PHS 2000B: Regression Discontinuity

Regression discontinuity design intuition

The regression discontinuity design (RDD) is used to estimate causal effects in settings where treatment eligibility depends on somewhat arbitrary cut-offs (or "thresholds"). The intuition that underlies RDD is that if the threshold for treatment assignment is arbitrary, then individuals within a narrow bandwidth on either side of the threshold are comparable.

When can we do a regression discontinuity analysis?

RDDs can only be applied in settings where we have information on a running variable (a.k.a., forcing variable) and a threshold/cut-off which is defined based on the running variable.

	Definition	Examples
Running variable	A continuous variable that is used to determine decisions about treatment. This variable may be associated with the outcome variable in the absence of treatment.	CD4 count; Test scores; GPA; Birth weight etc.
Threshold	A value of the running variable at which the probability of treatment changes discontinuously. These cut-offs are typically created by policies or administrative rules.	CD4 count below 350 to be eligible for ART; GPA above a certain value to be eligible for scholarships

Why are regression discontinuity analyses so appealing?

There are a couple of reasons for this. First, RDDs are one of the few study designs where an investigator knows the treatment assignment mechanism. Knowing the assignment mechanism (or "rule") allows us to develop models which more accurately capture the "selection mechanism", i.e., the mechanism that leads to certain people being treated while others not. In the real world, lots of treatment assignments are mechanistic and adhere to rules. Many of these rules are arbitrary (e.g., folks above a certain threshold get a benefit, while those below do not), allowing RDD studies to be applied to numerous real-world policy questions.

Second, studies have shown that RDD estimates have very high internal validity. For example, Chaplin et. al. (2018) test estimates from 15 RDD studies against comparable randomized controlled trials and find that the RDD studies very closely approximate the results from the RCTs.

Types of regression discontinuity designs

Broadly, there are two types of RDDs: sharp RDD and fuzzy RDD. The key distinction between these two designs is the way in which the **probability of treatment** changes at the cut-off.

Sharp regression discontinuity design

In a sharp RDD, the probability of receiving treatment jumps from 0 to 1 at the cut-off. In other words, treatment receipt is a deterministic (and discontinuous) function of the running variable. Figure 1 illustrates a sharp RDD. X represents the running variable, A a binary treatment, and the dashed vertical line at $X = c$ represents the threshold. The y-axis represents the probability of receiving the treatment A. Note that the probability jumps from 0 (left of the threshold) to 1 (right of the threshold) at the threshold itself.

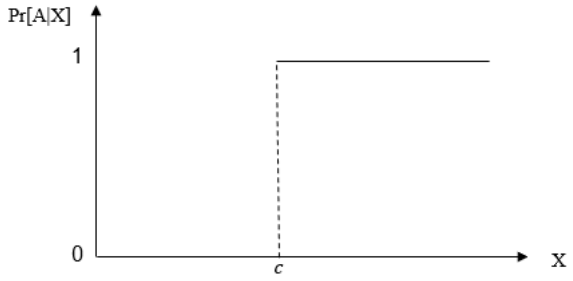


Figure 1: Sharp RDD

Fuzzy regression discontinuity design

In a fuzzy RDD, the probability of receiving treatment changes at the threshold, but *not* from 0 to 1. Thus, in a fuzzy RDD, the probability of receiving treatment is not a deterministic function of the running variable. Figure 2 illustrates a fuzzy RDD where A represents a binary treatment, X the running variable, and $X = c$ the threshold on the running variable.

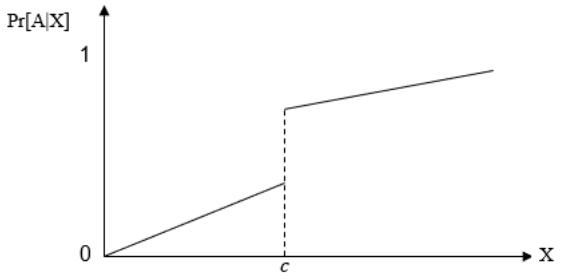


Figure 2: Fuzzy RDD

Challenge of causal inference in regression discontinuity designs

Sharp regression discontinuity design

Suppose we have a binary treatment ($A = \{0, 1\}$) where $A = 1$ implies treatment receipt and $A = 0$ implies treatment non-receipt. Suppose further that $A = 1$ with probability 100% when $X \geq c$ where c represents some arbitrary value of the running variable. Since we are in a sharp RDD setting, $A = 0$ when $X < c$.

The fundamental challenge of causal inference in a sharp RDD comes from the fact that we can only ever observe potential outcomes under treatment when $X \geq c$ and potential outcomes under no treatment when $X < c$. In other words, sharp RDDs suffer from a violation of structural positivity. This critical feature of the sharp RDD sets it apart from other regression-based methods of causal inference where we are able to contrast treated and control individuals within levels of covariates.

Figure 3, which is adapted from Bor et. al. (2014), illustrates the fundamental challenge of causal inference in a sharp RDD. Solid lines represent observed values while dotted lines represent unobserved values. In addition, blue represents potential outcomes under treatment ($Y^{a=1}$) while red represents potential outcomes under no treatment ($Y^{a=0}$). Note that the conditional expectation of the outcome under treatment is only observable when $X \geq c$ while the conditional expectation of the outcome under no treatment is observable only when $X < c$.

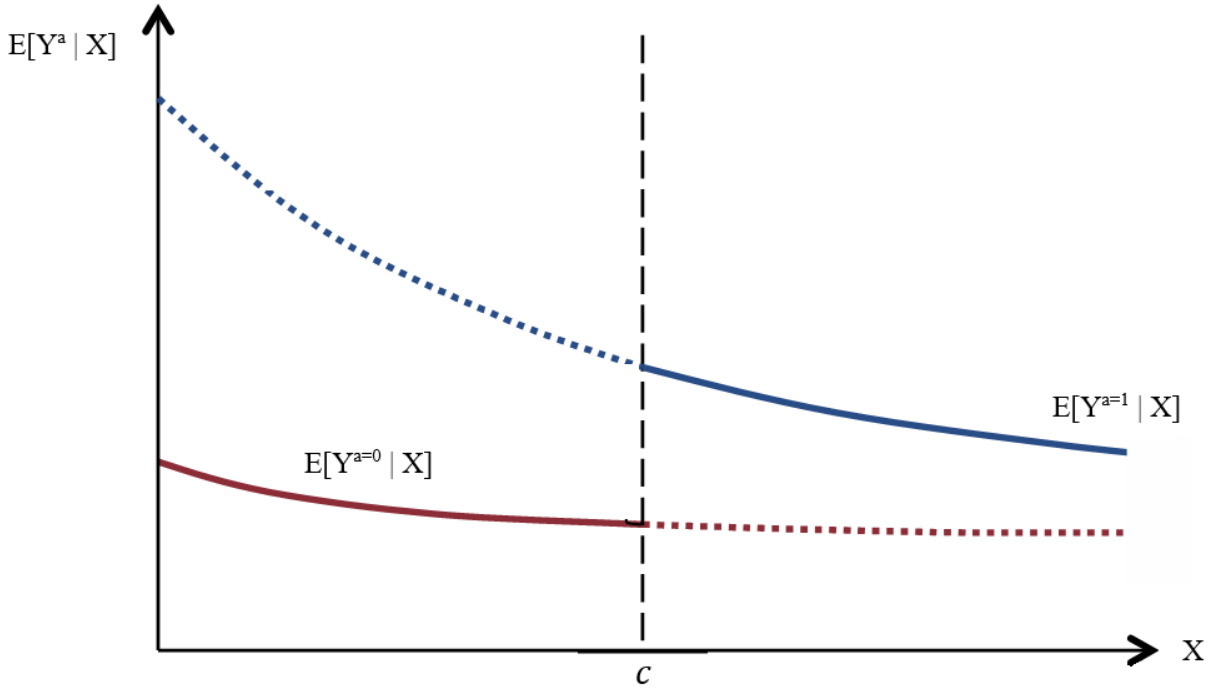


Figure 3: Fundamental challenge of causal inference in a sharp RDD

Fuzzy regression discontinuity design

In a fuzzy RDD, we observe both treated and untreated individuals on either side of the threshold since treatment receipt is not a deterministic function of the running variable. Thus, fuzzy RDDs do not face the same fundamental challenge of causal inference as sharp RDDs. However, the fact that there is variation in treatment status on either side of the cut-off suggests that there might be some sort of selection on unobservables issue which we need to account for when conducting a regression discontinuity analysis.

