ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Autumn 2019

Hongkyu Kim

Regression Discontinuity Design

Submission Date: 16.03.2020

Adviser: Prof. Dr. Marloes H. Maathuis

To my parents, Hye Jin Lee and Young Kwan Kim, for their support,
to my sister, So Hyun,
and to my love, Daniela

Abstract

The regression discontinuity (RD) design is a branch of the observational study, which locally resembles the randomized experiment and is used to estimate the local causal effect of a treatment that is assigned fully, or partly by the value of a certain variable and a threshold. As the RD design is the subject of the causal inference, important concepts of the causal inference are covered to properly proceed discussions. Based on those concepts, the fundamental idea and structure of the RD design are explained including two sub types of the design: the sharp and the fuzzy RD designs. Furthermore, the assumptions of the RD design is formulated, which have been slightly different in different fields. In order to accurately estimate the local causal effect without confounding, we introduce the bandwidth and use the data that are within the bandwidth away from a threshold only. Since there is still no settled way of finding a "good" bandwidth, we propose a novel approach for bandwidth selection along with two existing methods. Performances of these bandwidth selection methods are compared with simulated data, and it can be inferred that the newly proposed method may yield better results. At the end, we intentionally violate the unconfoundedness assumption and analyze three potential confounding models with simulated data.

Keywords: Regression discontinuity design; Causal inference; Observational study; Randomized Experiment; Bandwidth selection; Sensitivity analysis

Contents

List of Figures

Chapter 1

Introduction

The regression discontinuity (RD) design is a quasi-experimental methodology for estimating the causal effect of a treatment where the treatment is determined by whether a certain covariate exceeds a given threshold (Lee and Lemieux, 2010). As an alternative option in the observational study, the RD design was first proposed by Thistlethwaite and Campbell (1960), which analyzed the causal effect of the scholarship on students' future career choice, academic achievements, etc. The core idea of the design was that units with covariate values that are in a small neighborhood of a given threshold are expected to be unconfounded. Thus, one can mimic the random experiment locally, and get the credible result for the causal effect.

After the introduction of the RD design, it was further divided into two types according to the way the treatment is assigned. The first type is called the sharp RD design, which is the case of Thistlethwaite and Campbell (1960). It is characterized by the deterministic assignment of the treatment. The second type is called the fuzzy[1] RD design (Trochim, 1984), which shows the stochastic nature of the treatment assignment. Except these extensions and some works, the RD design have not drawn much attention.

Starting from the late 1990s, the RD design was actively used in the econometric field (Lee and Lemieux, 2010). The increased recognition is in some part due to the work by Hahn, Todd, and van der Klaauw (2001), which introduced the theoretical foundation of the RD design. Not only is the RD design used in the econometric field (Ludwig and Miller, 2005; Lee, 2008), but also in the epidemiology (Bor, Moscoe, Mutevedzi, Newell, and Bärnighausen, 2014; Moscoe, Bor, and Bärnighausen, 2015), medicine (Geneletti, O'Keeffe, Sharples, Richardson, and Baio, 2015; Geneletti, Ricciardi, O'Keeffe, and Baio, 2019), criminology (Berk and Rauma, 1983), educational psychology (Thistlethwaite and Campbell, 1960), and many other fields. It is highly recommended to read some review papers (Lee and Lemieux, 2010; Imbens and Lemieux, 2008; Berk, 2010; Cook, 2008; Cappelleri and Trochim, 2015; Cattaneo, Titiunik, and Vazquez-Bare, 2019) to have a big picture of the RD design.

In Chapter 2, the fundamental concepts of the causal inference including causality, Rubin causal model, randomized experiment, and observational studies are introduced. In Chapter 3, the basic structure of the RD design is introduced including the sharp and fuzzy RD design, and the assumptions of the design is formulated. In Chapter 4, the problem of the estimation and inference is covered. In Chapter 5, we analyze how the violation of the core assumption of the RD design affects results. Finally, we conclude in

[1] The term "fuzzy" was used by Campbell (1969) according to Trochim (1984).

Chapter 6 by summarizing the paper and suggesting some challenges that can be tackled in the future.

Chapter 2

Causal Inference

2.1 Causality

A cause can be defined as a thing that gave rise to a certain result. It is very straightforward to say "C causes R". In fact, the concept of causality is frequently used in everyday lives with people using the word "because". For example, it is common to say "I couldn't sleep last night because I drank coffee." or "He scored high on the test because he is a human, not a monkey.".

The first sentence is quite plausible in common sense, and can be investigated in academic sense. However, the second sentence is problematic in that it attributes the high score to him being a human not a monkey. This is problematic because the fact that he is a human cannot be changed in the sense that we cannot imagine him being a monkey with, for example, the same educational background, social relationships, etc. Instead, we can interpret the second sentence in correlational sense (Holland, 1986) as "Humans scored higher on the test than monkeys did." Therefore, we will restrict the notion of cause to be "a thing that produces a result, and can be (potentially) exposed[1] to other causes" as in Holland (1986).

In many introductory level statistics courses, the major theme is about the correlation. For instance, the linear regression, which is the most fundamental and simplest model in statistics, assumes the linear correlation between a response variable and predictor variables.

However, it can be easily deduced that the correlation does not necessarily imply the causality. For example, let's say we have a predictor variable X and a response variable Y, which are positively correlated. If X indeed causes Y, one will have a large value of Y given a large value of X because of the positive correlation. Unfortunately, we can construct a simple counter example where X and Y are positively correlated but no causal relation exists, by introducing a hidden variable Z. If there is a third hidden variable Z that is the common cause of both X and Y, then X and Y can have a positive correlation without having a causal relation. Thus, simply increasing the value of X may not influence the value of Y at all, and one can have a small value of Y even with a large value of X. Figure 2.1 illustrates this point more intuitively.

In order to infer about causality, we will introduce a causal model which will soon fol-

[1] Holland (1986) termed this "potential exposability".

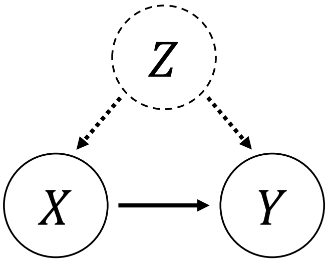


Figure 2.1: X is positively correlated with Y, but is not a cause of Y (The solid circles are observable variables, and the dotted circle is a hidden variable. The solid arrow is a correlation, and the dotted arrows are causal relations.)

low. Then we will define the causal effect[2] (or the treatment effect), and if the causal effect is nonzero with some statistical significance, we will say that the cause being considered indeed produced the result.

2.2 Rubin Causal Model

For a more formal discussion of causality, we proceed our discussion on the Rubin causal model (Holland, 1986; Rubin, 1974) or the Neyman-Rubin causal model. The main idea of the model in Rubin (1974) can be summarized as follows. Let U be a population of interest, u be an unit in the population U, T be a treatment (conceptually same as a cause) on which we want to conduct causal inference, and C be a control treatment. C can be simply an absence of T, or a default treatment in an experiment setting. For example, if a researcher wants to test an effect of a new medicine, then C would be taking a placebo. If an awakening effect of a coffee is to be tested, then C can be drinking water or drinking nothing at all. If we denote an outcome of the unit u as Y_u, the crucial assumption of the Rubin causal model is that we assume two potential outcomes $Y_u(T)$ and $Y_u(C)$ for the unit u. $Y_u(T)$ denotes a potential outcome if u received the treatment T, and $Y_u(C)$ denotes a potential outcome if u received the control treatment C. Then the causal effect of the treatment T relative to C[3] on u, τ_u, is defined as follows.

$$\tau_u \coloneqq Y_u(T) - Y_u(C), \tag{1}$$

or

$$\tau_u \coloneqq Y_u(1) - Y_u(0), \tag{2}$$

where 1 indicates that the treatment T is given and 0 indicates that the treatment T is not given.

However, there is one problem with this unit-level causal effect. Without proper assumption or justification[4], one can observe only one of the two potential outcomes. This is called the *fundamental problem of causal inference* (Holland, 1986; Cattaneo et al.,

[2] There are two important issues in causality. One is the effects of causes, and the other is the causes of effects (Holland, 1986). The latter topic has its own significance, but we will keep our focus on the effects of causes.

[3] As Holland (1986) noted, the causal effect of a treatment can only be defined in comparison with another treatment.

[4] For example, if one can be assured that effects of treatments vanish after some time and the effects does not change with time, then one can measure $Y_u(1)$ first and wait for the effect to vanish, and measure $Y_u(0)$

2019). One can circumvent the problem with a help of statistical approach, which is to estimate the average causal effect, rather than the unit-level causal effects (Holland, 1986; Rubin, 1974). If we denote τ to be the average causal effect of T relative to C, we can express τ as

$$\tau := \mathbb{E}[Y_u(1) - Y_u(0)] = \sum_{u \in U} (Y_u(1) - Y_u(0)) P(u), \tag{3}$$

where P is the probability mass function over the population U. By the linearity of expectation, the expectation term in (3) can be decomposed into a subtraction of two expectations as follows.

$$\tau = \mathbb{E}[Y_u(1)] - \mathbb{E}[Y_u(0)]. \tag{4}$$

By separately estimating $\mathbb{E}[Y_u(1)]$ and $\mathbb{E}[Y_u(0)]$ with units which received the treatments T and C respectively, we can circumvent the fundamental problem of causal inference and estimate τ.

The fundamental problem of causal inference can also be overcome with some assumptions such as temporal stability, causal transience, unit homogeneity, constant effect, etc (Holland, 1986). Temporal stability and causal transience assume that there is no confounding effect over time and prior exposure to the control treatment respectively. Thus, one measures $Y_u(0)$ first, and $Y_u(1)$ after some time to get $Y_u(1) - Y_u(0)$. Unit homogeneity assumes the existence of units u and v such that $Y_u(0) = Y_v(0)$ and $Y_u(1) = Y_v(0)$. Therefore, the causal effect for the unit u, $\tau_u (= \tau_v)$, can be written as

$$\begin{aligned} \tau_u &= Y_u(1) - Y_v(0) && (5) \\ &= Y_v(1) - Y_u(0). && (6) \end{aligned}$$

These assumptions make the problem easier, but they are strong assumptions and can only be validated under restricted situations. For example, temporal stability and causal transience are often valid in physical experiment (Holland, 1986).

