "msetwo", p = 1
)
rdplot(observed_outcome, x, c = discontinuity_point, p = 1)
# bandwidths are unequal
rdd_bandwidths <- rdrobust(observed_outcome, x,
  c = discontinuity_point,
  bwselect = "msetwo", p = 1
)$bws
rdd_bandwidths[1, 1] == rdd_bandwidths[1, 2]

# the estimated causal effect using rdrobust with a local linear specification
# (the correct model) at the threshold is:
rdd_implementation$coef[3]

# (for the scenario described, we should ideally take the negative of the
# treatment effect as rdrobust assumes that treatment has been given to
# those on the right-hand side of the threshold)
```

4. Assumptions

Assumption 1:

Being/not being eligible for free childcare (i.e. having an income of lower/higher than 96 dollars per week) in a narrow interval on either side of the threshold of 96 dollars is arbitrary and can be attributed to random chance. This can be assessed by checking the densities of the plot for income - they should not be different on either side of the threshold, on average. This is how this assumption was satisfied in worlds A and B. This assumption was violated in world C as there is a 'bump' in the number of observations one can see just before the income level of 96 (the density plot is shown later in the report). This might have happened due to various reasons, such as people may have falsely reported their income to be less than $96 per week to get free child care. In this case, getting free child care is not ignorable given income in a narrow income interval around 96; another factor, namely the reporting of false income levels, has violated the ignorability assumption.

In worlds where this assumption is satisfied, getting free child care is ignorable given income in a narrow interval of the income cutoff of 96 dollars.

This seems like a reasonable assumption to make; however, determining the exact 'narrow interval' for income may be tricky. There also might be other factors apart from being on a certain side of an income threshold which can effect treatment receipt (as discussed above when violating the structural assumption in world C) EVEN in that narrow interval. In that case, the assumption of ignorability of the treatment in the interval will not be satisfied resulting in a situation where RDD might not be the best approach to get a causal estimate.

Assumption 2:

To get the correct causal estimate, the functional form for the potential outcomes Y(1) and Y(0) given X must be accurate in the interval. This means that we must know the correct functional form of everyone's economic satisfaction in the specified income interval - both if they had received free child care and if they hadn't. This assumption was satisfied in worlds A and C since we included an interaction term in creating our potential outcomes (income * eligibility). Thus, when estimating our causal effect using linear regression

with an interaction term and rdrobust with a linear specification, we will be able to retrieve the correct causal estimates.

In world B, we specified our potential outcome Y(0) to have a quadratic form - leading to linear regression with an interaction term (income * eligibility) and rdrobust with a linear specification, being incorrect functional forms of the potential outcomes and thus leading us to the wrong causal estimates.

This does not seem like a reasonable assumption to make; as it may be difficult to know the exact functional trend which relates income to economic satisfaction in the situation where free child care is the intervention. There might not be enough literature to assess the functional forms one can assume in this context. Additionally, the larger the income interval, the less plausible this assumption becomes. MAYBE the functional form IS linear in a small interval but maybe it ISN'T in a larger one. One must choose the correct interval along with the right functional form; one choice is not independent of the other.

Assumption 3:

We cannot observe if SUTVA has been satisfied from the data. We need our study design to be made in such a way that SUTVA doesn't end up biasing our results. SUTVA would be satisfied in our example if the economic satisfaction due to getting free childcare for one person does not affect the economic satisfaction of another. This may not always be the case, though. Sharing satisfaction levels with others may affect what one thinks and reports of their own satisfaction.

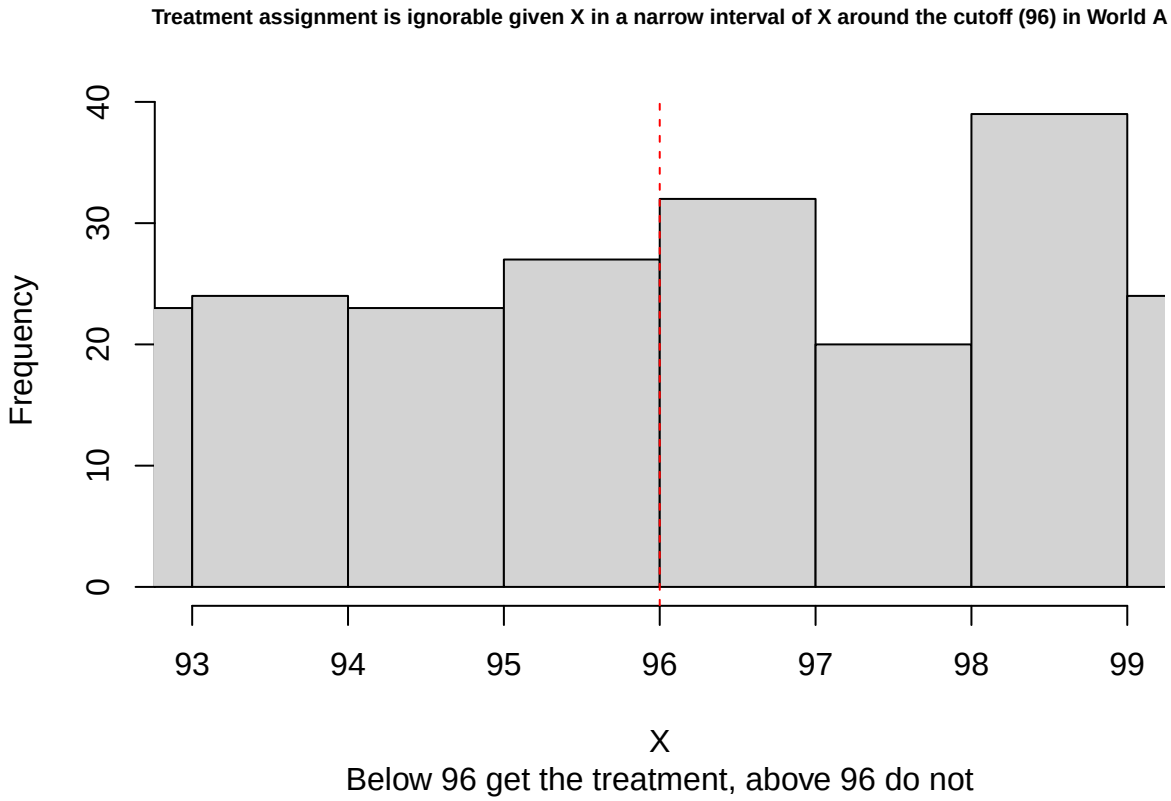
5. Code

World A

To ensure that the assumption of the treatment assignment being ignorable given X in an interval around 96 is met, we create our data in this manner:

```
n <- 1000
xA <- rnorm(n, mean = 100, sd = 15)
x_scaledA <- xA - 96
discontinuity_pointA <- 96
eligibleA <- ifelse(xA <= discontinuity_pointA, 1, 0)
```