Causal identification in regression discontinuity designs

Causal effect of interest

Sharp regression discontinuity design

In a sharp RDD, we are interested in the average treatment effect at the threshold. This is also known as the Average Causal Effect (ACE) in the RDD literature. Mathematically, this is expressed as:

$$\tau_{SRD} \equiv E[Y_i^{a=1} - Y_i^{a=0} | X = c] \tag{1}$$

where τ_{SRD} refers to the treatment effect in a sharp RDD. Note that this treatment effect is *only* at the threshold and may not be more generally applicable.

Using our data, the ACE is identified as:

$$\tau_{SRD} = lim_{x \downarrow c} E[Y_i|X_i = x] - lim_{x \uparrow c} E[Y_i|X_i = x] \tag{2}$$

Fuzzy regression discontinuity design

In a fuzzy RDD, we identify the causal effect among compliers at the threshold, i.e., the Local Average Treatment Effect at the threshold. This effect is often called the Complier Average Causal Effect (CACE) in the RDD literature. Mathematically, we can write this as:

$$\tau_{FRD} \equiv E[Y_i^{a=1} - Y_i^{a=0}|X = c, compliers] \tag{3}$$

where τ_{FRD} refers to the treatment effect in a fuzzy RDD.

Using our data, the CACE is identified as:

$$\tau_{FRD} = \frac{lim_{x \downarrow c} E[Y_i|X_i = x] - lim_{x \uparrow c} E[Y_i|X_i = x]}{lim_{x \downarrow c} E[A_i|X_i = x] - lim_{x \uparrow c} E[A_i|X_i = x]} \tag{4}$$

Exercise 1: True, false, uncertain

1. You are trying to identify the causal effect of a new antimalarial drug on child survival. You find that this drug is recommended for children who have malaria and a body temperature above 104F. You examine the data and find that utilization increases from 20 percent for children with body temperature <104F to 40 percent for children with body temperature ???104F. You make no assumption regarding the distribution of compliers at the threshold of 104F. A RDD analysis of these data will allow us to identify the Average Causal Effect (ACE) as well as the Average Treatment Effect (ATE).

   False. This is a fuzzy RDD so what we can identify using the data is the Complier Average Causal Effect (CACE). Under the practically implausible but theoretically plausible assumption that the compliers are a random sample of individuals at the cut-off, then we might argue that the CACE equals the ACE, but this is a very special case.

2. In a sharp RDD, the Average Causal Effect (ACE) will equal the Complier Average Causal Effect (CACE).

   True. In a sharp RDD, the ACE = CACE.

3. In a sharp RDD, the Average Causal Effect (ACE) will equal the Average Treatment Effect (ATE).

   False. While some papers use the terms ACE and ATE interchangeably, we use ACE to specifically refer to the causal effect at the threshold in a RDD study. Thus, the ACE will only equal the ATE if the treatment effect is assumed to be homogenous along the running variable (i.e., the causal effect at the threshold is equal to the causal effect at any point of the running variable.

Identification assumptions

Sharp RDD

We estimate the causal effect at the cut-off in a sharp RDD by extrapolating outcomes among treated and control individuals to the threshold. This extrapolation relies on only one identifying assumption:

1. **Expected potential outcomes are continuous in the running variable at the threshold:** Figure 4 below shows what it means for expected potential outcomes to be continuous at the threshold (and, in this case, along the running variable as well). Note that the expected potential outcome under either treatment or no treatment is a smooth function of X at and around the cut-off. This assumption is important because it allows us to extrapolate outcomes on either side of the threshold and identify the causal effect under a sharp RDD (more on this below).

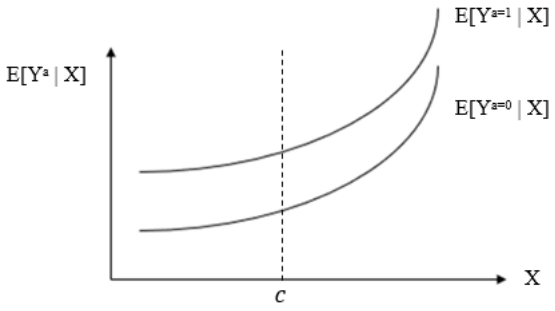


Figure 4: Continuous potential outcomes

Notice that the continuous potential outcomes imply that individuals directly above and below the threshold are exchangeable, and that the threshold does not coincide or cause changes in variables that are associated with the outcome (which implies exclusion holds as well.) Further, the discontinuity at the threshold implies relevance: the threshold is associated with a discontinuity in treatment. These are quite similar to our IV assumptions; indeed we could think of the sharp RD design as an extension of IV, but in this case we have perfect adherence at the threshold, so we do not need to adjust for non-compliance.

Fuzzy RDD

Fuzzy RDDs don't suffer from the issue of structural positivity. Rather, the challenge of causal inference in a fuzzy RDD is that the treated/untreated individuals we observe on either side of the cut-off within levels of the running variable may not be comparable, particularly in terms of unobserved variables that affect the outcome.

Remember that one way of identifying the causal effect in a scenario where we have selection on unobservables is to use instrumental variables. In trying to identify the causal effect in a fuzzy RDD, we use the logic of instrumental variables as a guide (indeed, Angrist and Pischke argue that fuzzy RDD can be thought of as a type of IV design). This requires thinking of the cut-off itself as a "treatment assignment mechanism". Assuming the cut-off is arbitrary (i.e., random), the fuzzy RDD is akin to a RCT with non-compliance because the treatment assignment mechanism is not perfectly applied. In our IV analysis, we attempted to get around the (possibly confounded) non-compliance issue by focusing our analysis on those who comply with the treatment assignment (i.e., the "compliers"). In a fuzzy RDD, we will import a similar logic and focus our analysis on identifying the causal effect among those individuals who comply with the threshold specific treatment assignment rule.[1]

The fact that we are focusing our analysis on compliers means that we need the same identifying assumptions we needed to estimate the Local Average Treatment Effect (LATE) in the heterogenous treatment effects world of IV. Additionally, note that among the compliers, the probability of treatment at the threshold jumps from 0 to 1. Therefore, to identify the effect among compliers at the threshold, we also need to make the same identifying assumption as in a sharp RDD.

So, in total, there are 2 identifying assumptions if we want to identify the CACE around a fuzzy discontinuity:

1. **Expected potential outcomes are continuous in the running variable around the threshold:** This is same as the assumption we make for sharp RDDs. Note that this implies that the threshold cannot have a direct effect on the outcome except through its effect on the treatmen and that the threshold must be entirely arbitrary, i.e., there should be no other variable that is associated with both the threshold and the outcome and people cannot select themselves to be above or below the threshold.
2. **Monotonicity:** The threshold must affect the probability of receiving treatment in the same direction for all individuals.

[1] A key difference between a standard IV analysis and a fuzzy RDD analysis is that in the former, we focus on all compliers in our study population whereas in the latter, we focus only on those compliers at the threshold.

Note, we are not explicitly assuming relevance here because if the treatment assignment probability did not jump at the cut-off, then we would not even have a RDD. In other words, we have already assumed relevance by assuming that we are in a setting where we can apply a RDD.

Exercise 2: True, false, uncertain

1. We can only conduct a regression discontinuity analysis when individuals are randomly sorted around the cut-off. In other words, we can only conduct a regression discontinuity analysis when we have a "locally randomized experiment" (i.e., local randomization is a necessary condition for identification).

   False. Treatment effects in a RDD are identified assuming that the potential outcomes are continuous at the cut-off. Although local randomization is a sufficient condition for continuity of potential outcomes at the cut-off, it is not a necessary condition.

2. Covariate imbalance on either side of the cut-off necessarily implies that the continuity of potential outcomes is invalid and therefore a RDD analysis is not feasible.

   False. Covariate imbalance does not necessarily imply that potential outcomes are not continuous at the cut-off. If covariates jump discontinuously at the cut-off, then we would be worried about there being a violation of the continuity assumption. The fact that the distribution of covariates are different on either side of the cut-off is not necessarily problematic.

3. While variables like self-reported income can be manipulated, biological variables like age or BMI are always free from manipulation.