2.3 Randomized Experiment

An intuitive and highly preferred way to estimate the average causal effect is to conduct an experiment, which divides the sub-population of U into two groups, the treatment group and the control group (usually with similar sizes) (Freedman, Pisani, and Purves, 2007; Rubin, 1974). Literally, the treatment group receives a treatment T, and the control group receives a control treatment C. A sample mean of the treatment group estimates $\mathbb{E}[Y_u(1)]$, and a sample mean of the control group estimates $\mathbb{E}[Y_u(0)]$. Then, the difference of the sample means is an estimator of the average causal effect τ. If we denote a sample mean of the treatment group by $\bar{\mu}_{\text{T}}$, a sample mean of the control group by $\bar{\mu}_{\text{C}}$, then an estimator $\hat{\tau}_{\text{Exp}}$ (of τ) from an experiment can be written as

$$\hat{\tau}_{\text{Exp}} := \bar{\mu}_{\text{T}} - \bar{\mu}_{\text{C}}. \tag{7}$$

However, this simple-minded approach has a weakness, namely, confounding effects. To estimate an effect of a treatment "accurately", samples from each group should be

to subtract it from $Y_u(1)$. However, this is a strong assumption because the treatment might not vanish, or the treatment can change the unit fundamentally, which makes the observed causal effect inaccurate. Furthermore, there can be a confounding effect through time, i.e., effects of the treatment changes over time. Refer temporal stability, and causal transience in Holland (1986) for more details.

similar in every aspect except a treatment received only. Instead, if the samples have substantial differences[5] other than the treatment, then it can be the case that the observed effect $\hat{\tau}_{\text{Exp}}$ is actually a mixture of many different effects. Thus, $\hat{\tau}_{\text{Exp}}$ is likely a biased estimator of τ. This confounding effect is called the sampling bias. The sampling bias is commonly observed unless the samples are selected cautiously.

A simple but powerful solution to the confounding is to assign the treatment randomly. For example, if a total sample size to be selected is $2N$, then pick N units randomly and assign a treatment T on them, and again pick N units randomly from the rest and assign a treatment C on them. Finally, one defines an estimator $\hat{\tau}_{\text{RCE}}$ (of τ) to be the subtraction of the two sample means as before. Therefore, $\hat{\tau}_{\text{Exp}}$ and $\hat{\tau}_{\text{RCE}}$ look exactly the same.

$$\hat{\tau}_{\text{RCE}} \coloneqq \bar{\mu}_{\text{T}} - \bar{\mu}_{\text{C}}. \tag{8}$$

The only difference is the way one assigns the treatment. This procedure is called the randomized experiment (Freedman et al., 2007), or the randomized trial. The term "randomized" is derived from this random assignment of the treatment. With the random assignment of treatments, one not only has the unbiasedness of an estimator $\hat{\tau}_{\text{RCE}}$, but also can precisely calculate the p-value of a hypothesis testing, especially one with the null hypothesis $H_0 : \tau = 0$ (Rubin, 1974).

The figure 2.2 is a simulation result that shows that the randomized experiment can indeed solve the confounding problem in an experiment. Appendix B contains detailed R-codes of the simulation. For the simulation, we generated 100 samples[6] which can be divided into two different groups of size 50. Let us denote those groups A and B. The group A was generated from a gaussian distribution with mean 5, and standard deviation 1, and the group B was generated from a gaussian distribution with mean 10, and standard deviation 1. Therefore, units from the group A and group B have substantial difference.

$$\text{Group A } \sim \mathcal{N}(5, 1) \quad \text{and} \quad \text{Group B } \sim \mathcal{N}(10, 1). \tag{9}$$

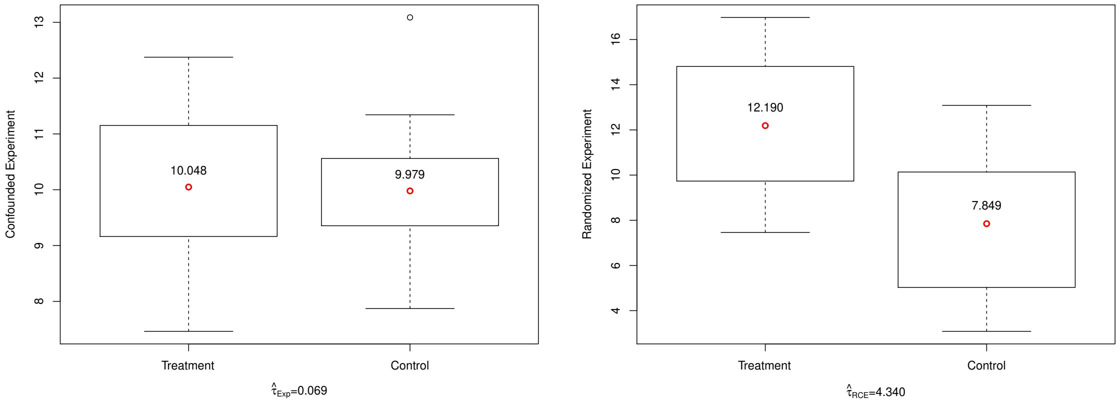
In addition, we assumed the true treatment effect to be 5, but to take randomness into account, we modeled the treatment effect to be distributed normally with mean 5 and standard deviation 0.5. Likewise, the control treatment effect was modeled to be distributed normally with mean 0 and standard deviation 0.5.

$$\text{Treatment Effect } \sim \mathcal{N}(5, 0.5^2) \quad \text{and} \quad \text{Control Effect } \sim \mathcal{N}(0, 0.5^2). \tag{10}$$

Suppose the experiment is designed to assign treatment for 50 unit and control treatment for the rest 50 units. In a non-randomized experiment, it can be the case that the treatment is given only to group A. The box plot (a) in figure 2.2 shows the result in this case. Even though the true treatment effect was normally distributed with mean 5, the observed effect was nearly 0 ($\hat{\tau}_{\text{Exp}} = 0.069$). This is because the treatment effect was compromised with the substantial difference between group A and group B. However, if we assign the treatment randomly, we get the result as the box plot (b) in figure 2.2. For the box plot (b), 50 units were randomly selected to receive the treatment where 27 were from the group A and 23 were from the group B. Just by randomly selecting treatment units, the group factor was well balanced, and produced a significantly better result. The observed effect $\hat{\tau}_{\text{RCE}}$ was 4.340 where the deviation was caused by the variability of units and effects.

[5] These are called the confounders.

[6] More precisely, the 100 samples are realizations of a certain covariate of 100 units.



(a) Confounded Experiment

(b) Randomized Experiment

Figure 2.2: Box plot results of confounded experiment and randomized experiment with fictitious data

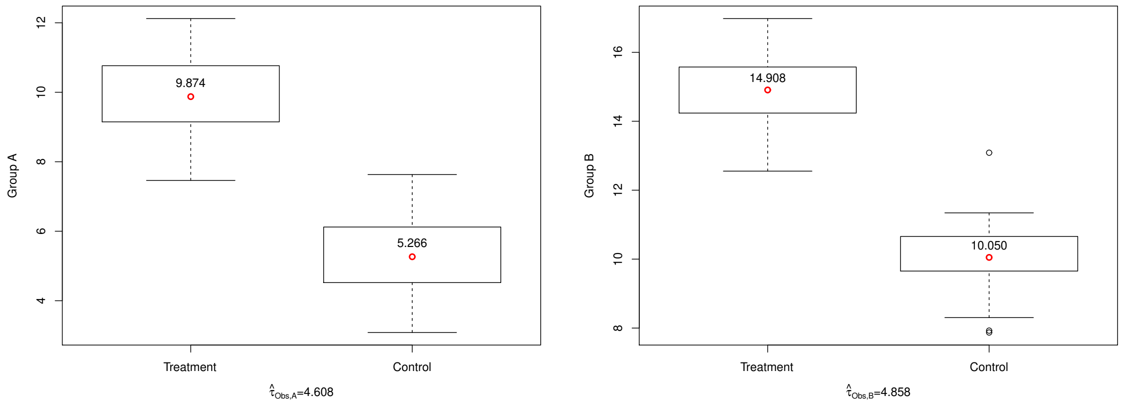
2.4 Observational Studies

It is advisable to implement a randomized experiment to infer about the causal effects whenever possible (Freedman et al., 2007; Rubin, 1974). Nevertheless, it is not always feasible to conduct an experiment for many reasons. As mentioned in Rubin (1974), there are three (and more) issues regarding the feasibility. First of all, it can be very costly to conduct an experiment. In order to get more accurate results, experiments sometimes need to be done in big scale. In a big scaled experiment, the required cost, such as personnel expenses, gets increased by a similar or larger scale. Furthermore, giving a treatment of interest itself can be expensive. Secondly, there is an ethical issue. For instance, let us suppose a researcher wants to know how much secondhand smoking affects the likelihood of getting lung cancer. It is obviously absurd to force people to be exposed to smoke only for the sake of getting the answer. Lastly, an experiment may take a long time to produce the results. Suppose a researcher is interested in the effect of high grade point average (GPA) from high school on the annual salary 20 years after graduation. Then, it needs at least 20 years in an experimental setting, just to have the data. This is an extremely inefficient way of inferring the causal effect.

An alternative option, when a randomized experiment is infeasible, is an observational study (Freedman et al., 2007). This method literally "observes" pre-collected data from some organizations or institutions rather than implementing a new experiment, hence no control over the treatment assignment. However, using pre-collected data solves the potential problems mentioned above. There is no cost and time needed in making data, and it is free of ethical issues because the pre-collected data is the result of something already happened, and is usually a consequence of agents' free will.

As explained earlier, good properties of the randomized experiment follow from the random assignment of a treatment. However, observational studies have no control over the treatment assignment, i.e., treatments are not randomly assigned. As a consequence, the unbiasedness of an estimator is no longer guaranteed, and p-value is not precise any more. Confounding matters again. The main goal of the observational study is to retrieve the causal effect with as little confounding as possible. Some techniques are developed for this purpose and sometimes a well designed observational study can provide as good result as a randomized experiment.

One example is the "matching" (Rubin, 1974; Rosenbaum and Rubin, 1983, etc) method. The crucial idea of matching is that there exists a variable such that units with a similar value for the variable are comparable to each other, i.e., there is no or less confounding within the units with a similar value for the variable. Although this is only an assumption and it cannot be justified in some cases, matching can retrieve the average causal effect when the assumption is true, or nearly true. This can be checked with the same data used for (b) of figure 2.2 (Appendix B). Let us suppose the data is given for our purpose, and there is no information whether the assignment was random or not. Instead, it will be assumed that the group category of each unit is known. To apply matching method, we will assume that units from the same group are comparable, i.e., the group variable will be "matched". As noted earlier, 27 units were from the group A and 23 units were from the group B among 50 units that were given treatment. Thus, within the group A, 27 units received the treatment and 23 units received the control treatment. It is exactly the opposite for the group B. The results are given in figure 2.3. There are two separate results because the binary group variable was matched. The observed effect for the group A ($\hat{\tau}_{\text{Obs, A}}$) was 4.608, and for the group B ($\hat{\tau}_{\text{Obs, B}}$) was 4.858. Interestingly, the results are less biased than that of the randomized experiment in figure 2.2, which is because there is one less variability by matching on the group variable. Therefore, when the assumption of the matching method is true, an observational study can provide even better results than a naive randomized experiment.