```
hist(xA,
  main = "Treatment assignment is ignorable given X in a narrow interval of X around the cutoff (96) in
  cex.main = 0.7, xlab = "X", breaks = 100, xlim = c(93, 99),
  sub = "Below 96 get the treatment, above 96 do not"
)
abline(v = discontinuity_pointA, lty = 2, col = "red")
```

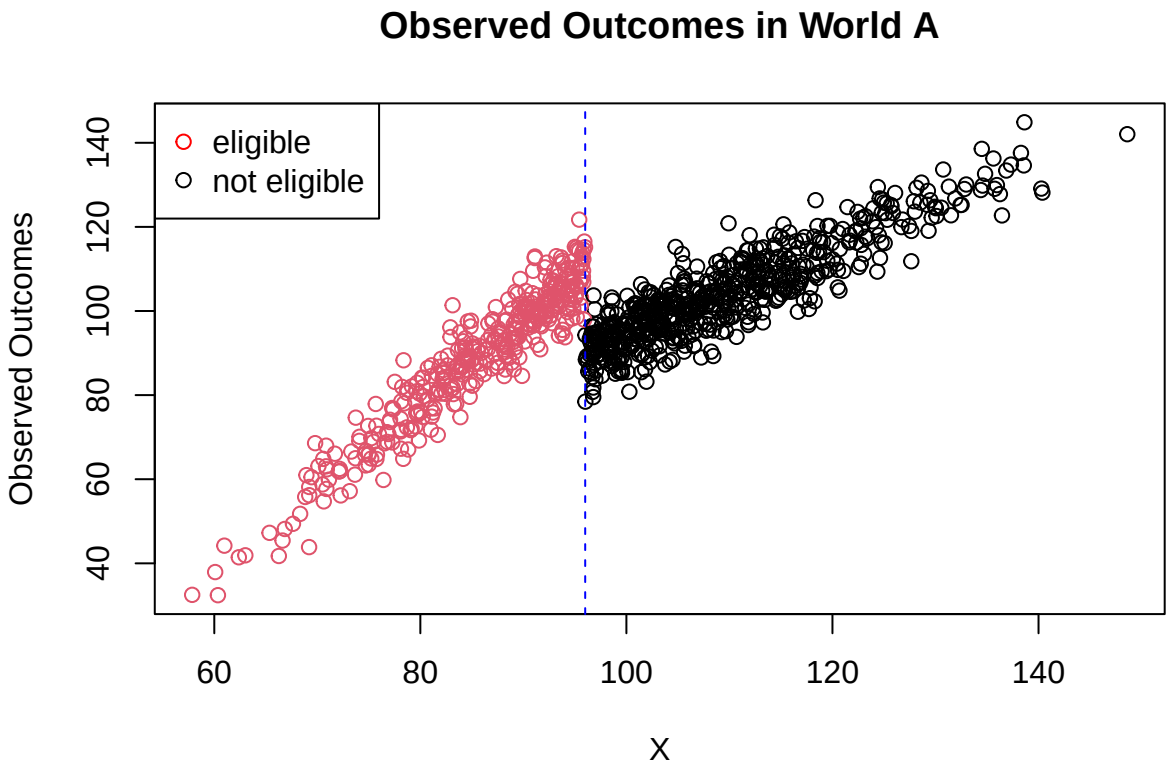


To ensure that the assumption of having the correct functional form for E[Y(0) | X, Z=0] and E[Y(1) |X,Z=1] in the interval around 96 is met (and since we will be using linear regression that includes an interaction term between the running variable and the treatment group - the correct model - to estimate the causal effect), we create potential outcomes in this manner:

```
potential_outcomeA_y0 <- x_scaledA + rnorm(n, mean = 90, sd = 5)
potential_outcomeA_y1 <- x_scaledA + rnorm(n, mean = 110, sd = 5) +
  x_scaledA * eligibleA
observed_outcomeA <- ifelse(eligibleA == 1, potential_outcomeA_y1, potential_outcomeA_y0)
worldA <- data.frame(xA, x_scaledA, eligibleA, potential_outcomeA_y0, potential_outcomeA_y1, observed_ou
head(worldA)
```

```
##          xA  x_scaledA eligibleA potential_outcomeA_y0 potential_outcomeA_y1
## 1  91.59287 -4.4071347         1              80.61387              98.62771
## 2  96.54734  0.5473377         0              85.34756             111.73203
## 3 123.38062 27.3806247         0             117.29072             134.67268
## 4 101.05763  5.0576259         0              94.39675             121.15376
## 5 101.93932  5.9393160         0              83.19260             116.81000
## 6 125.72597 29.7259748         0             124.92884             136.64963
##   observed_outcomeA
## 1          98.62771
## 2          85.34756
## 3         117.29072
## 4          94.39675
## 5          83.19260
## 6         124.92884
```

```r
plot(xA, worldA$observed_outcomeA,
  col = factor(worldA$eligible),
  main = "Observed Outcomes in World A",
  xlab = "X",
  ylab = "Observed Outcomes"
)
abline(v = discontinuity_pointA, lty = 2, col = "blue")
legend("topleft", legend = c("eligible", "not eligible"), col = c("red", "black"), pch = 1)
```

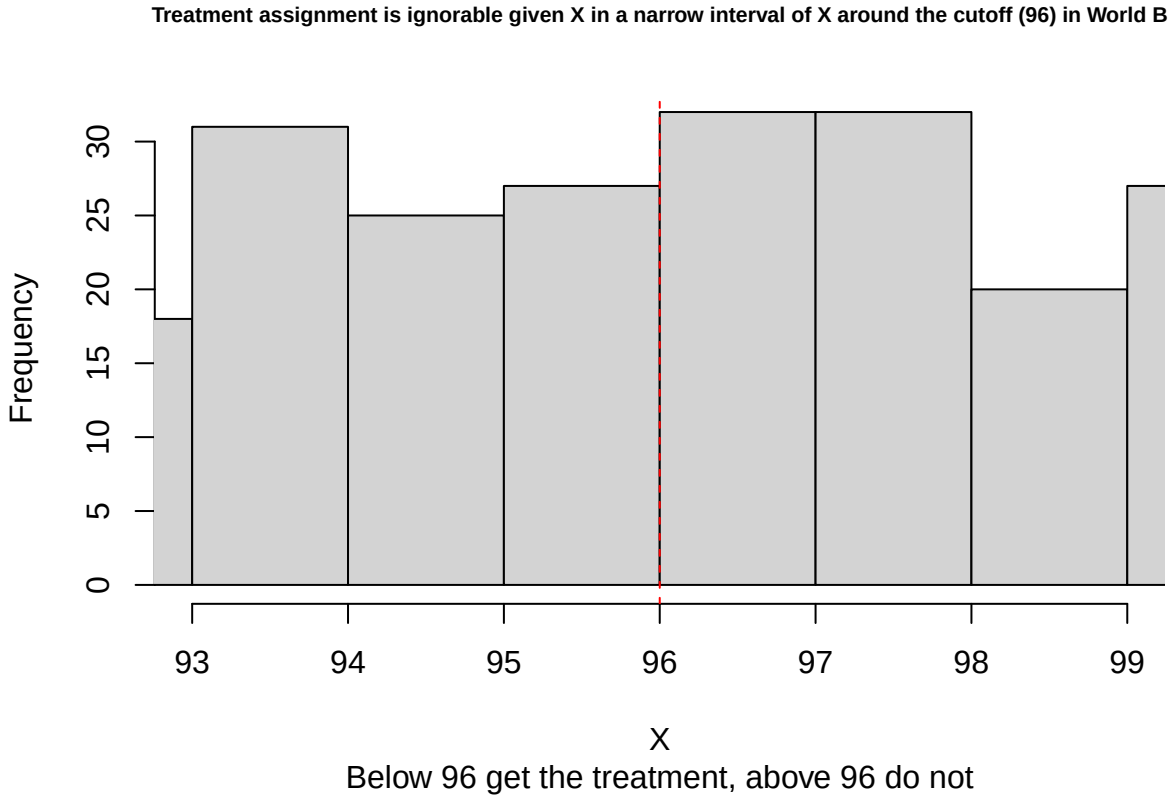


World B

To ensure that the assumption of the treatment assignment being ignorable given X in an interval around 96 is met, we create our data in this manner:

```r
n <- 1000
xB <- rnorm(n, mean = 100, sd = 15)
x_scaledB <- xB - 96
discontinuity_pointB <- 96
eligibleB <- ifelse(xB <= discontinuity_pointB, 1, 0)
```

```r
hist(xB,
  main = "Treatment assignment is ignorable given X in a narrow interval of X around the cutoff (96) in
  cex.main = 0.7, xlab = "X", breaks = 100, xlim = c(93, 99),
  sub = "Below 96 get the treatment, above 96 do not"
)
abline(v = discontinuity_pointB, lty = 2, col = "red")
```

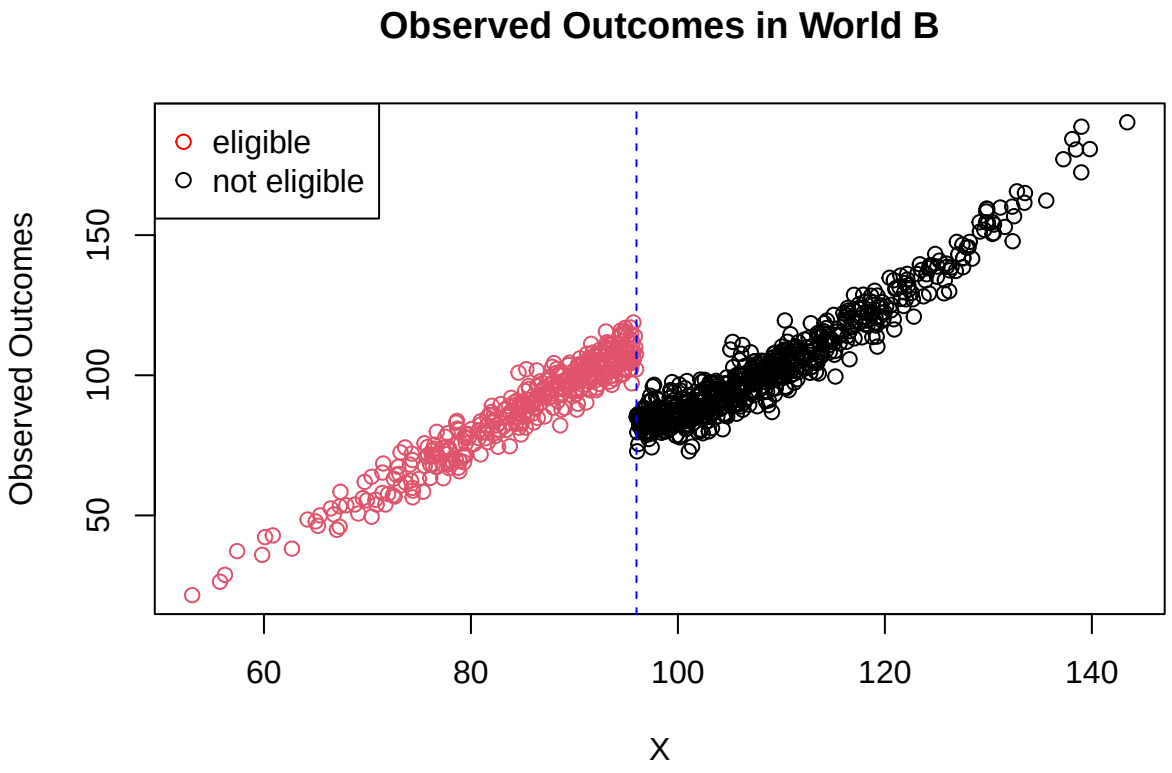


To ensure that the assumption of having the correct functional form for E[Y(0) | X, Z=0] and E[Y(1) |X,Z=1] in the interval around 96 is VIOLATED (and since we will be using linear regression that includes an interaction term between the running variable and the treatment group - the correct model - to estimate the causal effect), we create potential outcomes in this manner:

```
potential_outcomeB_y0 <- x_scaledB + rnorm(n, mean = 90, sd = 5) +
  10 * scale(x_scaledB^2)
potential_outcomeB_y1 <- x_scaledB + rnorm(n, mean = 110, sd = 5) +
  x_scaledB * eligibleB
observed_outcomeB <- ifelse(eligibleB == 1, potential_outcomeB_y1, potential_outcomeB_y0)
worldB <- data.frame(xB, x_scaledB, eligibleB, potential_outcomeB_y0, potential_outcomeB_y1, observed_ou
head(worldB)
```

```
##          xB  x_scaledB eligibleB potential_outcomeB_y0 potential_outcomeB_y1
## 1  97.74539   1.745388         0              85.55475             109.27452
## 2  95.08364  -0.916357         1              85.09295             113.80525
## 3  78.27752 -17.722479         1              78.02600              68.82029
## 4  89.54073  -6.459269         1              71.13613             104.48656
## 5 138.97735  42.977353         0             172.39067             157.55831
## 6  99.43877   3.438775         0              97.56203             115.11443
##   observed_outcomeB
## 1          85.55475
## 2         113.80525
## 3          68.82029
## 4         104.48656
## 5         172.39067
## 6          97.56203
```

```r
plot(xB, worldB$observed_outcomeB,
  col = factor(worldB$eligible),
  main = "Observed Outcomes in World B",
  xlab = "X",
  ylab = "Observed Outcomes"
)
abline(v = discontinuity_pointB, lty = 2, col = "blue")
legend("topleft", legend = c("eligible", "not eligible"), col = c("red", "black"), pch = 1)
```

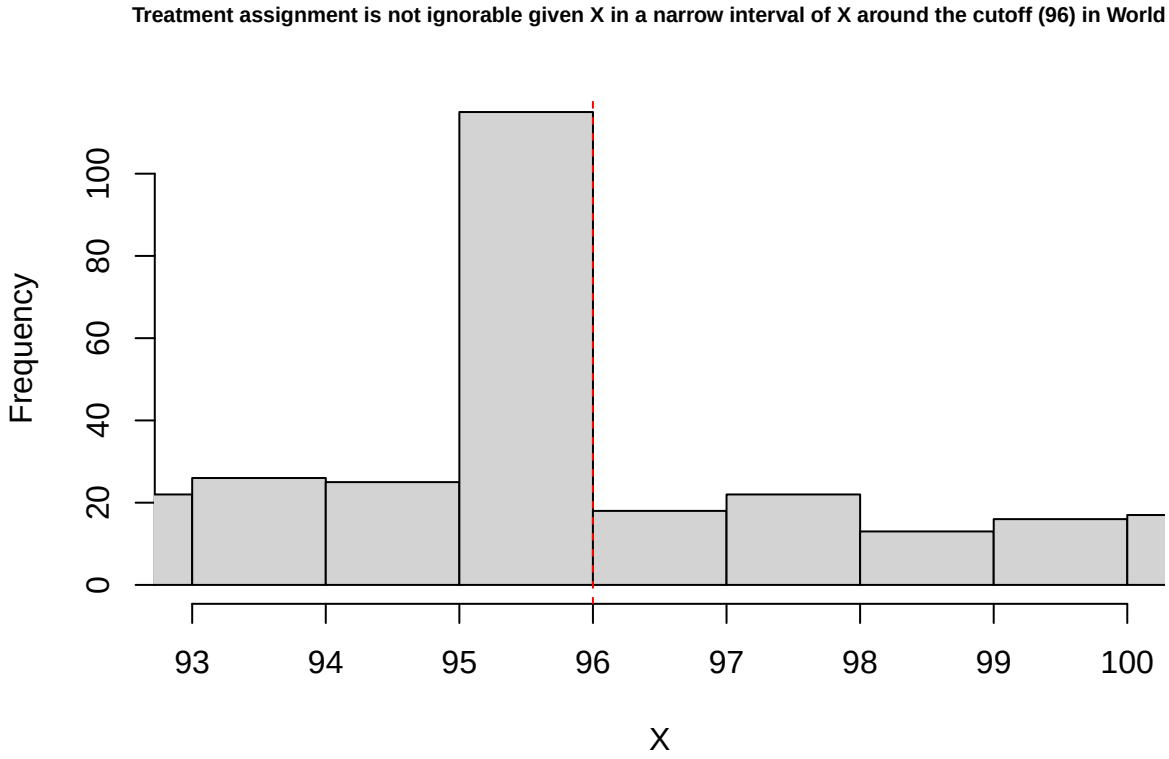


World C

To ensure that the assumption of the treatment assignment being ignorable given X in an interval around 96 is VIOLATED, we create our data in this manner:

```r
x1 <- rnorm(0.9 * n, mean = 90, sd = 15)
x2 <- runif(0.1 * n, 95.9, 96)
xC <- c(x1, x2)
x_scaledC <- xC - 96
discontinuity_pointC <- 96
eligibleC <- ifelse(xC <= discontinuity_pointC, 1, 0)
```

```r
hist(xC,
  main = "Treatment assignment is not ignorable given X in a narrow interval of X around the cutoff (96)",
  cex.main = 0.7, xlab = "X", breaks = 100, xlim = c(93, 100)
)
abline(v = discontinuity_pointC, lty = 2, col = "red")
```



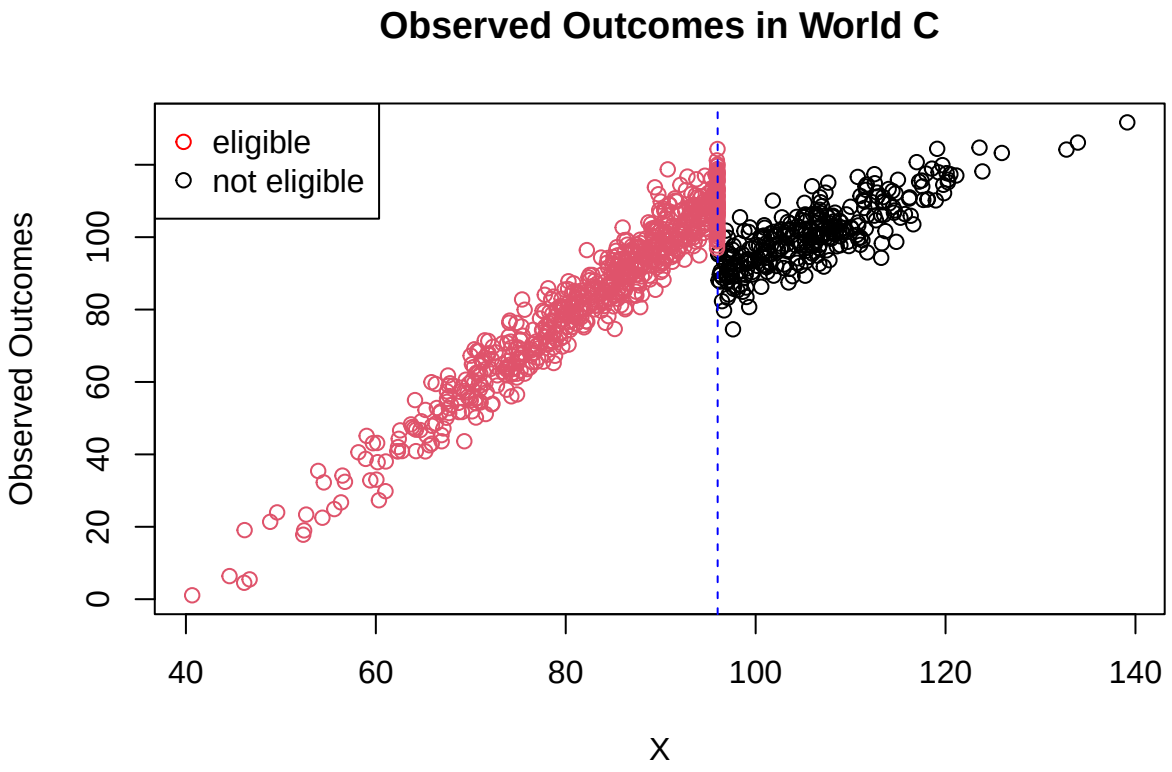
To ensure that the assumption of having the correct functional form for E[Y(0) | X, Z=0] and E[Y(1) |X,Z=1] in the interval around 96 is met (and since we will be using linear regression that includes an interaction term between the running variable and the treatment group - the correct model - to estimate the causal effect), we create potential outcomes in this manner:

```
potential_outcomeC_y0 <- x_scaledC + rnorm(n, mean = 90, sd = 5)
potential_outcomeC_y1 <- x_scaledC + rnorm(n, mean = 110, sd = 5) +
  x_scaledC * eligibleC
observed_outcomeC <- ifelse(eligibleC == 1, potential_outcomeC_y1, potential_outcomeC_y0)
worldC <- data.frame(xC, x_scaledC, eligibleC, potential_outcomeC_y0, potential_outcomeC_y1, observed_ou
head(worldC)
```

```
##          xC  x_scaledC eligibleC potential_outcomeC_y0 potential_outcomeC_y1
## 1  79.51158 -16.488422         1              69.80904              80.97910
## 2 104.94677   8.946773         0             104.65377             122.33280
## 3  79.60882 -16.391180         1              76.60897              79.95357
## 4  88.44775  -7.552246         1              91.52958              96.94299
## 5  99.05799   3.057991         0              93.60735             117.91459
## 6  80.87933 -15.120675         1              74.92447              78.00104
##   observed_outcomeC
## 1          80.97910
## 2         104.65377
## 3          79.95357
## 4          96.94299
## 5          93.60735
## 6          78.00104
```

```
plot(xC, worldC$observed_outcomeC,
  col = factor(worldC$eligible),
```

```
  main = "Observed Outcomes in World C",
  xlab = "X",
  ylab = "Observed Outcomes"
)
abline(v = discontinuity_pointC, lty = 2, col = "blue")
legend("topleft", legend = c("eligible", "not eligible"), col = c("red", "black"), pch = 1)
```



6. Causal Estimate and Interpretation

World A

Estimating a causal effect for the effect at the threshold using a linear regression that includes an interaction term between the running variable and the treatment group.

Choosing the bandwidth arbitrarily in this case, we have:

```
bandwidthA <- c(93, 99)
worldA_restricted <- worldA %>% filter(xA > bandwidthA[1] & xA < bandwidthA[2])
fitA <- lm(observed_outcomeA ~ x_scaledA * factor(eligibleA), data = worldA_restricted)
summary(fitA)
```