   False. Manipulation can arise even in biologic variables, either due to intentional manipulation (e.g., people try to gain or lose weight in order to qualify for a benefit) or misreporting (e.g., someone reports that they are older than they are in order to receive a benefit). While manipulation may be less likely when people are likely unaware of the level of their particular running variable (e.g., CD4 cell count), those taking the measurements (nurses and doctors, for example) might have incentives to over- or under-report these values.

Figure 5, adapted from de la Cuesta and Imai (2016) highlights the difference between the continuity assumption and the local randomization assumption. Local randomization implies continuity because local randomization assumes that the potential outcomes under different treatment levels are the same on either side of the cut-off. In other words, local randomization assumes that the potential outcome function under any treatment level is a flat line within the bandwidth under consideration on either side of the cut-off. In contrast, the RDD assumption of continuity is more flexible than this in that it allows the potential outcome function under any treatment level to take any arbitrary functional form. The only restriction is that these functions must be continuous at the cut-off.

When are RDD assumptions violated?

The assumption of potential outcomes being continuous at the cut-off can be violated if:

1. There is manipulation of the running variable. In other words, if individuals get to choose their value of the running variable, then they may be able to sort themselves onto one side of the cut-off or the other because they believe receiving a certain level of treatment may be more beneficial for them. Because of this sorting, folks around the cut-off may be fundamentally different from each other, meaning that there may be a discontinuous jump in the potential outcome at the cut-off.

2. Another variable which affects the potential outcome also jumps at the cut-off. If this were the case, then the potential outcome under whatever level of treatment is affected by this variable will also jump at the cut-off.

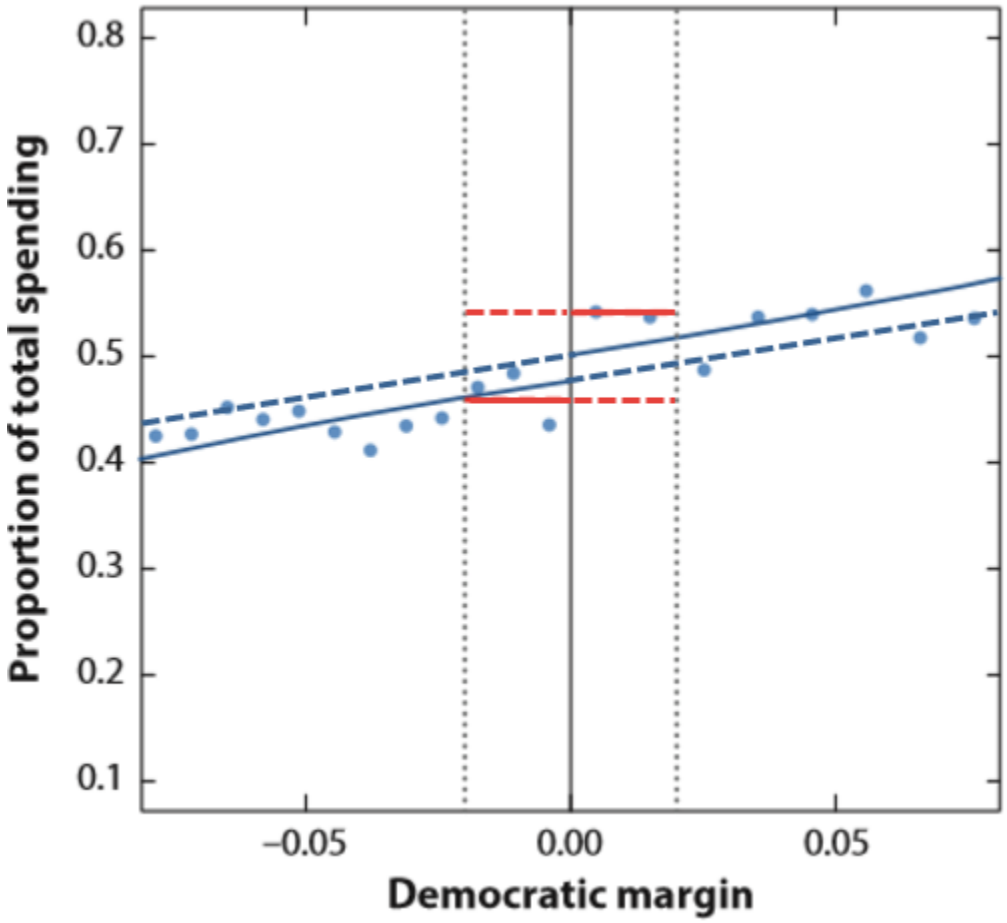


Figure 5: Continuous potential outcomes vs. local randomization

Estimating the causal effects using data

Note that Equation 2 and Equation 4 identify the sharp and fuzzy RDD treatment effects respectively in our data. Below we discuss the steps necessary to estimate the regression functions implied by these formulae.

Sharp RDD

There are three steps to estimating the causal effect in a sharp RDD. These are:

1. Center the running variable around the cut-off. Assuming that individuals with $X_i \geq c$ receive treatment, create an indicator variable for treatment A which equals 1 if $X_i \geq c$ and 0 otherwise.
2. Trim the sample to include only observations that are within a reasonable window around the cut-off c. This sample is often called the "discontinuity sample" in the RDD literature. In general, you want to choose a value h where $h > 0$ such that the discontinuity sample includes all observations that satisfy $c - h \leq X_i \leq c + h$ where X refers to the running variable. There are a few algorithms to choose the value of h. Alternatively, you could present results across different bandwidth specifications to demonstrate robustness of the results to bandwidth choice.
3. Fit a regression model of the outcome Y on the running variable X and the treatment indicator A. The sharp RDD model is generally written as: $y_i = f(x_i) + \tau a_i + \epsilon_i$. Here, τ represents the coefficient of interest and $f(x_i)$ represents any functional form imposed on the running variable when modeling its relationship with the outcome. The functional form can be different on either side of the cut-off c. You may present analyses with various functional form choices.

Fuzzy RDD

We can distill the estimation of the causal effect in a fuzzy RDD to three steps as well:

1. Center the running variable around the cut-off. Create an indicator variable for treatment A which equals 1 if $X_i \geq c$ and 0 otherwise. This is exactly the same as the second step in estimating the ATE in a sharp RDD set-up.
2. Create a "discontinuity sample" by trimming the dataset to include only observations that are within a reasonable window around the cut-off c.
3. Since a fuzzy RDD is essentially an IV analysis, we fit the outcome model using Two-Stage Least Squares (2SLS). A flexible functional form can be chosen to model the relationship between the running variable and the treatment as well as the running variable and the outcome. The functional form can differ on either side of the cut-off.[2]

Diagnostics in a RDD-based analysis

We can do the following diagnostic checks for both sharp and fuzzy RDDs:

1. **Balance checks:** We do not want other covariates influencing the outcome to jump at the threshold along with the probability of receiving treatment. We can check to see if this is the case empirically for measured variables. Suppose we are concerned that a variable M influences the outcome Y and is also discontinuous at the cut-off. We could check this in the data in two ways: first, we could plot $E[M_i|X_i]$ and visually inspect if there is a discontinuity in the expectation at the cut-off $X = c$. Second, we could run a regression with the same functional forms as the RDD but with the covariate M as the outcome. In this regression, the coefficient of interest would the coefficient on treatment indicator variable.
2. **Sorting around the threshold:** Sorting, or "bunching", around the threshold may indicate that individuals have manipulated their value of the running variable to receive/not receive treatment. We

[2] A challenge with the 2SLS estimator is that it is consistent but biased in finite samples. Because of this, ideally we want to fit the 2SLS estimator with as much data as possible, although this may involve increasing the size of the bandwidth which in turn could introduce bias in our estimates as a result of increasing the likelihood of incorrectly specifying the functional form.

can visually inspect for "bunching" by creating a histogram of the running variable and assessing if there is a discontinuous change in the density of the running variable around the threshold. A discontinous change would (likely) indicate manipulation of the running variable.

3. **Placebo thresholds:** We might be concerned that our estimated treatment effect at the cut-off is due to random chance or that the outcome, by nature, jumps at different points along the running variable. We could conduct a falsification check by constructing two datasets: one dataset includes all observations below the threshold of interest while the other includes all observations above. We can then choose "placebo thresholds" and fit the RDD model to test if there are jumps in the outcome at these thresholds in both datasets. While evidence of a jump in the outcome at placebo thresholds does not directly invalidate any of the identifying assumptions, they cast doubt on the result we obtain when fitting a model at the threshold of interest.