(a) Result for group A

(b) Result for group B

Figure 2.3: Box plot results of matching on groups

Chapter 3

Regression Discontinuity Design

3.1 The original work

The RD design first emerged as one alternative option in observational study methodologies in Thistlethwaite and Campbell (1960). In the original paper, Thistlethwaite and Campbell (1960) proposed the RD design and used it to analyze a causal effect of receiving a scholarship on future career choice, achievements, etc. The scholarship was given to students based on certain test scores which was grouped into 20 discrete scales from 1 to 20. Score 11 was the borderline so that students with score less than 11 didn't receive the scholarship, while those with 11 or greater received the scholarship.

The crucial idea of Thistlethwaite and Campbell (1960) was that, in a setting where a treatment[1] assignment is determined solely by a value of an observable covariate[2], units that lie near a cutoff[3] are good comparisons to each other (Lee and Lemieux, 2010). Just by considering units near a cutoff, one can find a trace of a randomized experiment and overcome the *fundamental problem of causal inference* (Holland, 1986; Cattaneo et al., 2019). With the data of students from 1 to 10, Thistlethwaite and Campbell (1960) extrapolated the data at score 11 by linear regression. The extrapolated data represents the counterfactual outcome that students with score 11 would give if they had not received the scholarship. Finally, the difference of the outcomes was used to estimate the causal effect of the scholarship on students with score 11.

Another important idea embedded in Thistlethwaite and Campbell (1960) was that, when the RD design is applicable, the treatment assignment can be considered as good as the random assignment near a cutoff with some mild assumptions (Cattaneo et al., 2019; Lee and Lemieux, 2010; Geneletti et al., 2015). If units have a full control on the assignment variable, then the random assignment near a cutoff is not justified, and confounding can arise (Lee and Lemieux, 2010). However, if there is some random error in a covariate, and if units have no control, or "imprecise" control over the value of their covariates, then the random assignment is still valid (Lee and Lemieux, 2010; Lee, 2008). Although they are untestable in many cases, the assumptions are conceptually intuitive and yield an attractive conclusion that a randomized experiment is observed locally. This

[1] As before, "treatment" means a treatment of interest, and the causal effect, or the treatment effect is always defined relative to a control treatment.

[2] This covariate is also referred to as score, assignment variable, forcing variable, running variable, etc.

[3] This is the threshold for receiving a treatment. For example, units with a covariate value greater than or equal to a cutoff receive a treatment.

is one reason why the RD design is preferred among other observational study methods when applicable.

In addition, the concept of the random treatment assignment around a cutoff may suggest an alternative framework to the classical RD design, which is a permutation test type framework. The alternative approach can be applied when an assignment variable is discrete, or can be used as a backup evidence for the design (Cattaneo et al., 2019; Cattaneo, Idorobo, and Titiunik, 2018).

3.2 Basics

As in a regression setting, there are a covariate X and an one-dimensional response (or an outcome) variable Y. For a simplicity of analysis and further discussion, we assume a covariate X is also a scalar[4]. In addition, there is a threshold value c for a covariate, which is called a cutoff.

We introduce two more variables which are related to a cutoff. One is a threshold indicator D. D is an indicator variable that attains a value 1 if $X \geq c$, or 0 otherwise.

$$D = \mathbb{1}\{X \geq c\} = \begin{cases} 1 & \text{if } X \geq c, \\ 0 & \text{if } X < c. \end{cases} \tag{11}$$

The other is a treatment variable, T. T is also an indicator variable, but for a treatment of interest. That is,

$$T = \begin{cases} 1 & \text{if a treatment is given,} \\ 0 & \text{if a control treatment is given.} \end{cases} \tag{12}$$

In the original work, the threshold indicator D was identical to the treatment indicator T, i.e., $D \equiv T$. However, they do not have to be identical. The two subclasses of the design will be explained in the following discussion according to whether D and T are identical or not. Another thing to note here is that the treatment does not affect the covariate value. Thus, if the covariate is given, then it is considered fixed. Additionally, we will denote unit-specific variables as X_u, Y_u, D_u and T_u which correspond to variables above for each unit u in a population of interest.

On the Rubin causal model, an observed outcome Y_u for an unit u can be expressed with variables defined above as follows.

$$Y_u = (1 - T_u) \cdot Y_u(0) + T_u \cdot Y_u(1) = \begin{cases} Y_u(0) & \text{if a treatment is given,} \\ Y_u(1) & \text{if a control treatment is given.} \end{cases} \tag{13}$$

This reformulation does not really help unless we can observe both of $Y_u(0)$ and $Y_u(1)$. However, this can be helpful in interpretation if we can give adequate counterfactuals in each case. As mentioned earlier, we are interested in the expectation of outcomes, not the unit-level outcomes. The RD design infers the counterfactuals based on the assumption that the conditional expectations of $Y_u(0)$ and $Y_u(1)$ given a covariate are continuous (Cattaneo et al., 2019; Imbens and Lemieux, 2008; Lee and Lemieux, 2010, etc). Then, with the continuity assumption, the local causal effect[5] at the threshold is estimated rather

[4] One can extend a discussion to a multidimensional covariate in a similar way (Rubin, 1977).

[5] This local causal effect can also be interpreted as a weighted average of causal effects according to Lee and Lemieux (2010).

than the average causal effect. The idea behind the continuity assumption is that if we assume the continuity then the discontinuity at the threshold can be interpreted as the local causal effect (Imbens and Lemieux, 2008).

The RD design can be used when a treatment is given strictly or loosely by the value of a certain variable, and the treatment assignment probability $\Pr(T_u = 1|X = x)$ is discontinuous at the threshold. This will be explained in more details in the following discussions.

3.3 Sharp Regression Discontinuity Design

There are two types of the RD design depending on whether the treatment assignment is deterministic or stochastic. One is the sharp RD (SRD for short) design, and the other is the fuzzy RD (FRD for short) design (see, for example, Trochim, 1984). The SRD design is the case where the treatment is given depending strictly on whether $X_u \geq c$ or not. This can be written equivalently as

$$T_u \equiv D_u \equiv \mathbb{1}\{X_u \geq c\}. \tag{14}$$

Now let us define two functions $\mu_0(x)$ and $\mu_1(x)$ (Cattaneo et al., 2019) where

$$\mu_0(x) \coloneqq \mathbb{E}[Y_u(0)|X_u = x] \quad \text{and} \quad \mu_1(x) \coloneqq \mathbb{E}[Y_u(1)|X_u = x]. \tag{15}$$

Both of $\mu_0(x)$ and $\mu_1(x)$ will be assumed to be continuous. Based on these two functions, we can define a function $\mu(x)$ (Cattaneo et al., 2019), which is a conditional expectation of observed outcomes.

$$\mu(x) \coloneqq \mathbb{E}[Y_u|X_u = x], \tag{16}$$

where $Y_u = (1 - T_u) \cdot Y_u(0) + T_u \cdot Y_u(1)$. Using a simple fact that T_u is binary, we can derive the following relationship from (16).

$$\begin{aligned} \mu(x) &= \mu_0(x) \cdot \Pr(T_u = 0|X_u = x) + \mu_1(x) \cdot \Pr(T_u = 1|X_u = x), && (17) \\ &= \begin{cases} \mu_0(x) & \text{if } x < c, \\ \mu_1(x) & \text{if } x \geq c. \end{cases} && (18) \end{aligned}$$

Because of the assignment rule, $\mu_0(x)$ is observed only for $x < c$, and $\mu_1(x)$ is observed only for $x \geq c$. Each unobserved part can be extrapolated, but are not really necessary for our purpose. What is estimated in the SRD design is the size of the jump (τ_{SRD}) at a cutoff, which is

$$\tau_{\text{SRD}} \coloneqq \lim_{x \downarrow c} \mathbb{E}[Y_u|X_u = x] - \lim_{x \uparrow c} \mathbb{E}[Y_u|X_u = x]. \tag{19}$$

However, what we want to infer about is the local causal effect at a cutoff c, which is

$$\begin{aligned} \tau(c) &\coloneqq \mathbb{E}[Y_u(1) - Y_u(0)|X_u = c], && (20) \\ &= \mu_1(c) - \mu_0(c). && (21) \end{aligned}$$

Here we use the assumption of continuity of $\mu_0(x)$ and $\mu_1(x)$. Then, we can easily deduce that $\tau_{\text{SRD}} = \tau(c)$ (Imbens and Lemieux, 2008). Actually, Hahn et al. (2001) showed that the continuity of $\mu_0(x)$ and $\mu_1(x)$ at $x = c$ is enough to identify the local causal effect $\tau(c)$.

As an educational purpose, exemplary figures of SRD are given (See Appendix B for R codes). The figure 3.1 corresponds to the treatment assignment probability, and the figure 3.2 corresponds to conditional expectation functions of both observed and potential outcomes. The cutoff is set to be at $x = 60$. As expressed in 18, the conditional expectation of observed outcomes follows exactly $\mu_0(x)$ when $x < 60$, and $\mu_1(x)$ when $x \geq 60$. The size of the jump at $x = 60$ is precisely the local causal effect of SRD τ_{SRD}.

The SRD design is frequently applied where there is a clear-cut rule to assign a treatment, or an incentive (Imbens and Lemieux, 2008). The fields range from criminology (Berk and Rauma, 1983), to educational psychology (Thistlethwaite and Campbell, 1960) and many other.

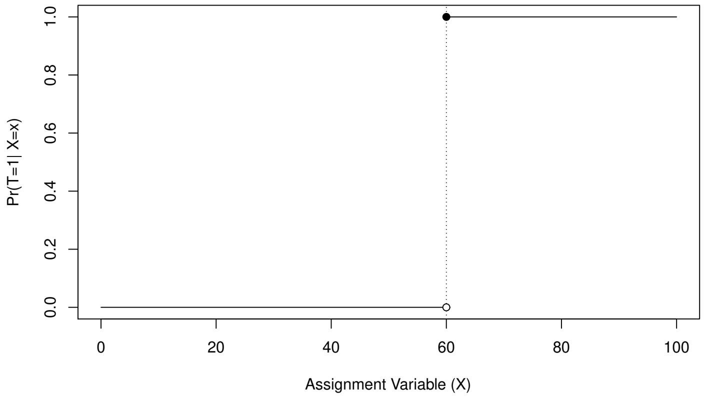


Figure 3.1: Treatment assignment probability in SRD

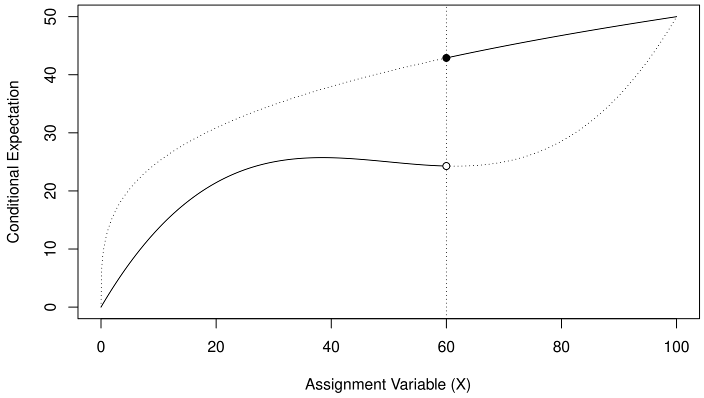


Figure 3.2: Conditional expectation of observed and potential outcomes in SRD (The upeer curve corresponds to $\mu_1(x)$, and the lower curve to $\mu_0(x)$. The solid curves are $\mu(x)$ which is observed and the dotted curves are counterfactuals.)

