We ultimately want to estimate the effect of D_i on Y. Suppose that D_i is determined, in part, by whether $X_i \geq c$. We call X_i the *running* or *forcing* variable. We also assume that $Y_i(0)$ and $Y_i(1)$ are related to X_i *continuously* to preclude a large and discontinuous jump in Y_i as X_i changes. Suppose further that the probability of treatment ($D_i = 1$) changes, based on how X_i is related to a threshold c. If we see that Y jumps disconinously, then we can estimate the effect of treatment around the threshold value.

As an example, researchers have used test score cutoffs as the threshold value in a regression discontinuity design. If a person failed a test, they were sent to summer school. The outcomes of students around the cutoff served to identify the efficacy of summer school.

We will never observe multiple treatments for the same individual, and are therefore unable to calculate $Y_i(1) - Y_i(0)$ directly. Instead, we can examine similar observations around the cutoff to estimate the expected impact. First, assume that $Y_i(1)$ and $Y_i(0)$ are smooth functions of X_i. Formally, $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ are continuous in X. Then

$$\tau_{srd} = \lim_{X \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{X \uparrow c} \mathbb{E}[Y_i|X_i = x] \tag{1}$$

We will rarely observe an individual with $X_i = c$ but in that case, assume that the treatment is granted. For a given individual, we will be unable to observe either $\mathbb{E}[Y_i(0)|X_i = x]$ or $\mathbb{E}[Y_i(1)|X_i = x]$ in order to estimate the true treatment effect, defined by

$$\tau_{srd} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c] = \mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c]$$

and so we must rely on the estimate in Equation (1). We are trying to estimate Equation 1 empirically by only looking at those individuals "near" the threshold c. We can look at this in code with the generating process

$$Y_i = 1 + X_i + 2 \cdot D_i + \epsilon \quad \text{with} \quad \epsilon \sim N(0, 1/2)$$

where $i \in \{1, 2, \ldots, N\}$. In particular, we set $\tau_{srd} = 2$. Can we estimate this treatment effect using regression discontinuity? First consider the sharp RD design — everyone with $X_i > c$ gets treated and everyone else does not.

```
c <- 0.5; N <- 10000
X <- runif(N)
D <- ifelse(X > c, 1, 0)
Y <- 1 + X + 2*D + rnorm(N, sd = 0.5)
```

Now define a bandwidth b, where we restrict our attention to observations with $X_i \in (c - b, c + b)$. The following code collects the indices for these individuals.

```
b <- 0.05
lower <- X < c & X > c - b
upper <- X > c & X < c + b
```

The total number of individuals in this group should be about 10% of the total sample, or about 1,000 when $N = 10,000$, given that the X_i's are drawn from a uniform distribution. This is shown to be true:

```
length(c(which(upper), which(lower)))
```

```
[1] 973
```

It follows from Equation (1) that we can simply difference the outcome variables for the upper and lower groups. Indeed the outcome reflects this shift – which can also be plotted.

```
mean(Y[upper]) - mean(Y[lower])
```

```
[1] 2.072374
```

Now consider the fuzzy RD design, where D_i is no longer determined *only* by X_i. Mathematically,

$$0 < \lim_{X \downarrow c} \mathbb{P}[D_i = 1 | X_i = x] - \lim_{X \uparrow c} \mathbb{P}[D_i = 1 | X_i = x] < 1 \tag{2}$$

The estimate for the treatment effect in a fuzzy regression discontinuity design is therefore given as

$$\tau_{frd} = \frac{\lim_{X \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{X \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{X \downarrow c} \mathbb{E}[D_i | X_i = x] - \lim_{X \uparrow c} \mathbb{E}[D_i | X_i = x]} \tag{3}$$

This is equivalent to the the instrumental variables estimator with instrument $Z_i = \mathbb{1}(X_i \geq c)$. There must also be a monotonicity assumption about the way that the treatment changes with X_i. If you increase the threshold, then there is a higher hurdle to get treated, so the probability of treatment should be non-increasing. We have already assumed that the probability of treatment is increasing in the running variable X. We need to assume that there is not strange behavior around the threshold.