```
##
## Call:
## lm(formula = observed_outcomeA ~ x_scaledA * factor(eligibleA),
##     data = worldA_restricted)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -11.713  -2.558  -0.121   3.252  13.288
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   89.0669     1.0706  83.197  < 2e-16 ***
## x_scaledA                      1.8080     0.5875   3.077  0.00245 **
## factor(eligibleA)1            20.8270     1.5445  13.484  < 2e-16 ***
## x_scaledA:factor(eligibleA)1   0.6934     0.8680   0.799  0.42559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.999 on 161 degrees of freedom
## Multiple R-squared:  0.6885, Adjusted R-squared:  0.6827
## F-statistic: 118.6 on 3 and 161 DF,  p-value: < 2.2e-16
```

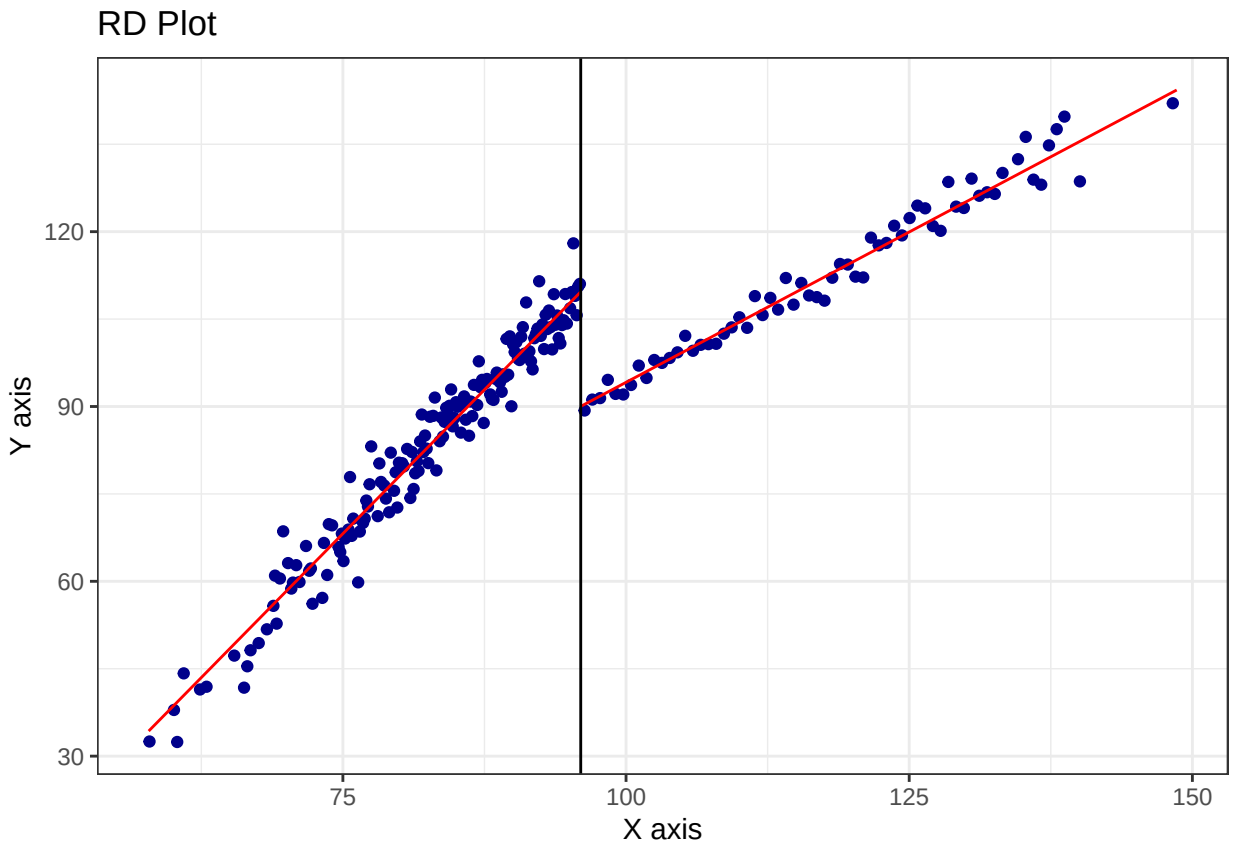
```
# Estimated causal effect using linear regression with interaction
# (the correct model for this data) at the threshold is:
summary(fitA)$coef[3, 1]
```

```
## [1] 20.82704
```

```
estimates_and_se <- data.frame(
  world_and_method = "world A: linear regression",
  estimate = summary(fitA)$coef[3, 1],
  standard_error = summary(fitA)$coef[3, 2]
)
```

Estimating a causal effect using rdrobust with a local linear specification and unequal bandwidth on the left and right sides of the cutoff.

```
rdd_implementationA <- rdrobust(observed_outcomeA, xA,
  c = discontinuity_pointA, bwselect = "msetwo", p = 1,
  all = T
)
rdplot(observed_outcomeA, xA, c = discontinuity_pointA, p = 1)
```




```
# unequal bandwidth
rdd_bandwidthsA <- rdrobust(observed_outcomeA, xA,
  c = discontinuity_pointA, bwselect = "msetwo", p = 1
)$bws
rdd_bandwidthsA[1, 1] == rdd_bandwidthsA[1, 2]
```

```
## [1] FALSE
```

```
# Estimated causal effect using rdrobust with a local linear specification
# (the correct model for this data) at the threshold is:
-rdd_implementationA$coef[3]
```

```
## [1] 20.02973
```

```
estimates_and_se <- estimates_and_se %>% add_row(
  world_and_method = "world A: rdrobust",
  estimate = -rdd_implementationA$coef[3],
  standard_error = rdd_implementationA$se[3, 1]
)
```

World B

Estimating a causal effect for the effect at the threshold using a linear regression that includes an interaction term between the running variable and the treatment group.

Choosing the bandwidth arbitrarily in this case, we have:

```r
bandwidthB <- c(93, 99)
worldB_restricted <- worldB %>% filter(xB > bandwidthB[1] & xB < bandwidthB[2])
fitB <- lm(observed_outcomeB ~ x_scaledB * factor(eligibleB), data = worldB_restricted)
summary(fitB)
```