4. **Sensitivity checks:** As noted earlier, a good RDD analysis will present results from regressions that model the X-Y relationship using various functional forms and over different bandwidth specifications.

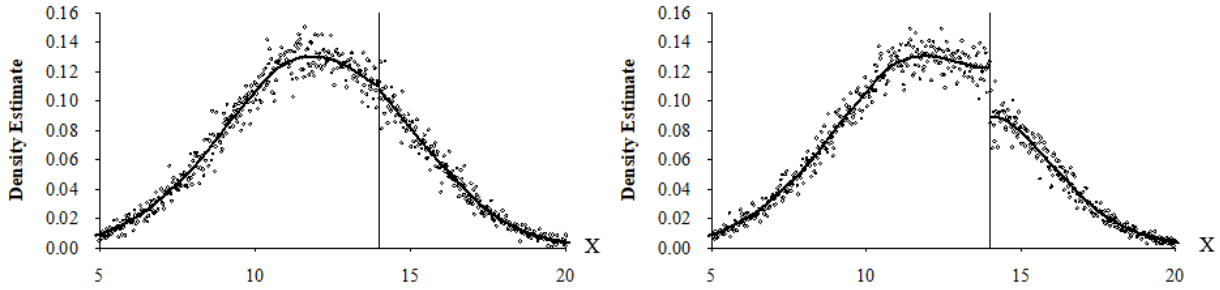


Figure 6: Manipulating the running variable

Limitations of RDDs

Holbein (2019) notes two key limitations of the RDD design:

1. RDD estimates are local estimates, which means that they might not be applicable to people who are farther away from the cut-off. Generalizing from one RDD would therefore require making some particularly strong assumptions about the nature of the treatment effect along the running variable. Furthermore, in the case of a fuzzy RDD, the effect estimates are "doubly local" in the sense that they are not only local to the cut-off but also only applicable to the compliers at the cut-off. Generalizing a RDD treatment effect is an active area of research in econometrics.

2. Since RDDs are based on identifying the effect estimate at the cut-off (generally) using data from individuals in a narrow bandwidth around the cut-off, the data requirements for being able to precisely estimate the treatment effect are very large. Deke and Dragoset (2012) note that RDDs may need as much as 9-17 times the sample size compared to a RCT to produce an effect estimate with the same degree of statistical precision.

Best practices for a RDD study

Moscoe et al (2015) suggest that a convincing RDD study will report the following:

- A discussion of whether the RDD is appropriate given the study context
- A clear presentation of the assignment rule
- Visual evidence of a discontinuity in the probability of receiving treatment and in the outcomes
- Covariate balance tests to assess discontinuity in pre-treatment variables

  - This can be done by running regressions using the same form as the RDD model, but with the relevant covariate as the outcome
  - This can also be done by showing histograms of covariate values to demonstrate that they are continuous around the threshold.
- A histogram of the running variable to show that there is no bunching around the threshold.
- Placebo tests, e.g.:
  - Testing whether there is a discontinuity in outcomes among groups that should not be affected by the treatment (e.g. in places that use a different assignment threshold.)
  - Testing whether there are significant discontinuities at other values of the running variable.
- Results using different functional forms and bandwidth

Conducting a regression discontinuity analysis in R: Impact of antiretroviral therapy on body weight of patients with HIV/AIDS

Our research question is: **What is the impact of initiating antiretroviral therapy (ART) at a CD4 count of 350 on body weight among patients living with HIV in South Africa?**

CD4 count measures the number of T lymphocyte cells (CD4 cells) per cubic millimeter of blood and is a measure of the severity of HIV infection. Lower CD4 counts correspond to a lower ability to combat disease infection and a worsening of HIV/AIDS symptoms and complications. There are several examples in the literature of studies using RDD to examine the impact of ART based on CD4 count thresholds. For example, Bor et al (2014) use RDD to estimate the causal effect of initiating ART at a CD4 count of 200 on mortality.

Because early ART regimens were costly and had significant toxic side effects, CD4 cell counts were used to determine eligibility to initiate ART: people living with HIV would only be initiated on ART after their CD4 cell count levels fell below a certain threshold. A brief historical overview of the evolving role of CD4 cell counts in HIV care can be found here.

As antiretroviral drugs became more widely available and as a growing body of literature indicated that early-initiation of ART was associated with better clinical outcomes, guidance on the initiation of ART was revised to include patients at higher and higher CD4 counts (in other words, to include patients at earlier and earlier stages of disease progression). Eventually, in 2015, the World Health Organization revised its guidance, abandoning the use of CD4 cell count thresholds to determine eligibility and instead recommended a "treat-all" approach. In South Africa, CD4 cell count thresholds were used to determine eligibility through 2017. The current national guidelines on ART eligibility state that "All people living with HIV (PLHIV) are eligible to start ART regardless of age, CD4 cell count and clinical stage."

Let's start by loading the data and relevant packages:

```
# Installing necessary packages
if (!require("rdrobust")) install.packages("rdrobust")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("aod")) install.packages("aod")
if (!require("AER")) install.packages("AER")
if (!require("ivpack")) install.packages("ivpack")

# Loading packages
library("ivpack")
library("aod")
library("AER")
library("ggplot2")
library("rdrobust")

# Setting work directory
setwd("C:/Users/danie/Dropbox/Dropbox (Harvard University)/PhD Course for PHS/Labs/PHS2000B_2021/Lab 8 -
```

```
# Loading data
load("RDdata.Rda")
```

Our dataset contains the following variables:

- `cd4_lab`: Lab measured CD4 count (the running variable)
- `art`: Indicator variable for ART receipt (1 if received ART, 0 otherwise)
- `Y`: Weight in lbs 12 months after CD4 count measurement (the outcome variable)
- `age`: Age measured in years (a covariate)
- `income`: Income in ZAR/month (a covariate)

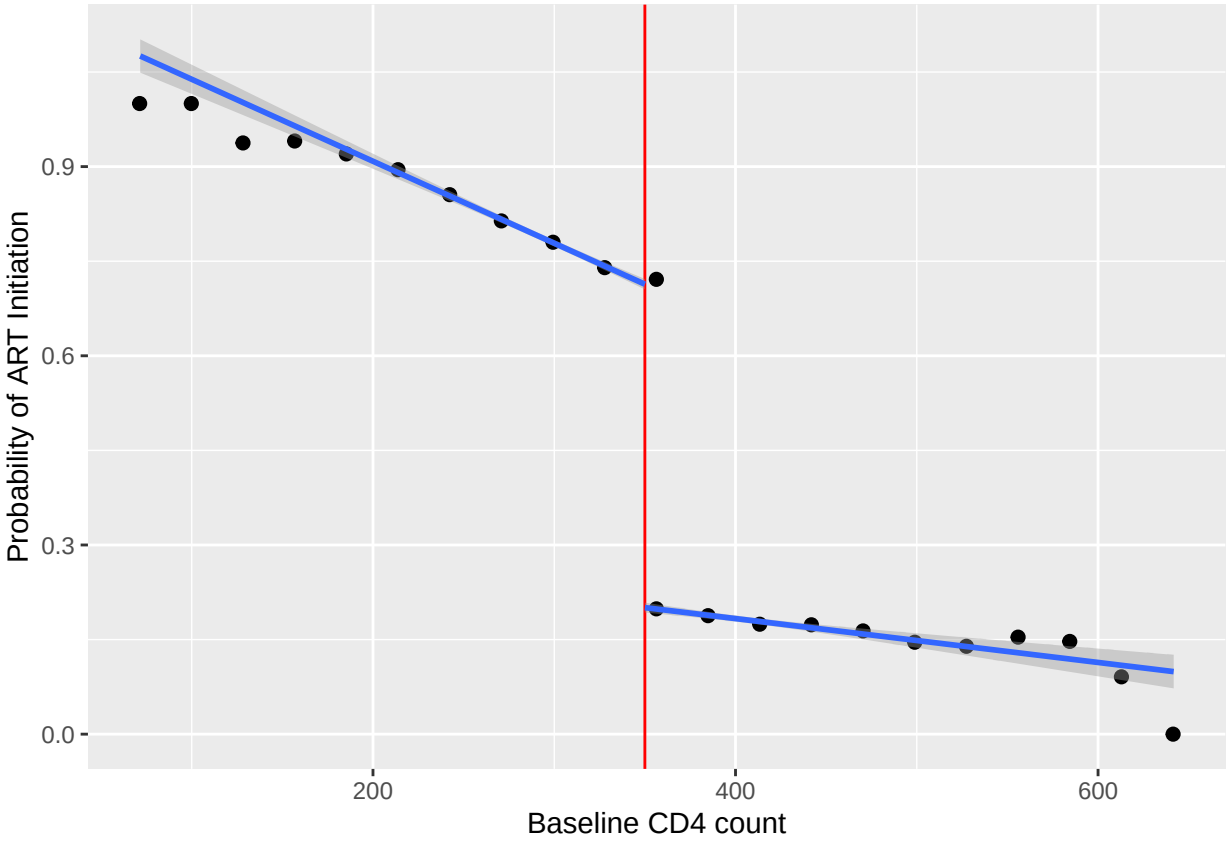
Since we will be measuring the changes in outcomes at the threshold, we start by creating a variable for the threshold called "elig." This indicator variable takes the value 1 if an individual has CD4 count less than 350 and 0 otherwise. A value of 1 for "elig" indicates that an individual is eligible to receive ART under an ART rationing schedule in South Africa.

```
rd.data$elig <- c(ifelse(rd.data$cd4_lab<350, 1, 0))
```