3.4 Fuzzy Regression Discontinuity Design

The fuzzy RD design is the generalization of the sharp RD design. In contrast to SRD design, the treatment assignment is stochastic in FRD design, i.e., $\Pr(T_u = 1|X_u = x)$ can be in the interval (0,1), not only $\{0,1\}$. As Hahn et al. (2001) noted, the FRD design only requires that

$$\lim_{x \downarrow c} \Pr(T_u = 1 | X_u = x) \neq \lim_{x \uparrow c} \Pr(T_u = 1 | X_u = x). \tag{22}$$

Therefore, the SRD design is just a special case of the FRD design with

$$\lim_{x \downarrow c} \Pr(T_u = 1 | X_u = x) - \lim_{x \uparrow c} \Pr(T_u = 1 | X_u = x) = 1. \tag{23}$$

In many occasions, the treatment status is not fully determined by a value of a certain covariate. Some portion of units with covariate values less than a cutoff may receive a treatment. Similarly, some portion of units with covariate values greater than or equal to a cutoff may not receive a treatment. The probabilistic treatment status can happen for several reasons. There might be some other factors than the covariate that affect the treatment assignment, and imperfect compliance of the rule by participants can also make the treatment status not "sharp" (Lee and Lemieux, 2010). An example of imperfect compliance can be found in Geneletti et al. (2015), which is the case of prescribing a new drug based on a risk score. In such an example, a doctor prescribes a drug after examining a certain risk score. There is a prescription guideline from the authorities, which works exactly like a threshold indicator D. However, it is not always the case that doctors strictly follow the guideline. Especially when patients' risk scores are around a cutoff, doctors may refer to other factors, or they can decide in a subjective way with their previous experiences (Geneletti et al., 2015). On top of that, the stochasticity of the treatment assignment can arise form patients' side. Even though doctors prescribe a drug with a given guideline, some patients may not take the drug for some reasons. These kinds of imperfect compliances make the FRD more common and important.

In order to estimate the local causal effect for the FRD design, we have to handle one subtlety. In SRD design, the size of the discontinuity of $\mu(x)$ at the threshold point was measured to estimate the effect of the treatment. The reason why the same estimator does not estimate the local causal effect for the FRD design, is that there are "violators" on both sides of the threshold. Violators are those which were given the treatment with $X_u < c$, or not given the treatment with $X_u \geq c$. The violators make $\lim_{x \downarrow c} \mathbb{E}[Y_u(1) | X_u = x]$ smaller, and make $\lim_{x \uparrow c} \mathbb{E}[Y_u(0) | X_u = x]$ bigger. In consequence, the SRD parameter $\tau_{\text{SRD}} (= \lim_{x \downarrow c} \mathbb{E}[Y_u | X_u = x] - \lim_{x \uparrow c} \mathbb{E}[Y_u | X_u = x])$ is smaller than the true local causal effect for the FRD. In the FRD design, τ_{SRD} represents another quantity, which is called the intent-to-treat effect (Geneletti et al., 2015). The intent-to-treat effect (τ_{ITT}) is nothing but the causal effect of the "intention" to give the treatment at the threshold, not the causal effect of the actual treatment. The intention here denotes whether $X_u \geq c$ or not, i.e., there is an intention to give the treatment if $X_u \geq c$, and no intention otherwise.

$$\begin{aligned} \tau_{\text{ITT}} &:= \lim_{x \downarrow c} \mathbb{E}[Y_u | X_u = x] - \lim_{x \uparrow c} \mathbb{E}[Y_u | X_u = x], && (24) \\ &= \tau_{\text{SRD}}. && (25) \end{aligned}$$

To retrieve the true local causal effect for the FRD, we have to divide τ_{ITT} by the size of the discontinuity of $\Pr(T_u = 1 | X_u = x)$ $(= \mathbb{E}[T_u | X_u = x])$ at $x = c$ (Lee and Lemieux, 2010; Imbens and Lemieux, 2008; Geneletti et al., 2015). Therefore, the local causal effect for the FRD τ_{FRD} is written as follows.

$$\tau_{\text{FRD}} := \frac{\lim_{x \downarrow c} \mathbb{E}[Y_u | X_u = x] - \lim_{x \uparrow c} \mathbb{E}[Y_u | X_u = x]}{\lim_{x \downarrow c} \mathbb{E}[T_u | X_u = x] - \lim_{x \uparrow c} \mathbb{E}[T_u | X_u = x]}. \tag{26}$$

The scale-up by the inverse of the discontinuity in the treatment assignment probability $\Pr(T_u = 1 | X_u = x)$ is explained in Lee and Lemieux (2010), but we give a simpler derivation as the following. Let us model the FRD design as

$$Y_u = \tau_{\text{FRD}} \cdot T_u + f(X_u) + \epsilon_Y, \tag{27}$$

$$\mathbb{E}[T_u|X_u] = \delta_{\text{FRD}} \cdot D_u + g(X_u), \tag{28}$$

$$D_u = \mathbb{1}\{X_u \geq c\}, \tag{29}$$

where τ_{FRD} is the local treatment effect at $x = c$, δ_{FRD} is the size of the discontinuity of the treatment assignment probability $\Pr(T_u = 1|X_u = x)(= \mathbb{E}[T_u|X_u = x])$ at $x = c$, and $\mathbb{E}[\epsilon_Y] = 0$ with ϵ_Y being independent of X_u. We need a further assumption that f and g are continuous at $x = c$ for τ_{FRD} to be identified as the local treatment effect at $x = c$. If we look at the discontinuity of $\mu(x)$ at $x = c$, then we have

$$\lim_{x \downarrow c} \mathbb{E}[Y_u|X_u = x] - \lim_{x \uparrow c} \mathbb{E}[Y_u|X_u = x] = \tau_{\text{FRD}} \cdot \big\{ \lim_{x \downarrow c} \mathbb{E}[T_u|X_u = x] - \lim_{x \uparrow c} \mathbb{E}[T_u|X_u = x] \big\}. \tag{30}$$

Therefore, we get (26), or

$$\tau_{\text{FRD}} = \frac{\tau_{\text{ITT}}}{\delta_{\text{FRD}}}. \tag{31}$$

As before, exemplary figures of FRD are given below (See Appendix B for R codes). The figure 3.3 is a treatment assignment probability of FRD with a cutoff again at $x = 60$. In contrast to SRD design, the size of the discontinuity of the treatment assignment probability can be less than 1. The figure 3.4 shows three different types of conditional expectation functions. The upper dotted curve represents $\mu_1(x)$, the lower dotted curve $\mu_0(x)$, and the middle solid curve $\mu(x)$ which is observed. Because of the relation 17 and the fact that a treatment assignment probability of FRD design can be strictly in between 0 and 1, $\mu(x)$ is, in general, lie in between $\mu_1(x)$ and $\mu_0(x)$. The local causal effect of FRD τ_{FRD} can be retrieved by dividing the discontinuity of the figure 3.4 by the discontinuity of the figure 3.3.

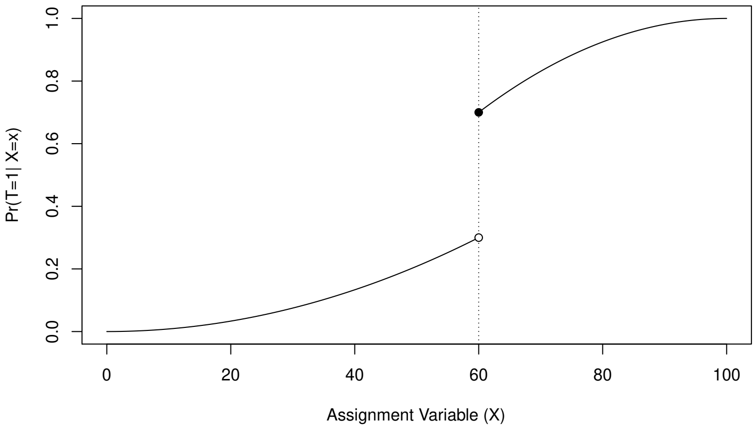


Figure 3.3: Treatment assignment probability in FRD

3.5 Assumptions of Regression Discontinuity Design

There are a number of assumptions to be satisfied for the RD design to successfully estimate the local causal effect, and these assumptions are described with slight differences

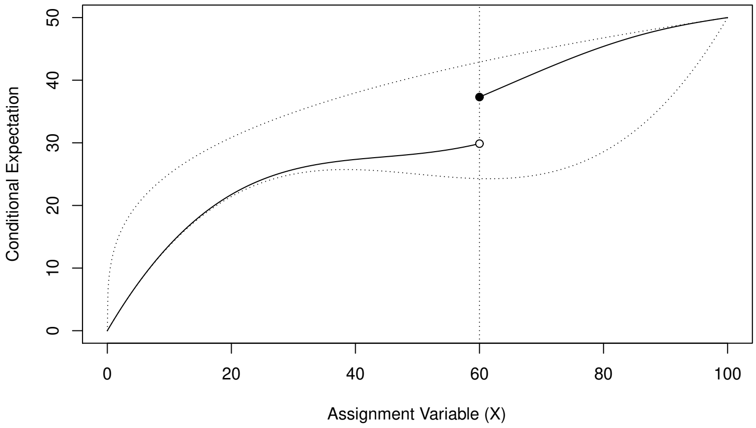


Figure 3.4: Conditional expectation of observed and potential outcomes in FRD (The upper dotted curve corresponds to $\mu_1(x)$, and the lower dotted curve to $\mu_0(x)$. The solid curves are $\mu(x)$ which is observed.)

depending on the fields (Geneletti et al., 2015). In essence, however, there are only a little differences among those assumptions, and can be written in more universal languages. For that purpose, we give two applicability criteria which decides whether the RD design is applicable or not for a given problem. Then, four assumptions for the internal validity of the design are given. These assumptions are based on the assumptions given in Geneletti et al. (2015), Hahn et al. (2001), and Lee and Lemieux (2010).