```
##
## Call:
## lm(formula = observed_outcomeB ~ x_scaledB * factor(eligibleB),
##     data = worldB_restricted)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0678  -3.3581   0.0243   2.6900  11.9617
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    83.3169     0.9570  87.059   <2e-16 ***
## x_scaledB                       0.7765     0.6063   1.281    0.202
## factor(eligibleB)1             25.2824     1.3941  18.136   <2e-16 ***
## x_scaledB:factor(eligibleB)1    0.1149     0.8372   0.137    0.891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.609 on 163 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8612
## F-statistic: 344.4 on 3 and 163 DF,  p-value: < 2.2e-16
```

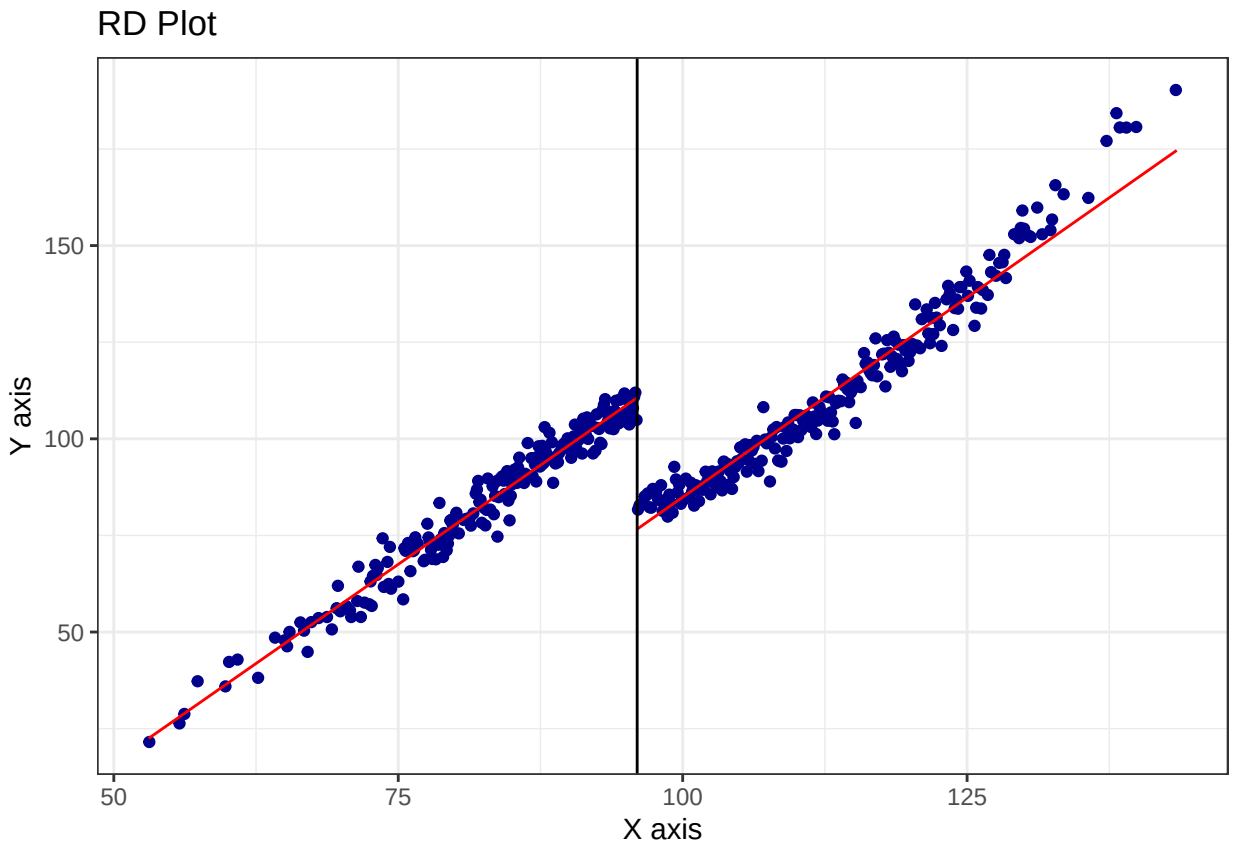
```r
# Estimated causal effect using linear regression with interaction
# (the incorrect model for this data) at the threshold is
summary(fitB)$coef[3, 1]
```

```
## [1] 25.28236
```

```r
estimates_and_se <- estimates_and_se %>% add_row(
  world_and_method = "world B: linear regression",
  estimate = summary(fitB)$coef[3, 1],
  standard_error = summary(fitB)$coef[3, 2]
)
```

Estimating a causal effect using rdrobust with a local linear specification and unequal bandwidth on the left and right sides of the cutoff.

```r
rdd_implementationB <- rdrobust(observed_outcomeB, xB,
  c = discontinuity_pointB, bwselect = "msetwo", p = 1
)
rdplot(observed_outcomeB, xB, c = discontinuity_pointB, p = 1)
```




```
# unequal bandwidth
rdd_bandwidthsB <- rdrobust(observed_outcomeB, xB,
  c = discontinuity_pointB, bwselect = "msetwo", p = 1
)$bws
rdd_bandwidthsB[1, 1] == rdd_bandwidthsB[1, 2]
```

```
## [1] FALSE
```

```
# Estimated causal effect using rdrobust with a local linear specification
# (the incorrect model for this data) at the threshold is:
-rdd_implementationB$coef[3]
```

```
## [1] 25.85813
```

```
estimates_and_se <- estimates_and_se %>% add_row(
  world_and_method = "world B: rdrobust",
  estimate = -rdd_implementationB$coef[3],
  standard_error = rdd_implementationB$se[3, 1]
)
```

World C

Estimating a causal effect for the effect at the threshold using a linear regression that includes an interaction term between the running variable and the treatment group.

Choosing the bandwidth arbitrarily in this case, we have:

```r
bandwidthC <- c(93, 99)
worldC_restricted <- worldC %>% filter(xC > bandwidthC[1] & xC < bandwidthC[2])
fitC <- lm(observed_outcomeC ~ x_scaledC * factor(eligibleC), data = worldC_restricted)
summary(fitC)
```

```
##
## Call:
## lm(formula = observed_outcomeC ~ x_scaledC * factor(eligibleC),
##     data = worldC_restricted)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9821  -3.4860   0.1462   3.4388  14.6600
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    90.6107     1.4493  62.520   <2e-16 ***
## x_scaledC                       0.5686     0.9159   0.621    0.535
## factor(eligibleC)1             19.1075     1.5405  12.404   <2e-16 ***
## x_scaledC:factor(eligibleC)1    1.0843     1.0165   1.067    0.287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.421 on 215 degrees of freedom
## Multiple R-squared:  0.6602, Adjusted R-squared:  0.6555
## F-statistic: 139.3 on 3 and 215 DF,  p-value: < 2.2e-16
```

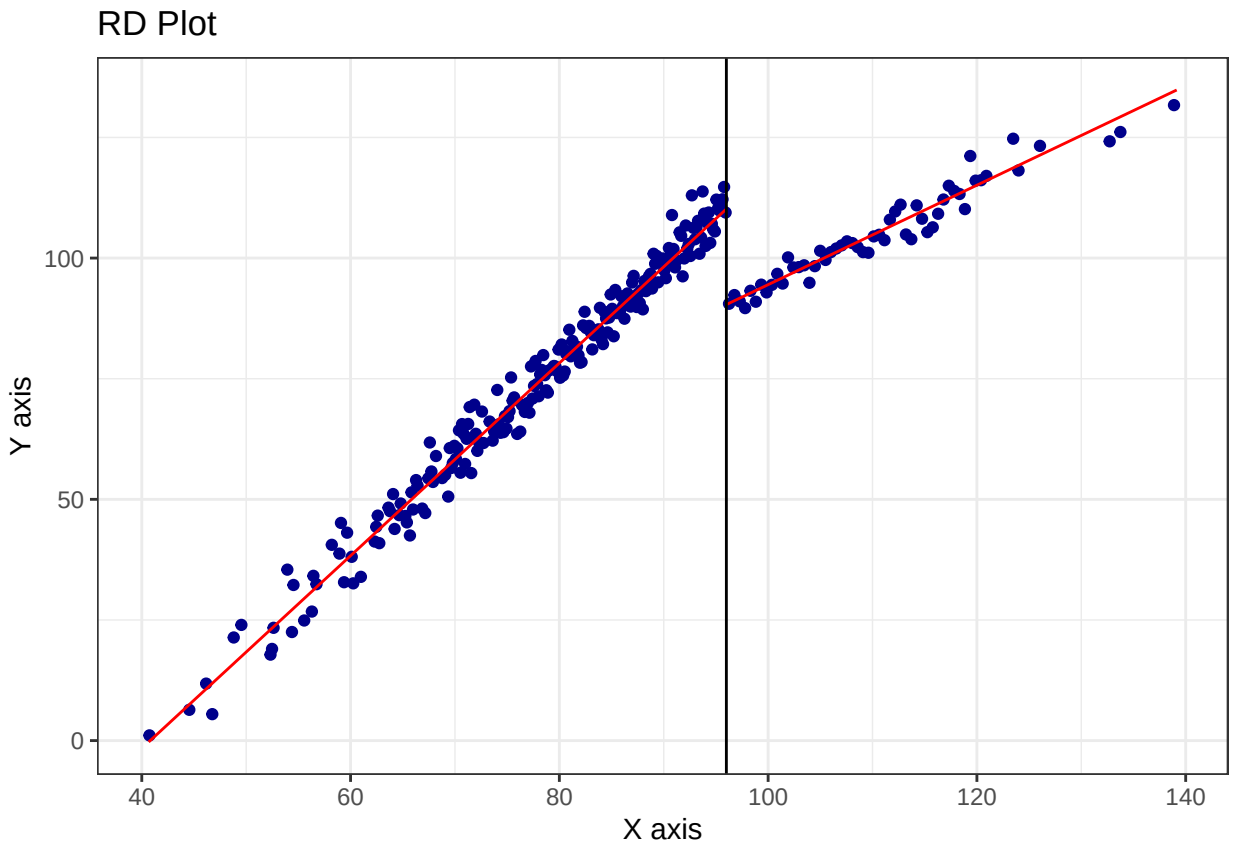
```r
# Estimated causal effect using linear regression with interaction
# (the incorrect model for this data) at the threshold is:
summary(fitC)$coef[3, 1]
```