Lots of graphs and regressions!

An RDD analysis normally involves presenting a lot of visual evidence. The first graph we should think about creating in any RDD is a graph showing the probability of receiving treatment with the running variable.

```
ggplot(rd.data, aes(x=cd4_lab,y=art, group=elig)) +
  stat_summary_bin(fun='mean', bins=20, size=2, geom='point') +
  geom_vline(xintercept = 350, color = "red") +
  geom_smooth(method='lm',formula=y~x) +
  labs(x = "Baseline CD4 count", y = "Probability of ART Initiation")
```

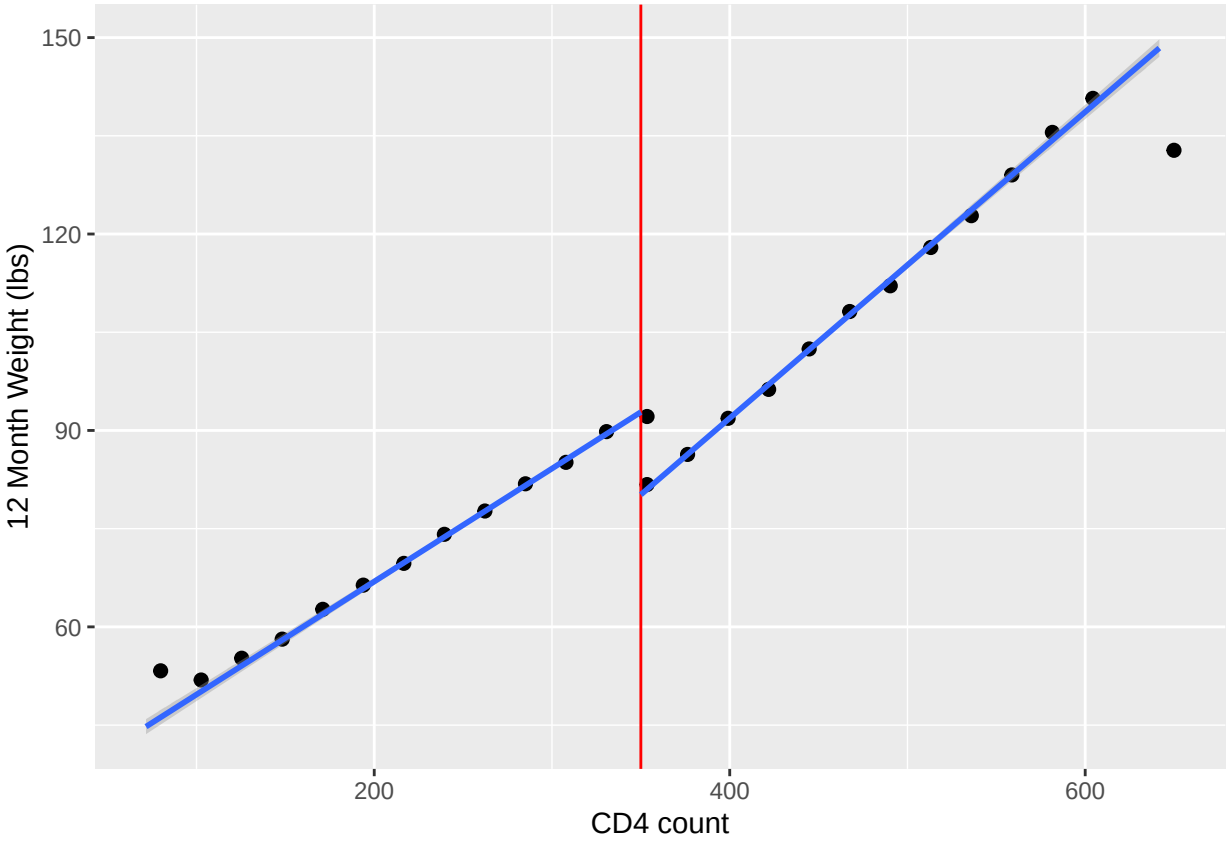


Is this a sharp or a fuzzy RDD? Since the probability of treatment does not go from zero to 1 at the threshold, this is an example of **fuzzy RDD**.

As discussed before, the method of estimating a treatment effect in a fuzzy RDD is the same as estimating a treatment effect in an IV analysis. Therefore, we can use the 2SLS estimator to estimate the treatment effect. However, before we do this, let's make some more graphs to convince ourselves (and our readers) that the outcome also jumps discontinuously at the threshold and that the assumptions necessary for estimating effects in a fuzzy RDD analysis are satisfied.

The graph below plots average weight in lbs at 12 months against the running variable. The graph appears to show that the outcome jumps at the cut-off value of 350. A linear functional form to describe the relationship between the outcome and the running variable also seems to be appropriate, at least near the cut-off.

```
ggplot(rd.data, aes(x=cd4_lab,y=Y, group=elig)) +
  stat_summary_bin(fun='mean', bins=25, size=2, geom='point') +
  geom_vline(xintercept = 350, color = "red") +
  geom_smooth(method='lm',formula=y~x) +
  labs(x = "CD4 count", y = "12 Month Weight (lbs)")
```