Consider the following data generating process, where the treatment is no longer a deterministic function of Y_i and that there is some additional randomness that determines treatment:

```
g <- 0.5; gamma <- ifelse(X > c, 1, 0) + rnorm(N)
D <- ifelse(gamma > g, 1, 0)
Y <- 1 + X + 2*D + rnorm(N, sd = 0.5)
```

The probability of treatment is increasing in X, specifically, $D_i = \mathbb{1}(\gamma_i > g)$, where $gamma_i = \mathbb{1}(X_i > c) + \epsilon_i$, with $\epsilon_i \sim N(0, 1)$ and where g is a predetermined threshold distinct from c. We can estimate the treatment effect in Equation (3) by scaling the sharp regression discontinuity with the difference in probability on either side of the c threshold.

```
(mean(Y[upper]) - mean(Y[lower])) / (mean(D[upper]) - mean(D[lower]))
```

```
[1] 2.174791
```

Note that, assuming that (3) yields an unbiased estimator, applying the sharp RD in (1) will underestimate the treatment impact, given that

$$\lim_{X \downarrow c} \mathbb{E}[D_i | X_i = x] - \lim_{X \uparrow c} \mathbb{E}[D_i | X_i = x] < 1$$

How is the estimator from the fuzzy regression discontinuity design distributed? The following code runs the exact specification 1,000 times and plots the histogram in Figure 1.

```
sim.fn <- function(repitition, c = 0.5, b = 0.05, g = 0.5, N = 10000) {
  X <- runif(N)
  gamma <- ifelse(X > c, 1, 0) + rnorm(N)
  D <- ifelse(gamma > g, 1, 0)
  Y <- 1 + X + 2*D + rnorm(N, sd = 0.5)
  lower <- X < c & X > c - b
  upper <- X > c & X < c + b
  (mean(Y[upper]) - mean(Y[lower])) / (mean(D[upper]) - mean(D[lower]))
}
```

The resulting histogram suggests that the fuzzy RD design is biased upwards. Additionally, the bias is a function of almost all of the parameters in `sim.fn`, including c, g, b, and the standard error on the disturbance terms. Why would this be? It seems that the bias is a result of attenuation, given two strict thresholds and two continuous distributions.

```
x <- sapply(1:1000, sim.fn)
hist(x, xlab = "", border = "grey", col = "grey", breaks = 40, main = "")
```

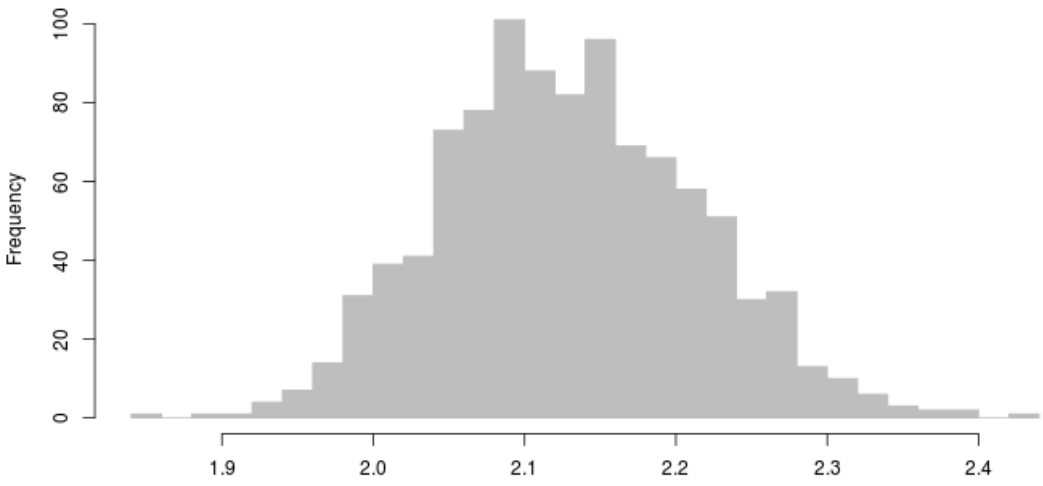

Figure 1: Histogram of estimated impact, simulated fuzzy regression discontinuity