```
## [1] 19.10753
```

```r
estimates_and_se <- estimates_and_se %>% add_row(
  world_and_method = "world C: linear regression",
  estimate = summary(fitC)$coef[3, 1],
  standard_error = summary(fitC)$coef[3, 2]
)
```

Estimating a causal effect using rdrobust with a local linear specification and unequal bandwidth on the left and right sides of the cutoff.

```r
rdd_implementationC <- rdrobust(observed_outcomeC, xC,
  c = discontinuity_pointC, bwselect = "msetwo", p = 1
)
rdplot(observed_outcomeC, xC, c = discontinuity_pointC, p = 1)
```




```
# unequal bandwidth
rdd_bandwidthsC <- rdrobust(observed_outcomeC, xC,
  c = discontinuity_pointC, bwselect = "msetwo", p = 1
)$bws
rdd_bandwidthsC[1, 1] == rdd_bandwidthsC[1, 2]
```

```
## [1] FALSE
```

```
# Estimated causal effect using rdrobust with a local linear specification
# (the incorrect model) at the threshold is:
-rdd_implementationC$coef[3]
```

```
## [1] 20.14306
```

```
estimates_and_se <- estimates_and_se %>% add_row(
  world_and_method = "world C: rdrobust",
  estimate = -rdd_implementationC$coef[3],
  standard_error = rdd_implementationC$se[3, 1]
)
```

Final table of causal estimates and standard errors:

```
estimates_and_se <-
  estimates_and_se %>%
  rename(
```

```
    "World and Method" = "world_and_method",
    "Causal Estimate" = "estimate",
    "Standard Error" = "standard_error"
  ) %>%
  gt() %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels(
      columns = c(
        "World and Method",
        "Causal Estimate",
        "Standard Error"
      )
    )
  )

estimates_and_se
```

World and Method	Causal Estimate	Standard Error
world A: linear regression	20.82704	1.544535
world A: rdrobust	20.02973	1.261084
world B: linear regression	25.28236	1.394050
world B: rdrobust	25.85813	1.158953
world C: linear regression	19.10753	1.540482
world C: rdrobust	20.14306	1.212877

Causal interpretation of the linear regression estimate in world A: For people who earned 96 dollars per week, getting free child care resulted in economic satisfaction levels 20.827 units higher on average than they would have been if they hadn't got free child care.

Non-causal interpretation of the linear regression estimate in world B: If we compared two groups of people, both of which have an income of 96 dollars per week, where one group received child care and one did not, then the average economic satisfaction of those who received childcare would be 25.28236 units, on average, more, than that for those who did not receive childcare.

7. Bias

```
true_value <- 20
iter <- 10000
```

World A: Linear Regression

```
estimate_linear_regression_worldA <- rep(NA, iter)

for (i in 1:iter) {
  n <- 1000
  xA <- rnorm(n, mean = 100, sd = 15)
```

```
  x_scaledA <- xA - 96
  discontinuity_pointA <- 96
  eligibleA <- ifelse(xA <= discontinuity_pointA, 1, 0)
  potential_outcomeA_y0 <- x_scaledA + rnorm(n, mean = 90, sd = 5)
  potential_outcomeA_y1 <- x_scaledA + rnorm(n, mean = 110, sd = 5) +
    x_scaledA * eligibleA
  observed_outcomeA <- ifelse(eligibleA == 1, potential_outcomeA_y1, potential_outcomeA_y0)
  worldA <- data.frame(xA, x_scaledA, eligibleA, potential_outcomeA_y0, potential_outcomeA_y1, observed
  worldA_restricted <- worldA %>% filter(xA > bandwidthA[1] & xA < bandwidthA[2])
  fitA <- lm(observed_outcomeA ~ x_scaledA * factor(eligibleA), data = worldA_restricted)
  estimate_linear_regression_worldA[i] <- summary(fitA)$coef[3, 1]
}

bias_linear_regression_worldA <- true_value - abs(mean(estimate_linear_regression_worldA))
```

World A: Rdrobust

```
estimate_rdrobust_worldA <- rep(NA, iter)

for (i in 1:iter) {
  n <- 1000
  xA <- rnorm(n, mean = 100, sd = 15)
  x_scaledA <- xA - 96
  discontinuity_pointA <- 96
  eligibleA <- ifelse(xA <= discontinuity_pointA, 1, 0)
  potential_outcomeA_y0 <- x_scaledA + rnorm(n, mean = 90, sd = 5)
  potential_outcomeA_y1 <- x_scaledA + rnorm(n, mean = 110, sd = 5) +
    x_scaledA * eligibleA
  observed_outcomeA <- ifelse(eligibleA == 1, potential_outcomeA_y1, potential_outcomeA_y0)
  rdd_implementationA <- rdrobust(observed_outcomeA, xA,
    c = discontinuity_pointA, bwselect = "msetwo", p = 1,
    all = T
  )
  estimate_rdrobust_worldA[i] <- -rdd_implementationA$coef[3]
}

bias_rdrobust_worldA <- true_value - abs(mean(estimate_rdrobust_worldA))
```

World B: Linear Regression

```
estimate_linear_regression_worldB <- rep(NA, iter)

for (i in 1:iter) {
  n <- 1000
  xB <- rnorm(n, mean = 100, sd = 15)
  x_scaledB <- xB - 96
  discontinuity_pointB <- 96
  eligibleB <- ifelse(xB <= discontinuity_pointB, 1, 0)
  potential_outcomeB_y0 <- x_scaledB + rnorm(n, mean = 90, sd = 5) +
```

```
    10 * scale(x_scaledB^2)
  potential_outcomeB_y1 <- x_scaledB + rnorm(n, mean = 110, sd = 5) +
    x_scaledB * eligibleB
  observed_outcomeB <- ifelse(eligibleB == 1, potential_outcomeB_y1, potential_outcomeB_y0)
  worldB <- data.frame(xB, x_scaledB, eligibleB, potential_outcomeB_y0, potential_outcomeB_y1, observed
  worldB_restricted <- worldB %>% filter(xB > bandwidthB[1] & xB < bandwidthB[2])
  fitB <- lm(observed_outcomeB ~ x_scaledB * factor(eligibleB), data = worldB_restricted)
  estimate_linear_regression_worldB[i] <- summary(fitB)$coef[3, 1]
}

bias_linear_regression_worldB <- true_value - abs(mean(estimate_linear_regression_worldB))
```

World B: Rdrobust