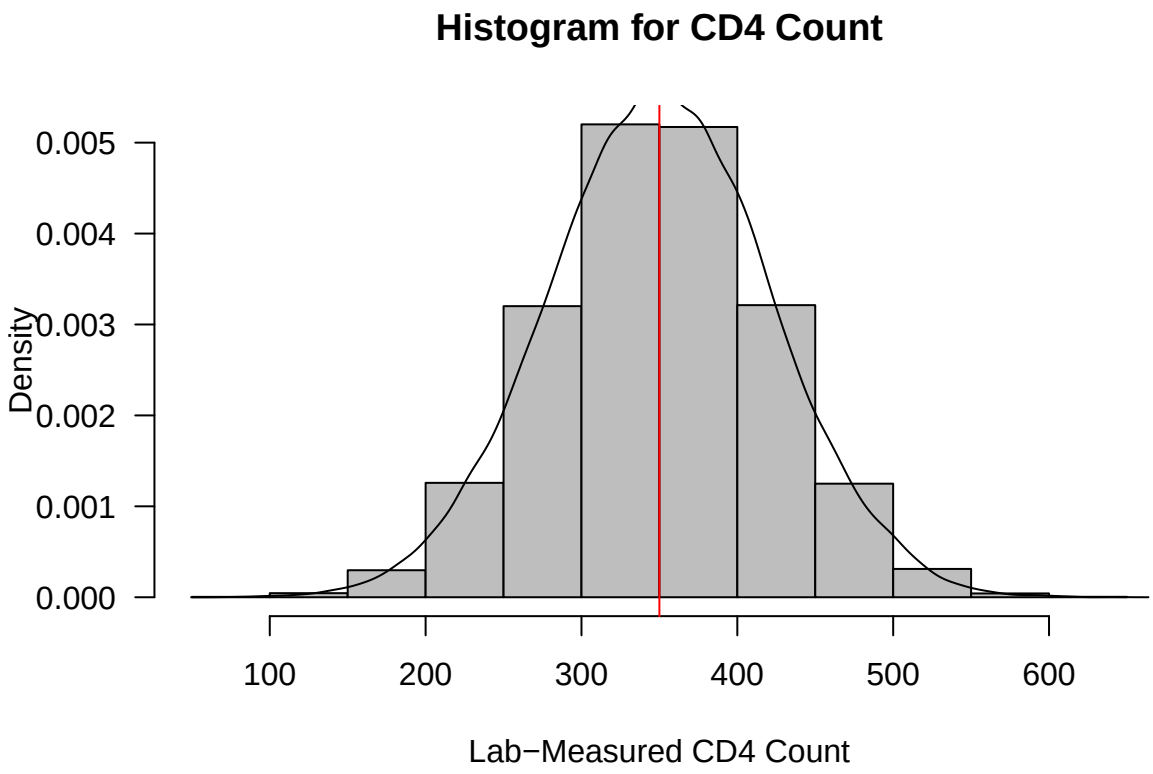


Testing the continuity assumption

Earlier, we argued that we can assess the continuity assumption by creating a histogram of the running variable. We are specifically interested in seeing if there is a discontinuous shift in the density of the running variable around the cut-off. Such a discontinuous shift would indicate that individuals might be manipulating their value of the running variable, which in turn would give us less confidence in asserting that the continuity assumption holds.

Below, we create a histogram of the running variable and impose a density function over it. Visually, we do not see much evidence for "bunching": the density function appears to be relatively smooth around the cut-off. This gives us some confidence that the continuity assumption holds in our analysis.

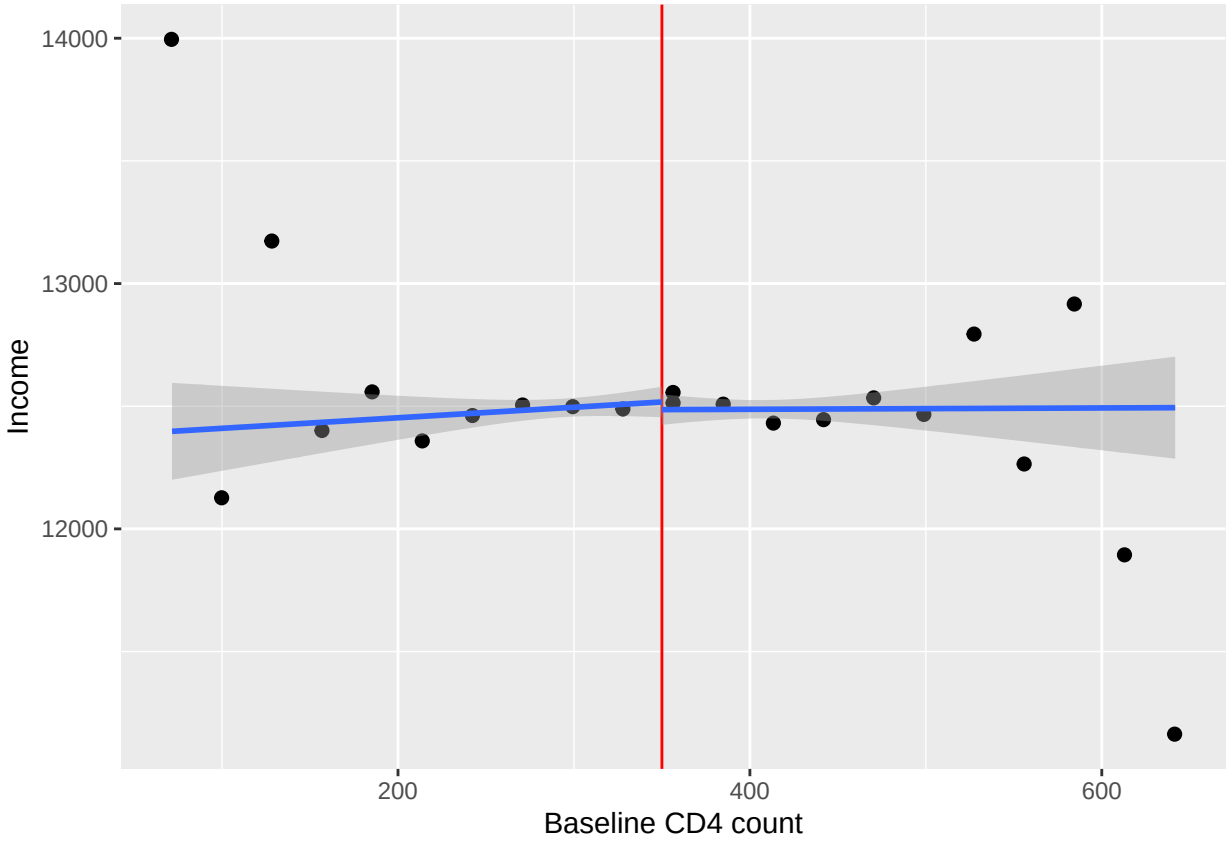
```
hist(rd.data$cd4_lab,
     main="Histogram for CD4 Count",
     xlab="Lab-Measured CD4 Count",
     col="gray",
     las=1,
     breaks=20,
     prob=TRUE)
  abline(v=350, col="red")
  lines(density(rd.data$cd4_lab))
```



However, we are also concerned that other covariates which affect the outcome may be changing discontinuously at the cut-off. Let's test to see if there are discontinuous changes in income at the threshold. We can do this in two ways: by plotting income the same way we plotted treatment above, and by regressing income on the running variable and the threshold.

The graph below seems to suggest that income does not change discontinuously at the cut-off.

```
ggplot(rd.data, aes(x=cd4_lab,y=income, group=elig)) +
  stat_summary_bin(fun='mean', bins=20, size=2, geom='point') +
  geom_vline(xintercept = 350, color = "red") +
  geom_smooth(method='lm',formula=y~x) +
  labs(x = "Baseline CD4 count", y = "Income")
```



The regression results below corroborate what we saw visually when we plotted income against the running varaible.

```
RD.model.incomecheck <- lm(formula = income ~ cd4_lab + elig, data=rd.data)
summary(RD.model.incomecheck)$coefficients
```

```
##                 Estimate  Std. Error   t value  Pr(>|t|)
## (Intercept) 1.239455e+04 129.4454374 95.751138 0.0000000
## cd4_lab     2.289751e-01   0.3147412  0.727503 0.4669213
## elig        3.124580e+01  44.7694452  0.697927 0.4852261
```

If we have other potential confounders in the data, we can similarly test to see if they change discontinuously at the cut-off.

Assessing the IV assumptions

Since we have a fuzzy RDD, our cut-off ("instrument") must satisfy the IV assumptions under heterogenous treatment effects to identify the causal effect of interest (which, in this case, would be the CACE).

Unforunately, the only IV assumption we can empirically test is the relevance assumption. To do so, we can fit a first stage regression of the treatment on the instrument (and the running variable) as follows:

```
# Fitting the first stage model
rd.first <- lm(formula = art ~ cd4_lab + elig, data=rd.data)
summary(rd.first)$coefficients
```

```
##                  Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)  0.5166866073 1.694652e-02  30.48924 2.607750e-202
## cd4_lab     -0.0008254388 4.120477e-05 -20.03261  6.403208e-89
```

```
## elig         0.5125672601 5.861052e-03  87.45311  0.000000e+00
```

```
# Estimating robust standard errors
coeftest(rd.first, vcov = vcovHC(rd.first, "HC1"))
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  5.1669e-01  1.5826e-02  32.648 < 2.2e-16 ***
## cd4_lab     -8.2544e-04  3.8168e-05 -21.626 < 2.2e-16 ***
## elig         5.1257e-01  5.9900e-03  85.570 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this regression, we observe that the probability of treatment jumps by 51 percentage points at the cutoff. Furthermore, even under heteroskedasticity robust estimates of the standard error, we get a t-value on the instrument of 85.570, which implies that the F-statistic of interest is well over the rule-of-thumb value of 10 (recall that when we have only one instrument, the F-statistic is the square of the t-statistic).

The regression of income on the cut-off earlier suggested that the cut-off was not associated with income. In a more formal RDD analysis, we would want to estimate as many regressions of this nature as possible to see if confounders are "balanced" around the cut-off. We could then use the results from these regressions as arguments for or against the independence assumption.

For the time being, let us assume that the exclusion, independence, and monotonicity assumptions hold. Monotonicity is plausible in this setting because it seems unlikely that the probability of ART receipt would decrease if individuals had CD4 count under 350.

Conducting the RDD analysis

Earlier, we claimed that we should conduct the RDD analysis by centering the running variable at the cut-off. Let's first center the CD4 count variable. Note that this will not affect the ART eligibility indicator variable we created earlier.