RDD Criterion 1. (Association of X and T)
Treatment assignment variable T is fully or partly determined by a covariate X, or

$$X \not\perp\!\!\!\perp T. \tag{32}$$

RDD Criterion 2. (Discontinuity)

- $\mathbb{E}[T|X = x]$ is right-continuous at $x = c$.
- $\lim_{x\uparrow c} \mathbb{E}[T|X = x]$ exists, and $\lim_{x\downarrow c} \mathbb{E}[T|X = x] \neq \lim_{x\uparrow c} \mathbb{E}[T|X = x]$.

These two criteria rule out some problems where the RD design cannot be applied. The RDD criterion 1 must be met because it is the underlying structure of the design. The RDD criterion 2 is crucial in that it allows us to identify the local causal effect. If the RDD criterion 2 is not satisfied, then we can still try to apply the RD design, but there is no point in using the design. Suppose there is a positive local causal effect τ_{FRD}, which is expressed as in (26). If we let the denominator of τ_{FRD} approach to zero, then the numerator approaches to zero as well. This means that, in case where $\lim_{x\downarrow c} \mathbb{E}[T|X = x] = \lim_{x\uparrow c} \mathbb{E}[T|X = x]$, we cannot observe the discontinuity of $\mu(x)$, hence no use of the design. In this perspective, the term "discontinuity" in the RD design should be accounted for the discontinuity of $\mathbb{E}[T|X = x]$ at $x = c$.

Once the two RDD criteria are satisfied, then the RD design is a good option to deploy for estimating the local causal effect. The following are four assumptions to guarantee the goodness of the estimator from the RD design.

Assumption 1. (Unconfoundedness)
For some small $\epsilon, \delta > 0$ and a given metric d[a], we have

$$d(C_u, C_v) < \epsilon, \text{ if } |X_u - X_v| < \delta.$$

u and v are units, and C_w is the vector of confounders of an unit w.

[a] A natural choice for d would be the Euclidean norm, or the L^2-norm.

The first assumption is the unconfoundedness among units with similar covariate values. In Lee and Lemieux (2010), the unconfoundedness is described as the continuity of the conditional expectation of confounders given the covariate, but essentially there is only a little difference between the two. In short, both of them say that if units have similar covariate values, then their confounders are also similar, hence comparable to each other. This is particularly useful when the covariate value is close to a cutoff. The assumption guarantees that the confounders are similar near the cutoff so that the only noteworthy difference among units is whether they received the treatment or not. As a consequence, if we consider units near the cutoff only, then we are locally reproducing a randomized experiment[6]. The unconfoundedness assumption, thus, makes the RD design highly appealing than many other forms of observational studies.

One thing to note is that, along with the continuity assumption which will follow, the unconfoundedness assumption is the key concept that allows us to interpret the discontinuity of the potential outcomes as the local causal effect (in SRD case, and the intent-to-treat effect in FRD case). In the sharp RD design, what we want to know is (20), i.e., $\mathbb{E}[Y_u(1)|X_u = c] - \mathbb{E}[Y_u(0)|X_u = c]$. This quantity can be interpreted as the effect of a treatment only if the treatment status is the sole difference between the treatment group and the control group. If there is a systematic change, for example, the treated units have high values of a certain attribute, but the controlled units have low values of that attribute, then $\mathbb{E}[Y_u(1)|X_u = c] - \mathbb{E}[Y_u(0)|X_u = c]$ contains the effect of the attribute as well. Therefore, the unconfoundedness assumption excludes such a situation, and allows us to estimate the "pure" local causal effect.

Assumption 2. (Unit invariant distribution)

- For all x, $\Pr(T_u = 1|X_u = x)$ is identical for all units u.
- For a given x, $Y_u, Y_u(1), Y_u(0)$ are independent and identically distributed with respect to units, given $X_u = x$.

The second assumption is about the distributions which are invariant over units. The assumption is needed mainly for the technical reasons. If the distributions are different for each unit, then the local causal effects must be defined for each unit as well, which makes the problem more complicated. Instead, one regression function will represent $\mu(x)$ for all units. Furthermore, unit invariant distribution assumption is compatible with the unconfoundedness assumption and can be viewed as a corollary of it. If we are given units with the same covariate value, then their confounders are very similar according to

[6] In connection with the Chapter 4, if we choose the constant function function for the regression function, then we are precisely reproducing a local random experiment.

the unconfoundedness assumption. Then, it is natural to deduce that outcomes of units with the same covariate value and similar confounders have similar distributions, which is indeed slightly weak version of the unit invariant distribution assumption.

Assumption 3. (Continuity)

$$\mu_0(x) = \mathbb{E}[Y(0)|X = x] \text{ and } \mu_1(x) = \mathbb{E}[Y(1)|X = x] \text{ are continuous in } x. \tag{33}$$

The third assumption is the continuity of the conditional expectations of potential outcomes. Actually, the continuity at a single point ($x = c$) is enough (Hahn et al., 2001), but we will assume the continuity in a whole domain for the regression analysis. This allows us to identify the discontinuity of $\mu(x)$ at $x = c$ as the local causal effect (Imbens and Lemieux, 2008), or intent-to-treat effect. If the conditional expectations $\mu_0(x)$ and $\mu_1(x)$ are discontinuous at $x = c$, then the discontinuity of $\mu(x)$ at the cutoff is confounded by the discontinuities of $\mu_0(x)$ and $\mu_1(x)$. Therefore, the continuity assumption is one of the most important building blocks of the design.

Assumption 4. (Optional) (Monotonicity)

$$\mathbb{E}[T|X = x] \text{ is increasing at } x = c. \tag{34}$$

The last assumption is the monotonicity of the treatment assignment probability at a cutoff. However, we note that the assumption is optional. The estimate of the design is not harmed even without the monotonicity. One issue without the monotonicity is that the local causal effect can be negative. The negative value of the local causal effect is not a problem at all in any sense even when the true average causal effect is positive, because large positive causal effects elsewhere can compensate the negative causal effect at the cutoff.

Notwithstanding the unnecessary nature, we add the monotonicity here for the coherence of our discussion. Throughout the paper, the average causal effect of a treatment is implicitly assumed to be non-negative. In case where we expect the causal effect to be negative, we can simply multiply -1 to all observed outcomes to get the non-negative causal effect. Therefore, without loss of generality, we may assume that the local causal effect is non-negative. If we assume the monotonicity of the treatment probability assignment at the cutoff, we can keep the coherence of the non-negativity of the local causal effect (26)[7].

Unfortunately, most of the assumptions above are not fully testable. One reason is that they are based on things that are not observable, or that cannot be tested fundamentally. The unconfoundedness assumption relies partly on unobservable confounders, and there are a lot of factors that affect the conditional expectations of potential outcomes, which makes the continuity assumption untestable. Additionally, we only have one response for each unit, but we need more responses to approximate the distributions of $Y_u, Y_u(0), Y_u(1)$ to show the unit invariance distribution assumption.

Some of them may be partly justified. The RDD criterion 1 can be shown directly by the data provider, or sometimes by the discovery of the hidden assignment variable and checking up the high sample correlations between the assignment variable and the treatment indicator ($\hat{\rho}_{\text{X, T}}$), or between the threshold indicator and the treatment indicator ($\hat{\rho}_{\text{D, T}}$). The RDD criterion 2 and the monotonicity assumption can partly be supported

[7] The numerator of (26) can be negative, but then it is likely to be a sign of zero effect.

with using non-parametric regression[8] and observing a big jump at the cutoff. Lastly, the unconfoundedness assumption may not be rejected by considering only the observable confounders. The unconfoundedness among observable confounders is a clearly a partial clue to the complete unconfoundedness. Therefore, we may keep the assumption if we can be certain that unconfoundedness holds with the observable confounders.

[8] Originally, Berk (2010) mentions the non-parametric regression as one option to estimate τ_{SRD} in the SRD design, but we can use it on $\Pr(T = 1|X = x)$ as well.

Chapter 4

Estimation and Inference in the Regression Discontinuity Design

4.1 Estimation

4.1.1 Estimation with Bandwidth

The major process of the RD design is the estimation of the local causal effect. There are two approaches in the estimation: parametric and non-parametric regression (Jacob, Zhu, Somers, and Bloom, 2012). Parametric regression, here, means to determine a functional form first and then fit the whole data to get the coefficients. Any functional form can be chosen, but most commonly polynomial functions or linear function is selected. Non-parametric regression, however, does not assume any specific functional form, but use only the portion of data near the threshold. Any non-parametric method such as k-nearest neighbors method, or local linear regression can be used. Both approaches will fit two different functions; one for the left side and the other for the right side of the threshold. Then, the estimator for the local causal effect (in SRD case) will be defined as the difference of the function values at a threshold c (see (21)).

As mentioned in Jacob et al. (2012), there is only a minor difference between the two approaches and it is the matter of one's perspective, even though the asymptotic behavior of the non-parametric method is better. Throughout this paper, we will focus on the "locally" parametric regression which fixes the functional form at first and then fit the data that are only within some neighborhood of the threshold. In order to define the neighborhood, we introduce the quantity called the bandwidth, or the window, w. Given a window w, local data within the window w ($\mathcal{D}_w$) that we are going to fit are those with a covariate X in $[c-w, c+w]$,

$$\mathcal{D}_w := \{u \in U | X_u \in [c-w, c+w]\}. \tag{35}$$

Since we are going to fit the data separately for the left and right side of the threshold, we define two data sets $\mathcal{D}_{w-}$ and $\mathcal{D}_{w+}$ for the left and right side respectively.

$$\mathcal{D}_{w-} := \{u \in U | X_u \in [c-w, c)\}, \tag{36}$$
$$\mathcal{D}_{w+} := \{u \in U | X_u \in [c, c+w]\}. \tag{37}$$

There are many options for the functional forms for the regression. Polynomial functions are commonly used and we will particularly focus on the linear function. We are interested only in small neighborhood of a threshold, and, in general, linear functions are locally good approximations. Then, for a given bandwidth w, an estimator for the local causal effect τ_{SRD}[1] is defined as

$$\hat{\tau}_{\text{SRD}} = \hat{\alpha}_{w+} - \hat{\alpha}_{w-}, \tag{38}$$

where

$$(\hat{\alpha}_{w+}, \hat{m}_{w+}) = \arg\min_{\alpha,m} \sum_{u \in \mathcal{D}_{w+}} (Y_u - \alpha - m(X_u - c))^2, \tag{39}$$

$$(\hat{\alpha}_{w-}, \hat{m}_{w-}) = \arg\min_{\alpha,m} \sum_{u \in \mathcal{D}_{w-}} (Y_u - \alpha - m(X_u - c))^2. \tag{40}$$

The following figure shows the exemplary implementation of the locally parametric estimation for a given bandwidth w. Units that received a treatment were described in circles, and those that did not receive a treatment were described in crosses. c is a threshold, and $\mathcal{D}_{w-}, \mathcal{D}_{w+}, \hat{\alpha}_{w-}, \hat{\alpha}_{w+}$ are as defined above. For the details of the simulation, refer Appendix B.