```
estimate_rdrobust_worldB <- rep(NA, iter)

for (i in 1:iter) {
  n <- 1000
  xB <- rnorm(n, mean = 100, sd = 15)
  x_scaledB <- xB - 96
  discontinuity_pointB <- 96
  eligibleB <- ifelse(xB <= discontinuity_pointB, 1, 0)
  potential_outcomeB_y0 <- x_scaledB + rnorm(n, mean = 90, sd = 5) +
    10 * scale(x_scaledB^2)
  potential_outcomeB_y1 <- x_scaledB + rnorm(n, mean = 110, sd = 5) +
    x_scaledB * eligibleB
  observed_outcomeB <- ifelse(eligibleB == 1, potential_outcomeB_y1, potential_outcomeB_y0)
  rdd_implementationB <- rdrobust(observed_outcomeB, xB,
    c = discontinuity_pointB, bwselect = "msetwo", p = 1,
    all = T
  )
  estimate_rdrobust_worldB[i] <- -rdd_implementationB$coef[3]
}

bias_rdrobust_worldB <- true_value - abs(mean(estimate_rdrobust_worldB))
```

World C: Linear Regression

```
estimate_linear_regression_worldC <- rep(NA, iter)

for (i in 1:iter) {
  x1 <- rnorm(0.9 * n, mean = 100, sd = 15)
  x2 <- runif(0.1 * n, 95.9, 96)
  xC <- c(x1, x2)
  x_scaledC <- xC - 96
  discontinuity_pointC <- 96
  eligibleC <- ifelse(xC <= discontinuity_pointC, 1, 0)
  potential_outcomeC_y0 <- x_scaledC + rnorm(n, mean = 90, sd = 5)
  potential_outcomeC_y1 <- x_scaledC + rnorm(n, mean = 110, sd = 5) + x_scaledC * eligibleC
```

```
  observed_outcomeC <- ifelse(eligibleC == 1, potential_outcomeC_y1, potential_outcomeC_y0)
  worldC <- data.frame(xC, x_scaledC, eligibleC, potential_outcomeC_y0, potential_outcomeC_y1, observed
  worldC_restricted <- worldC %>% filter(xC > bandwidthC[1] & xC < bandwidthC[2])
  fitC <- lm(observed_outcomeC ~ x_scaledC * factor(eligibleC), data = worldC_restricted)
  estimate_linear_regression_worldC[i] <- summary(fitC)$coef[3, 1]
}

bias_linear_regression_worldC <- true_value - abs(mean(estimate_linear_regression_worldC))
```

World C: Rdrobust

```
estimate_rdrobust_worldC <- rep(NA, iter)

for (i in 1:iter) {
  x1 <- rnorm(0.9 * n, mean = 100, sd = 15)
  x2 <- runif(0.1 * n, 95.9, 96)
  xC <- c(x1, x2)
  x_scaledC <- xC - 96
  discontinuity_pointC <- 96
  eligibleC <- ifelse(xC <= discontinuity_pointC, 1, 0)
  potential_outcomeC_y0 <- x_scaledC + rnorm(n, mean = 90, sd = 5)
  potential_outcomeC_y1 <- x_scaledC + rnorm(n, mean = 110, sd = 5) + x_scaledC * eligibleC
  observed_outcomeC <- ifelse(eligibleC == 1, potential_outcomeC_y1, potential_outcomeC_y0)
  rdd_implementationC <- rdrobust(observed_outcomeC, xC,
    c = discontinuity_pointC, bwselect = "msetwo", p = 1,
    all = T
  )
  estimate_rdrobust_worldC[i] <- -rdd_implementationC$coef[3]
}

bias_rdrobust_worldC <- true_value - abs(mean(estimate_rdrobust_worldC))
```

```
results <- data.frame(
  World = rep(c("A", "B", "C"), each = 2),
  Estimator = rep(c("Linear Regression", "Rdrobust"), 3),
  Bias = c(
    abs(bias_linear_regression_worldA), abs(bias_rdrobust_worldA),
    abs(bias_linear_regression_worldB), abs(bias_rdrobust_worldB),
    abs(bias_linear_regression_worldC), abs(bias_rdrobust_worldC)
  )
)
results <- results %>%
  gt() %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels(
      columns = c(
        "World",
        "Estimator",
        "Bias"
      )
```

```
    )
  )

results
```

World	Estimator	Bias
A	Linear Regression	0.033608980
A	Rdrobust	0.002397606
B	Linear Regression	7.145005648
B	Rdrobust	7.121262403
C	Linear Regression	0.017979950
C	Rdrobust	0.002259055

We note that in world C above, the pile-up before the cutoff of 96 is only problematic if it suggests cheating or manipulation. Thus, in this world, the income one observes is not what one would see in the case of no manipulation that the true DGP should be based upon. In the simulations above, even though world C provided us with a valid causal estimate, it may not always be the case. Consider an extreme example below:

```
set.seed(123)
n <- 1000
x1 <- rnorm(0.01 * n, mean = 100, sd = 15)
x2 <- runif(0.99 * n, 95.9, 96.0)
xC <- c(x1, x2)
x_scaledC <- xC - 96
discontinuity_pointC <- 96
eligibleC <- ifelse(xC <= discontinuity_pointC, 1, 0)
potential_outcomeC_y0 <- x_scaledC + rnorm(n, mean = 90, sd = 5)
potential_outcomeC_y1 <- x_scaledC + rnorm(n, mean = 110, sd = 5) + x_scaledC*eligibleC
observed_outcomeC <- ifelse(eligibleC == 1, potential_outcomeC_y1, potential_outcomeC_y0)
worldC <- data.frame(xC, x_scaledC, eligibleC, potential_outcomeC_y0,
                     potential_outcomeC_y1, observed_outcomeC)
bandwidthC <- c(93, 99)
worldC_restricted <- worldC %>% filter(xC > bandwidthC[1] & xC < bandwidthC[2])
fitC <- lm(observed_outcomeC ~ x_scaledC + factor(eligibleC), data = worldC_restricted)
# the causal estimate for world C (an extreme assumption violation example)
# linear regression
summary(fitC)$coef[3, 1:2]
```

```
##   Estimate Std. Error
##  25.443925   5.031354
```

As we can see above, the extreme violation of the structural assumption in world C has led to large standard errors as there are only very few observations after the cutoff than before; in addition to this, the estimate of the causal ATE at the cutoff is also off by around 5 units.

- In world A, I did not violate any assumptions; thus, the estimates from linear regression and rdrobust were very close to the true causal effect at the cutoff. Thus, bias was small.
- In world B, I violated the assumption of linearity between the outcome and covariates - this led to inaccurate estimates as given by linear regression (which assumes linearity - thus it gave us an inaccurate result) and rdrobust (which we specified that it estimate the causal effect using local linear specification - thus it gave us an inaccurate result). Thus, bias was much larger in these cases.

- In world C, I violated the structural assumption of treatment assignment being ignorable at the cutoff. This surprisingly did not lead to inaccurate results (bias was small) using linear regression or rdrobust as the implication of violating this assumption is that the running variable that one observes is not what one would see in the case of no manipulation that the true DGP should be based upon. The result of getting accurate results using linear regression was especially surprising to me as having a large number of data points on one side of the threshold as compared to the other within a certain bandwidth led me to believe that linear regression would fail but it did not. Maybe if the bandwidth was increased, linear regression would lead to inaccurate results. Additionally, in the extreme case of having 99% of our observations just before the cutoff of 96, we see that linear regression estimates are especially off at 25.443925 with a large standard error of 5.031354. Thus, substantially violating this assumption can lead to inaccurate causal estimates while violating this assumption in a more believable manner can reflect negatively on the design of the study because one is making causal claims without accounting for manipulations in the data. This can also have implications when it comes to the generalizability of the study's results.