```
# Centering CD4 count
rd.data$cd4_lab_centered <- rd.data$cd4_lab - 350
```

Since we are essentially conducting an IV analysis, we have a choice in terms of using the Wald estimator or the 2SLS estimator. Recall that the Wald Estimator is the ratio of the reduced form to the first-stage. Furthermore, recall that while the Wald estimator gives us the correct point estimate, it gives us incorrect standard errors.

Let us therefore estimate the CACE using the 2SLS estimator. In R, we can use a canned package called "ivreg" to do this. Let's also estimate a model where we allow the relationship between the running variable and the treatment/outcome to be different on either side of the cut-off. We will also compute heteroskedasticity-robust standard errors for the 2SLS model.

```
# Creating an indicator variable for being on the right of the cut-off
rd.data$cd4_lab_high_centered <- rd.data$cd4_lab_centered*(1-rd.data$elig)

# Estimating the 2SLS model
rdreg1_centered <- ivreg(Y ~ art + cd4_lab_centered + cd4_lab_high_centered , ~ elig + cd4_lab_centered

# Estimating robust standard errors
robust.se(rdreg1_centered)
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           75.2546895  0.2709148 277.780 < 2.2e-16 ***
## art                   24.5996957  0.4978151  49.415 < 2.2e-16 ***
## cd4_lab_centered       0.2045275  0.0027698  73.841 < 2.2e-16 ***
## cd4_lab_high_centered  0.0378542  0.0035762  10.585 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient on ART is the coefficient of interest. Assuming all RDD and IV assumptions hold, we can interpret this as the CACE at the cut-off. Thus, for the compliers at the cut-off, initiating ART increased body weight by 24.5997 lbs.

The following code shows that we would have gotten the same point estimate had we decided to use the Wald estimator:

```
# Estimating the reduced form model
reducedform.centered <- lm(Y ~ cd4_lab_centered*elig, data=rd.data)
summary(reducedform.centered)
```

```
##
## Call:
## lm(formula = Y ~ cd4_lab_centered * elig, data = rd.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.971 -12.465   0.262  12.653  76.733
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           80.191279   0.193601  414.21   <2e-16 ***
## cd4_lab_centered       0.233824   0.002721   85.93   <2e-16 ***
## elig                  12.615109   0.273239   46.17   <2e-16 ***
## cd4_lab_centered:elig -0.061274   0.003842  -15.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.45 on 49996 degrees of freedom
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.2357
## F-statistic:  5140 on 3 and 49996 DF,  p-value: < 2.2e-16
```

```
# Estimating the first stage model
firststage.centered <- lm(art ~ cd4_lab_centered*elig, data=rd.data)
summary(firststage.centered)
```

```
##
## Call:
## lm(formula = art ~ cd4_lab_centered * elig, data = rd.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0055 -0.1892 -0.1487  0.2339  0.8865
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)            2.007e-01  4.147e-03  48.387  < 2e-16 ***
## cd4_lab_centered      -3.479e-04  5.829e-05  -5.968 2.41e-09 ***
## elig                   5.128e-01  5.853e-03  87.611  < 2e-16 ***
## cd4_lab_centered:elig -9.520e-04  8.230e-05 -11.568  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3953 on 49996 degrees of freedom
## Multiple R-squared:  0.3745, Adjusted R-squared:  0.3744
## F-statistic:  9977 on 3 and 49996 DF,  p-value: < 2.2e-16
```

```
# Wald estimate
coef(reducedform.centered)[3] / coef(firststage.centered)[3]
```

```
##    elig
## 24.5997
```

While we can recover the same point estimate without centering the running variable, the procedure is a little bit more involved:

```
# Estimating the reduced form model
reducedform.uncentered <- lm(Y ~ cd4_lab*elig, data=rd.data)
summary(reducedform.uncentered)
```

```
##
## Call:
## lm(formula = Y ~ cd4_lab * elig, data = rd.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.971 -12.465   0.262  12.653  76.733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.647020   1.112929   -1.48    0.139
## cd4_lab       0.233824   0.002721   85.93   <2e-16 ***
## elig         34.061020   1.373119   24.81   <2e-16 ***
## cd4_lab:elig -0.061274   0.003842  -15.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.45 on 49996 degrees of freedom
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.2357
## F-statistic:  5140 on 3 and 49996 DF,  p-value: < 2.2e-16
```

```
# Estimating the first stage model
firststage.uncentered <- lm(art ~ cd4_lab*elig, data=rd.data)
summary(reducedform.uncentered)
```

```
##
## Call:
## lm(formula = Y ~ cd4_lab * elig, data = rd.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.971 -12.465   0.262  12.653  76.733
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.647020   1.112929   -1.48    0.139
## cd4_lab       0.233824   0.002721   85.93   <2e-16 ***
## elig         34.061020   1.373119   24.81   <2e-16 ***
## cd4_lab:elig -0.061274   0.003842  -15.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.45 on 49996 degrees of freedom
## Multiple R-squared:  0.2357, Adjusted R-squared:  0.2357
## F-statistic:  5140 on 3 and 49996 DF,  p-value: < 2.2e-16
```

```
# Recovering the correct point estimate
(coef(reducedform.uncentered)[3] + 350*coef(reducedform.uncentered)[4]) / (coef(firststage.uncentered)[
```

```
##    elig
## 24.5997
```

Sensitivity of the point estimate to bandwidth and functional form specifications

Estimating the RDD model in a narrow bandwidth around the cut-off allows us to reduce our dependency on getting the functional form right. However, smaller bandwidths imply smaller sample sizes which in turn mean lower precision around the point estimates.

Let's see how we can operationalize different bandwidth specifications. First, let's fit the RDD model with a bandwidth of +/- 10 counts around the cut-off:

```
# Estimating a 2SLS model with data restricted to 340 < CD4 count < 360
rdreg2_centered <- ivreg(Y ~ art + cd4_lab_centered + cd4_lab_high_centered , ~ elig + cd4_lab_centered

# Estimating robust standard errors
robust.se(rdreg2_centered)
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           74.852082   0.913278 81.9598  < 2e-16 ***
## art                   25.158793   1.689686 14.8896  < 2e-16 ***
## cd4_lab_centered       0.202700   0.109905  1.8443  0.06519 .
## cd4_lab_high_centered  0.019882   0.156848  0.1268  0.89913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suppose we had a bandwidth of +/- 50 points around the cut-off:

```
# Estimating a 2SLS model with data restricted to 300 < CD4 count < 400
rdreg3_centered <- ivreg(Y ~ art + cd4_lab_centered + cd4_lab_high_centered , ~ elig + cd4_lab_centered

# Estimating robust standard errors
robust.se(rdreg3_centered)
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
```

```
##
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)             74.861043   0.439938 170.1628   <2e-16 ***
## art                     25.767927   0.808011  31.8906   <2e-16 ***
## cd4_lab_centered         0.225321   0.010550  21.3573   <2e-16 ***
## cd4_lab_high_centered    0.021976   0.014858   1.4791   0.1391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we observe difference in both the point estimate and the standard errors across these two estimations. As expected, when the sample size is small, our standard errors are large.