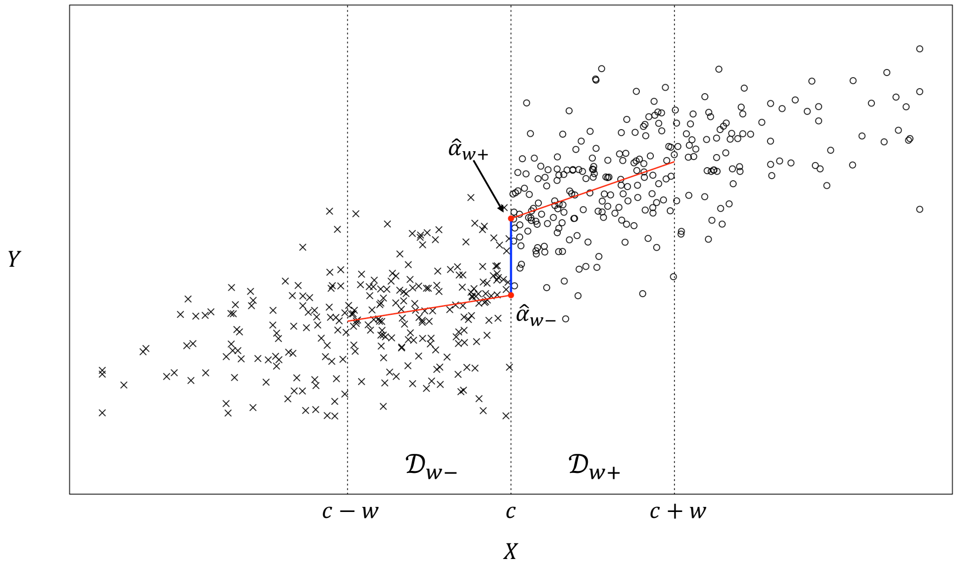


Figure 4.1: An example of locally parametric estimation with simulated data.

4.1.2 Bandwidth Selection

It is still an open problem in the RD design to find an "optimal" bandwidth which makes the unconfoundedness assumption valid and yields an accurate estimator of the local causal effect (Geneletti et al., 2019). In the following, we introduce two existing methods that are often used in practice to find a bandwidth, and propose a new one. The results of the methods will be compared using simulated data with a known optimal bandwidth.

[1] Throughout this chapter, the sharp RD design is assumed for the discussion. However, some of the concepts can also be understood in the fuzzy RD design with proper counterparts, for example, τ_{ITT} for τ_{SRD}.

Cross Validation Approach

There are several methods used by econometricians to choose a bandwidth w. Imbens and Lemieux (2008) and Lee and Lemieux (2010) introduces some of them, and the one we introduce here is the cross validation approach developed by Ludwig and Miller (2005) and Imbens and Lemieux (2008), which exploits the idea of local linear regression. Simply put, this approach finds the bandwidth w that minimizes the loss function, $\mathcal{L}_{\text{CV}}(w)$, defined as

$$\mathcal{L}_{\text{CV}}(w) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_{u_i} - \hat{\mu}_w(X_{u_i})\right)^2, \tag{41}$$

where

$$\hat{\mu}_w(x) = \begin{cases} \hat{\mu}_{w-}(x) & \text{if } x < c, \\ \hat{\mu}_{w+}(x) & \text{if } x \geq c, \end{cases} \tag{42}$$

with

$$(\hat{\mu}_{w-}(x), \hat{\nu}_{w-}(x)) = \arg\min_{\mu,\nu} \sum_{i=1}^{N} \mathbb{1}\{X_{u_i} < x\}(Y_{u_i} - \mu - \nu(X_{u_i} - x))^2 K\left(\frac{X_{u_i} - x}{w}\right), \tag{43}$$

and

$$(\hat{\mu}_{w+}(x), \hat{\nu}_{w+}(x)) = \arg\min_{\mu,\nu} \sum_{i=1}^{N} \mathbb{1}\{X_{u_i} > x\}(Y_{u_i} - \mu - \nu(X_{u_i} - x))^2 K\left(\frac{X_{u_i} - x}{w}\right). \tag{44}$$

The local causal effect τ_{SRD} is then estimated as

$$\hat{\tau}_{\text{SRD}} = \hat{\mu}_{w^*+}(c) - \hat{\mu}_{w^*-}(c), \tag{45}$$

where

$$w^* = \arg\min_{w} \mathcal{L}_{\text{CV}}(w). \tag{46}$$

Here, $K(\cdot)$ is a kernel function on $[-1, 1]$ which assigns a weight for each data point. Since we are going to use the uniform kernel[2], they can be written as

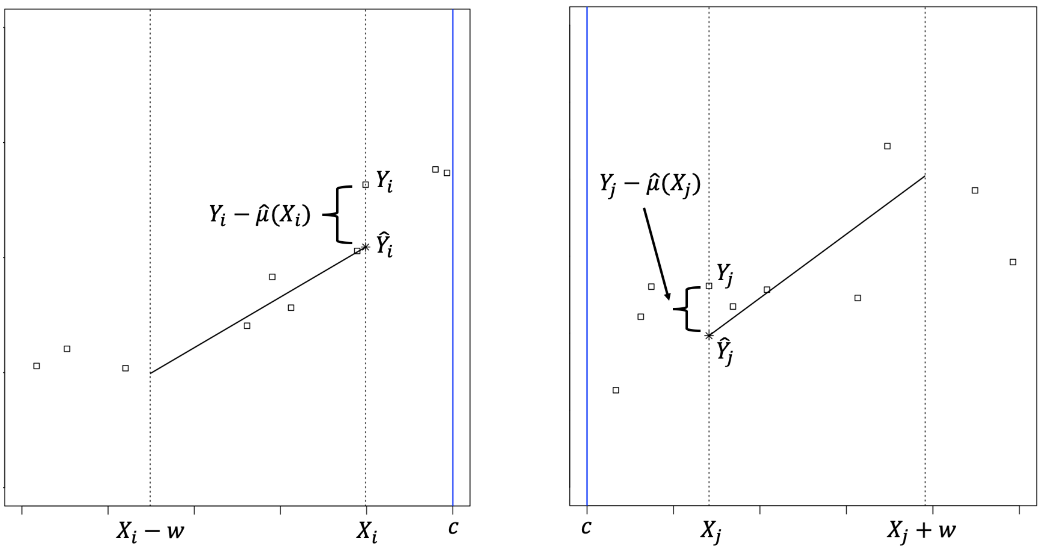
$$(\hat{\mu}_{w-}(x), \hat{\nu}_{w-}(x)) = \arg\min_{\mu,\nu} \sum_{i=1}^{N} \mathbb{1}\{x - w \leq X_{u_i} < x\}(Y_{u_i} - \mu - \nu(X_{u_i} - x))^2, \tag{47}$$

$$(\hat{\mu}_{w+}(x), \hat{\nu}_{w+}(x)) = \arg\min_{\mu,\nu} \sum_{i=1}^{N} \mathbb{1}\{x < X_{u_i} \leq x + w\}(Y_{u_i} - \mu - \nu(X_{u_i} - x))^2. \tag{48}$$

If we note the strict inequalities and the fact that x will be replaced with covariates X_{u_i}, we can interpret this method as a variant of leave-one-out cross validation (Lee and Lemieux, 2010). This will be more clear with the exemplary illustrations of Figure 4.2.

The purpose of this approach explained in Appendix B of Ludwig and Miller (2005) is to find a bandwidth that yields best boundary estimation. However, there are some flaws in this approach. First of all, what we need is a bandwidth that works good at a single point, i.e., at a cutoff, but this approach measures the performance of a bandwidth

[2] According to Lee and Lemieux (2010), Fan and Gijbels (2003) has shown that the triangular kernel is an optimal kernel. However, the difference in result is minor, and for the sake of interpretability, the uniform kernel is preferred (Lee and Lemieux, 2010).



(a) Cross Validation Loss Term below a threshold (b) Cross Validation Loss Term above a threshold

Figure 4.2: Cross Validation Loss Terms

over all data points. This might seem reasonable for data near a cutoff, but the good performance of boundary estimation with data far away from a threshold may not lead us to a bandwidth we want[3]. Secondly, the bandwidth Ludwig and Miller (2005) tries to find is a bandwidth that works well for the local linear regression. Therefore, the method is not fully exploiting the core assumption of the RD design: the unconfoundedness assumption. Lastly, the computational complexity of the method can be high, because the procedure contains fitting a regression line for each data point.

Alternative Bandwidth Selection Method

The previous approach is often used in practice to determine the bandwidth, however, there are still some shortcomings as mentioned above. The previous cross validation approach measures the performance of a bandwidth over all data points, even though we are interested only at a single point, and it does not fully appreciate the unconfoundedness near a threshold. Technically, it tries to find a bandwidth for the local linear regression with little context of the unconfoundedness assumption. To fill this contextual gap, we propose an alternative way of determining a bandwidth.

There are two issues to be considered when determining a bandwidth w. The first

[3] To solve this issue, Imbens and Lemieux (2008) suggests ignoring some portions of data which are far away from a threshold. Therefore, the cross validation loss term does not add all the errors but smaller amount, i.e.,

$$\mathcal{L}^{\delta}_{\text{CV}}(w) = \frac{1}{N_\delta} \sum_{u: q_{\delta -} \le X_u \le q_{(1-\delta)+}} (Y_u - \hat{\mu}(X_u))^2, \tag{49}$$

where $q_{\delta-}$ and $q_{(1-\delta)+}$ are the δ and $(1-\delta)$-quantiles of X on the left and right side of a threshold, and N_δ is the number of units within those quantiles. Imbens and Lemieux (2008) suggests $\delta = 0.5$ in practice.

issue is the volatility of an estimator. If the choice of a bandwidth is too small so that there are only little data inside, then an estimator for local causal effect tends to have a high variance which tends to decrease with larger bandwidths. This tendency can be observed with the linear regression analysis. Recall that an estimator for the local causal effect is defined to be $\hat{\tau}_{\text{SRD}} = \hat{\alpha}_{w+} - \hat{\alpha}_{w-}$ as in (38), where $\hat{\alpha}_{w+}$ and $\hat{\alpha}_{w-}$ are intercepts of each corresponding fitted linear functions. With the fundamental assumptions of the linear regression and the normality assumption of errors (Groß, 2003), we can describe the distributions of $\hat{\alpha}_{w+}$ and $\hat{\alpha}_{w-}$, hence the distribution of $\hat{\tau}_{\text{SRD}}$. If we express (39) and (40) as matrix forms, then they are written as

$$\begin{pmatrix} \hat{\alpha}_{w-}(x) \\ \hat{m}_{w-}(x) \end{pmatrix} = \arg\min_{\beta} \|Y_{w-} - X_{c-}\beta\|^2, \text{ and } \begin{pmatrix} \hat{\alpha}_{w+}(x) \\ \hat{m}_{w+}(x) \end{pmatrix} = \arg\min_{\beta} \|Y_{w+} - X_{c+}\beta\|^2, \tag{50}$$

where Y_{w-} is the $|\mathcal{D}_{w-}|$-dimensional vector with entries $\{Y_u, u \in \mathcal{D}_{w-}\}$, X_{c-} is the $|\mathcal{D}_{w-}| \times 2$ matrix with rows $\{(1, X_u - c) : u \in \mathcal{D}_{w-}\}$, and similarly for Y_{w+} and X_{c+} with $\mathcal{D}_{w-}$ being replaced by $\mathcal{D}_{w+}$. Then, by the Theorem 2.7 in Groß (2003), we get

$$\begin{pmatrix} \hat{\alpha}_{w-}(x) \\ \hat{m}_{w-}(x) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha_{w-}(x) \\ m_{w-}(x) \end{pmatrix}, \sigma^2 (X_{c-}^T X_{c-})^{-1} \right), \tag{51}$$

and

$$\begin{pmatrix} \hat{\alpha}_{w+}(x) \\ \hat{m}_{w+}(x) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha_{w+}(x) \\ m_{w+}(x) \end{pmatrix}, \sigma^2 (X_{c+}^T X_{c+})^{-1} \right). \tag{52}$$

Here, σ^2 is the true variance of the error of outcomes, and α's and m's without hat are the true coefficients if the true functional form is linear. In particular, we note that

$$\text{Var}(\hat{\alpha}_{w-}) = \sigma^2 (X_{c-}^T X_{c-})^{-1}_{11} = \frac{\sigma^2}{|\mathcal{D}_{w-}|} \cdot \frac{\frac{\sum_{u \in \mathcal{D}_{w-}} (X_u)^2}{|\mathcal{D}_{w-}|}}{\frac{\sum_{u \in \mathcal{D}_{w-}} (X_u)^2}{|\mathcal{D}_{w-}|} - \left(\frac{\sum_{u \in \mathcal{D}_{w-}} X_u}{|\mathcal{D}_{w-}|} \right)^2}, \tag{53}$$

and

$$\text{Var}(\hat{\alpha}_{w+}) = \sigma^2 (X_{c+}^T X_{c+})^{-1}_{11} = \frac{\sigma^2}{|\mathcal{D}_{w+}|} \cdot \frac{\frac{\sum_{u \in \mathcal{D}_{w+}} (X_u)^2}{|\mathcal{D}_{w+}|}}{\frac{\sum_{u \in \mathcal{D}_{w+}} (X_u)^2}{|\mathcal{D}_{w+}|} - \left(\frac{\sum_{u \in \mathcal{D}_{w+}} X_u}{|\mathcal{D}_{w+}|} \right)^2}. \tag{54}$$

If we assume the independence of the two estimators, the variance of $\hat{\tau}_{\text{SRD}}$ is the sum of the variances of $\hat{\alpha}_{w+}$, and $\hat{\alpha}_{w-}$, i.e. $\text{Var}(\hat{\tau}_{\text{SRD}}) = \text{Var}(\hat{\alpha}_{w+}) + \text{Var}(\hat{\alpha}_{w-})$. Therefore, $\text{Var}(\hat{\tau}_{\text{SRD}})$ tend to decrease as a bandwidth gets larger by the law of large numbers.

The second issue is the confounding. We cannot simply keep increasing a bandwidth to get an estimator with small variance. If the bandwidth is too large, the data that are far away from a threshold affects the estimation of the local causal effect at a threshold. The problem of those data are that they may have significantly different confounders compared to those near the threshold. If these data are included to fit a regression line, the estimator of the local causal effect can be highly biased. For the sake of the unconfoundedness assumption in the chapter 3, we need a small bandwidth.

With the two issues in mind, we propose an alternative bandwidth selection method with the following loss function,

$$\mathcal{L}_{\text{ALT}}^{\lambda}(w) = \lambda \cdot \frac{w}{(X_{\max} - X_{\min})} + \hat{\sigma}_{\text{SRD}}^2, \tag{55}$$

where λ is the tuning parameter, $\hat{\sigma}^2_{\text{SRD}}$ is the sum of estimated variances of $\hat{\alpha}_{w-}$ and $\hat{\alpha}_{w+}$, and $X_{\max}$ and $X_{\min}$ are upper and lower bounds of a covariate X respectively. The reason we divide w by $X_{\max} - X_{\min}$ is because we want it to be free from the scale of the covariate X.

Intuitively, the alternative loss term makes a bandwidth large enough so that $\hat{\sigma}^2_{\text{SRD}}$ is not large, but small enough to retain the unconfoundedness. We can interpret this as a bias-variance trade-off of an estimated local effect. With a small bandwidth, the bias of the estimated effect is small but the variance is large. On the other hand, with a large bandwidth, the bias becomes large, but the variance is small.

One issue to be considered with this alternative approach is about determining the value of λ in the loss (55). Indeed, if λ is too large, then the loss penalties too much on the large bandwidth so that a bandwidth selected will be very small. On the other hand, if λ is too small, then the penalty on the bias term is negligible so that the bandwidth gets large only to reduce $\hat{\sigma}^2_{\text{SRD}}$. One rule of thumb is to rely on domain knowledge, or prior knowledge on data, to set a maximum bandwidth $w_{\max}$, and adjust λ not too large such that this method does not choose a bandwidth exceeding $w_{\max}$.

Asymptotically Optimal Bandwidth Selection

When there are large samples, there is a better way of determining the bandwidth. Imbens and Kalyanaraman (2011) developed the idea of Ludwig and Miller (2005) and derived a bandwidth selection algorithm which depends solely on the data. A bandwidth from the algorithm is asymptotically optimal under squared loss and optimal with respect to Li (1987) (Imbens and Kalyanaraman, 2011). In a nutshell, Imbens and Kalyanaraman (2011) finds the bandwidth that minimizes what they call "asymptotic mean squared error" or AMSE, which is a first order approximation of MSE with $\text{MSE}(h) = \mathbb{E}[(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})^2]$. This can also be extended to the fuzzy RD design.

Imbens and Kalyanaraman's algorithm has been favored over the cross validation approach and often been applied after its publication. However, one should note that the algorithm is optimal with large data, and requires some technical assumptions and proper setting. The authors made the source code public in a personal webpage[4].

Bandwidth Selection Results with Simulated Data

To see how well each bandwidth selection works given the presence of a confounder, we compare the results of each bandwidth selection process with simulated data. Three processes, the cross validation approach, the quantile cross validation approach (49) with $\delta = 0.5$, and the newly proposed method, were compared. For the simulation, we introduce a single hypothetical confounder H which also affects the outcome Y. The confounder H will be set to be almost constant around a threshold with a known bandwidth so that the unconfoundedness assumption is valid within this known bandwidth. Consequently, a good bandwidth selection algorithm should return a bandwidth close to the known bandwidth.

The λ in the proposed alternative selection was set to be 5.5[5]. 5000 data sets were generated, and each data set contained 200 samples with $X_{\min} = 0$, $X_{\max} = 10$, and the

[4] MATLAB source code of the algorithm can be found on `https://imbens.people.stanford.edu/software`.

[5] The value of λ was selected in such a way that the alternative bandwidth selection method returns a bandwidth close to the known bandwidth with a single data set.

cutoff $c = 5$. In each data set, bandwidths[6] ranging from 0.2 to 1 with interval 0.05 were tested with each bandwidth selection method. The effect of X on Y is modeled as

$$\mu(x) = \begin{cases} 0.4x + 1 & \text{if } x < 5, \\ 0.2x + 3 & \text{if } x \geq 5, \end{cases} \tag{56}$$

and Y, combined with the effect of H, is written as

$$Y = \mu(X) + H + \epsilon_Y, \tag{57}$$

where ϵ_Y is normally distributed with mean 0, and variance 1. The distribution of the confounder H is given as

$$H = \begin{cases} \frac{8}{9}X + \epsilon_H & \text{if } X \in [0, 4.5), \\ 4 + \epsilon_H & \text{if } X \in [4.5, 5.5], \\ \frac{8}{9}(X - 5.5) + 4 + \epsilon_H & \text{if } X \in (5.5, 10], \end{cases} \tag{58}$$

where ϵ_H is a gaussian error with mean 0, and standard deviation 0.25.

With the given distribution of the confounder H, the unconfoundedness assumption is valid around the threshold with bandwidth 0.5. Note that any bandwidth less than or equal to 0.5 makes the unconfoundedness assumption still valid. Therefore, we can say that a certain bandwidth selection performs well if it returns the bandwidth of size close to 0.5, or less.

Figure 4.3 shows the distribution of 5000 results (See Appendix A for tabular data). It is obvious that the newly proposed method showed better results. With 49.66% of the data sets, the new method yielded bandwidths of size 0.5 or less. On the other hand, it was only 11.98% for the cross validation approach, and 14.56% for the quantile cross validation approach. Furthermore, the new method tended to select bandwidths less and less which are bigger than 0.6, and only 9 selections for bandwidth 1, but the other two methods tended to select more and more as bandwidth increases. The cross validation approach selected bandwidth 1 for 41.66% of the total data sets, and the quantile version selected bandwidth 1 for 32.08%. The quantile cross validation method was slightly better than the normal cross validation method, and the alternative method we proposed was far better than the two. Although the way we chose λ is not fair, the result tells us that if we can find a proper λ, then the new method can outperform the existing methods.

Estimation in the Fuzzy Regression Discontinuity

All the procedures explained above were formulated under the sharp RD design, but we can extend it to the fuzzy RD design with ease. Simply we replace $\hat{\tau}_{\text{SRD}}$ with $\hat{\tau}_{\text{ITT}}$, and apply the same procedures again for estimating the discontinuity in $\Pr(T_u = 1|X = x)$. Technically, we can set four different bandwidths for the regressions of $\hat{\tau}_{\text{ITT}}$ and $\Pr(T_u = 1|X = x)$, but it is favored to choose the same bandwidth for all of them (Imbens and Lemieux, 2008). There are two options to choose for the single bandwidth. One is for minimizing $\mathcal{L}_{\text{CV}}$ for τ_{ITT}, and the other is for minimizing $\mathcal{L}_{\text{CV}}$ for δ_{FRD}. Imbens and Lemieux (2008) suggests to choose the smaller one to avoid an asymptotic bias, that is,

$$w_{\text{FRD}} = \min(\arg\min_w \mathcal{L}^Y_{\text{CV}}(w), \arg\min_w \mathcal{L}^D_{\text{CV}}(w)), \tag{59}$$

[6] The lower bound 0.2 was set because there were some data sets that had too few data in $\mathcal{D}_{w-}$ or $\mathcal{D}_{w+}$. The upper bound 1 was set because the neighborhood around a threshold with bandwidth 1 already accounts for 20% of the range $X_{\max} - X_{\min}$, which is thought to be too large.