So far, we have assumed a linear relationship between the running variable and the treatment/outcome. Let's relax this assumption and fit a model where we add a quadratic term. This may allow us to capture some of the non-linearities we noticed in the tails of the graphs of the probability of receiving treatment and the running variable and the outcome and the running variable. For the purposes of lab, we fit a quadratic functional form which does not differ on either side of the cut-off.

```
# Estimating a quadratic relationship between the running variable and the treatment/outcome
rdreg4_centered <- ivreg(Y ~ art + cd4_lab_centered + I(cd4_lab_centered^2), ~ elig + cd4_lab_centered +

# Estimating robust standard errors
robust.se(rdreg4_centered)
```

```
## [1] "Robust Standard Errors"
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value  Pr(>|t|)
## (Intercept)             7.5784e+01 2.5977e-01 291.733 < 2.2e-16 ***
## art                     2.4588e+01 4.9797e-01  49.377 < 2.2e-16 ***
## cd4_lab_centered        2.2341e-01 2.1353e-03 104.627 < 2.2e-16 ***
## I(cd4_lab_centered^2) 1.0855e-04 1.0820e-05  10.032 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do our results change relative to the model with the linear trend specification?

Apendix 1: Some comments on functional form

There are two relatively standard methods of estimating the RDD model. First, following Lee and Lemieux (2008), researchers estimate models on either side of the cut-off as follows:

$$y_i = \sum_{k=0}^{K} x_i^k \beta_j + \epsilon_i. \tag{5}$$

In other words, researchers estimate a model which is linear in the parameters but which also posits the functional form relationship between the outcome and the running variable to be a K^{th} order polynomial. Normally, researchers estimate this model across all values of $X < c$ and $X \geq c$ separately and then estimate the difference in the predicted value of Y at the cut-off $X = c$. This method is referred to as using global, K^{th} order polynomials.

A second popular method of estimating the RDD model is to use a local linear regression in a narrow bandwidth around the cut-off.[3] A local linear regression is one which estimates a weighted linear regression in subsets of the data. After estimating a local linear regression on either side of the cut-off, we can then estimate the value of the outcome at the cut-off from both above and below $X = c$ and calculate the difference between these two values as our estimate of the treatment effect at the cut-off.

Which one of these two methods is best to use? Gelman and Imbens (2018) strongly come down in favor of using the local linear regression (or local quadratic regression) method. They provide three arguments for this:

1. Any coefficient estimate from a global polynomial regression model can be considered to be a weighted average of the outcome amongst the treated and control groups. However, the higher the order of the global polynomial function, the more extreme and unreasonable the values of the weights along the running variable far away from the cut-off. In other words, these models tend to more heavily (and unreasonably) weight values that are farther away from the cut-off, which is a problem because we want to get the functional form as correct as possible near the cut-off to identify the treatment effect of interest.
2. Effect estimates from models that use polynomial regressions are sensitive to the order of the polynomial. There are no methods yet in terms of choosing what the right order of the polynomial is.
3. Statistical inference based on models that use higher order polynomials are often misleading.

In general, the RDD literature has moved towards using locally linear regressions in a bandwidth close-ish to the cut-off. You can find a short discussion on bandwidth selection in Appendix 3.

Note that the selection of the functional form can be quite consequential for your results. Imagine, for example, that the distribution of the expected value of a variable Y_0 is nonlinear. If we mistakenly model this nonlinearity using a linear functional form above and below the threshold, we may accidentally mistake the nonlinearity for a discontinuity:

[3] In some cases, researchers may also estimate a local regressions linear in parameters but with a K^{th} order polynomial.

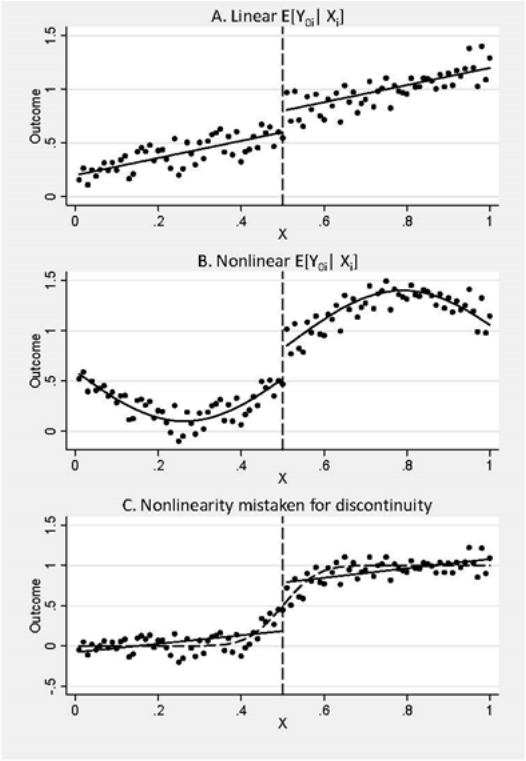


Figure 7: Mistaking nonlinearity for discontinuity in a RDD (from Angrist and Pischke 2009)

Appendix 2: Some comments on controlling for covariates

Many RDD analyses involve the use of local linear regressions with covariate adjustment. However, the consequences of adjusting for covariates, particular in terms of unbiasedness, consistency, and efficiency have not been particularly well studied. Calonico, Cattaneo, Farrell, and Titiunik (2019) provide some guidance on conducting RDD analyses when adjusting for covariates under the assumption that the covariate adjustment is being conducted to increase precision of the RDD estimator.[4]

The authors argue for the use of covariates in a manner which is linear in the parameters and additively separable from the treatment variable. Furthermore, the authors also argue that the same covariates should be adjusted for both above and below the cut-off. These recommendations are conditional on the covariates not jumping at the cut-off as a result of change in treatment status. Although the authors do not provide guidance on how to use covariates when the covariates are being used to ensure the veracity of the continuity assumption in a RDD, they do say that when covariates are being adjusted for such purposes, greater thought needs to be given to the functional form assumptions being made in the model.

[4] As the authors note, this is similar to covariate adjustment in an ideal RCT. I was unable to find similar work on controlling for covariates with the purpose of confounding adjustment.

Appendix 3: Some comments on bandwidth selection algorithms

In class, we discussed manually selecting different bandwidths, estimating our RDD model, and reporting our results. A problem with this, of course, is that researchers can simply cherry pick the bandwidths they want based on the results they get and report this in their papers. In contrast, there are now a number of algorithms available to select the size of the bandwidth for a RDD analysis. These algorithms generally try to optimize the bias-variance trade-off in choosing the bandwidth size. Specifically, the wider the bandwidth, the more data we are using to estimate our models, which means the lower our variance is around the estimate of interest. However, wider bandwidths also imply that we are less sure about the functional form of the regression models below and above the cut-off, which in turn means a higher likelihood of bias.

One popular bandwidth selection algorithm is the one designed by Calonico, Cattaneo, and Titiunik (2014). This bandwidth selection algorithm is applicable to local polynomial regressions of order p and has been discussed in a number of papers by the authors, all of which can be found on their site here. Note that when $p = 1$, we are estimating a local linear regression. Broadly, the idea behind this bandwidth selection algorithm is as follows:

1. Choose a polynomial order p and choose a kernel function $K(.)$. Note, that a kernel regression is a weighted regression model where the weights depend on the choice of the kernel function. For more on kernel functions and kernel regressions, consider reading this.
2. Choose a bandwidth size h such that the mean square error of the estimator is asymptotically minimized. This is called choosing a bandwidth that is "mean squared error optimal". Note that the mean squared error is a function of the bandwidth size because error is related to the bias in an estimator and the bias in the estimator is, as we discussed before, related to the size of the bandwidth.

A key question after the selection of the bandwidth size is estimating confidence intervals around the estimate of the causal effect. Calonico, Cattaneo, and Titiunik (2014) lay out the theory for this in their paper. We will not go into detail here, but broadly speaking, their method involves accounting for the variance in the estimated bias of an estimator when constructing the confidence intervals. Note that the bias needs to be estimated in order to: a) estimate the bandwidth size; and b) to construct “bias-corrected” point estimates.

The Calonico, Cattaneo, and Titiunik bandwidth selection algorithm has been coded as the “rdrobust” package, which is available in both R and Stata formats.

Appendix 4: Some comments on time as a running variable

Imagine that our running variable is not a characteristic of individuals (like their income or CD4 cell count) but one of the intervention itself: time. Before a particular date, the intervention did not exist; after it, everyone was treated. Could we think of time as a running variable?

The answer is yes. Time is continous, after all, and it seems reasonable the potential outcomes can vary continously over time. In fact, RD studies that use time as the running variable have their own name: regression discontinuity in time, or RDiT.

How do RDiT studies compare or differ from other methods, such as interrupted time series (or ITS, which we will cover later in this course) or event studies? An excellent empirical review of RDiT studies (and how they compare to other methods available for time-series data) can be found here. In brief, RDiT studies can be very similar to ITS designs; they differ, primarily, in their specifications. RDiT, for example, may allow for flexibility in specifying the functional form and modeling the time trends before and after the threshold.