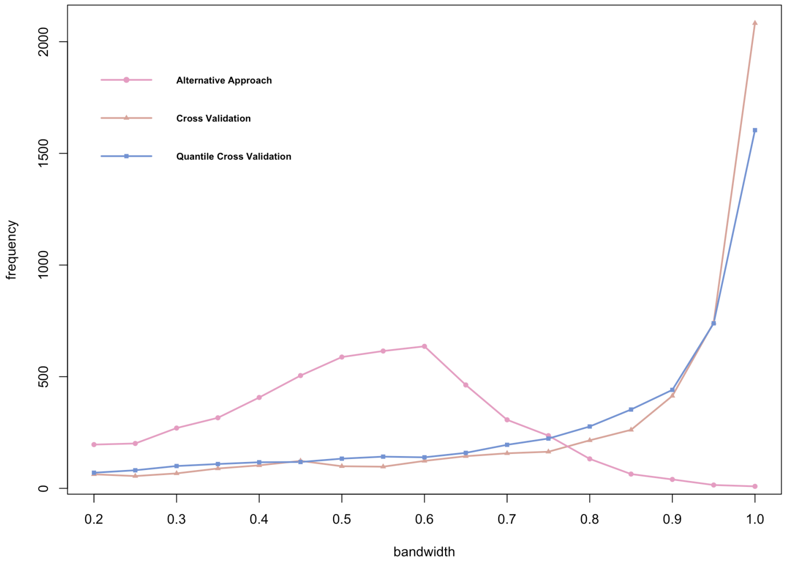


Figure 4.3: Results of Three Bandwidth Selection Methods

where $\mathcal{L}_{\text{CV}}^{Y}$ is the cross validation loss for estimating τ_{ITT}, and $\mathcal{L}_{\text{CV}}^{D}$ is the cross validation loss for estimating δ_{FRD}.

We present an exemplary estimation procedure for the fuzzy RD design. For this example, we assume that a bandwidth is already given and the true model is as follows.

$$\mathbb{E}[T|X] = \begin{cases} \frac{0.3}{c}X, & \text{if } X < c, \\ \frac{0.3}{c}X + 0.4, & \text{if } X \geq c, \end{cases} \tag{60}$$

and

$$\mathbb{E}[Y|X] = \begin{cases} \frac{0.3}{c}X + 0.2, & \text{if } T = 0, \\ \frac{0.3}{c}X + 0.7, & \text{if } T = 1, \end{cases} \tag{61}$$

where c is the threshold and set to be 5 for the simulation. We can easily calculate that δ_{FRD} is 0.4, τ_{FRD} is 0.5, and τ_{ITT} is 0.2 from (31). In the simulation, we set a bandwidth w to be 2, sample size to be 500, and the results are shown in figures 4.4 and 4.5 (refer Appendix B for details). Units with a control treatment are denoted by crosses, and units with a treatment are denoted by circles in both figures. $\hat{\delta}_{\text{FRD}}$ was estimated to be 0.4253, $\hat{\tau}_{\text{ITT}}$ to be 0.2543, and thus $\hat{\tau}_{\text{FRD}}$ to be 0.5981.

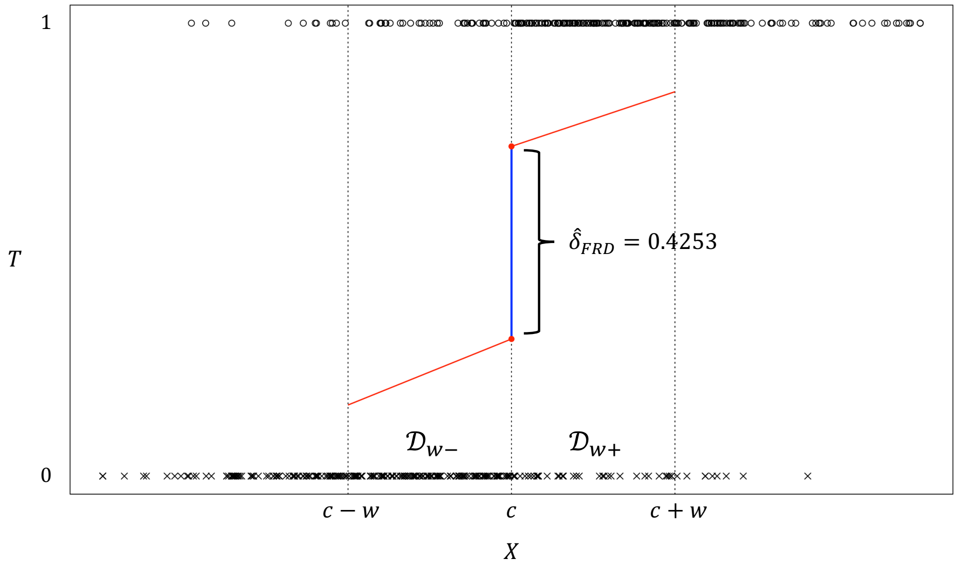


Figure 4.4: Estimation of the Discontinuity of the Treatment Probability

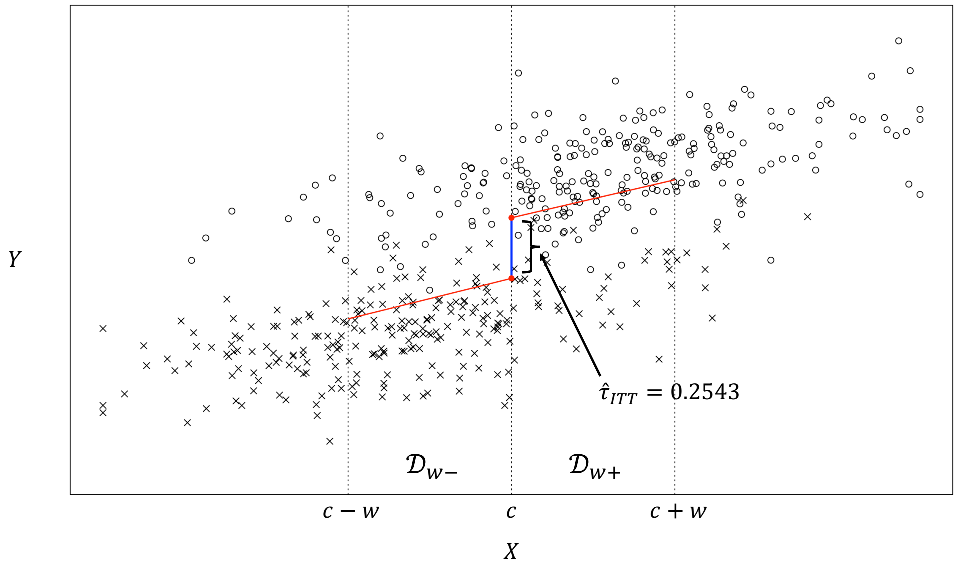


Figure 4.5: Estimation of Intent-To-Treat Effect

4.2 Inference

Inference in the RD design is already hinted in the estimation process, but we give separate and more clear description here. Two important subjects in the inference would be the hypothesis testing and the confidence interval regarding the local causal effect τ_{SRD}.

As mentioned before, we assume the linear model in a small neighborhood of a threshold as

$$Y = \alpha_{w-} + m_{w-}(X - c) + \epsilon_Y \quad \text{on } [c - w, c), \tag{62}$$

$$Y = \alpha_{w+} + m_{w+}(X - c) + \epsilon_Y \quad \text{on } [c, c + w], \tag{63}$$

$$\tau_{\text{SRD}} = \alpha_{w+} - \alpha_{w-}, \tag{64}$$

where ϵ_Y follows a gaussian distribution mean 0 and variance σ_Y^2. These can be simplified, if we introduce the treatment indicator variable T, as

$$Y = \alpha_{w-} + m_{w-}X + (\alpha_{w+} - \alpha_{w-})T + (m_{w+} - m_{w-})X \cdot T + \epsilon_Y \quad \text{on } [c - w, c + w]. \tag{65}$$

One advantage of writing equations in a simpler way is that a single linear regression is needed to have an estimator of τ_{SRD} and the information about it, because what we try to estimate is the coefficient of T. Furthermore, there is no need to assume the independence of two estimators $\hat{\alpha}_{w-}$ and $\hat{\alpha}_{w+}$ as before. Again, with the fundamental assumptions of the linear regression (Groß, 2003), we can describe the distribution of the estimator of $\tau_{\text{SRD}}(= \alpha_{w+} - \alpha_{w-})$ as

$$\hat{\tau}_{\text{SRD}} \sim \mathcal{N}(\tau_{\text{SRD}}, \sigma_Y^2 (X_c^T X_c)^{-1}_{33}), \tag{66}$$

where X_c is $|\mathcal{D}_w| \times 4$ matrix with rows $\{(1, X_u, T_u, X_u T_u) : u \in \mathcal{D}_w\}$. However, the true value of σ_Y^2 is not known, therefore, the normal distribution above is meaningless. Instead, it is known that if we use the estimate of the variance $\hat{\sigma}_Y^2$[7], then we have the t distribution with $n_w - 4$ degrees of freedom (Bühlmann and Mächler, 2016), i.e.,

$$\frac{\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}}{\hat{\sigma}_Y^2 (X_c^T X_c)^{-1}_{33}} \sim \mathcal{T}_{n_w - 4}, \tag{67}$$

where $n_w = |\mathcal{D}_w|$.

Hypothesis Testing

The most important question we want to have an answer is whether a treatment has a positive effect or not. We can have an evidence for the answer using a hypothesis testing. Let us consider a hypothesis test with a significance level α where the null hypothesis is that there is no effect, and the alternative hypothesis is that there is a positive effect[8], i.e.,

$$H_0 : \tau_{\text{SRD}} = 0 \quad \text{vs} \quad H_1 : \tau_{\text{SRD}} > 0. \tag{68}$$

Using the distribution (67), we can calculate the p-value of the test, which is the probability of observing more extreme events than the one observed under the null hypothesis, i.e.,

$$p = \Pr(\hat{\tau}_{\text{SRD}} \geq \hat{\tau}^{\text{obs}}_{\text{SRD}} | H_0), \tag{69}$$

where $\hat{\tau}^{\text{obs}}_{\text{SRD}}$ is an observed value with given samples. Under H_0, we have

$$\frac{\hat{\tau}_{\text{SRD}}}{\hat{\sigma}_Y^2 (X_c^T X_c)^{-1}_{33}} \sim \mathcal{T}_{n_w - 4}, \tag{70}$$

therefore, the p value of the test is

$$p = \Pr(T \geq \frac{\hat{\tau}^{\text{obs}}_{\text{SRD}}}{\hat{\sigma}_Y^2 (X_c^T X_c)^{-1}_{33}}), \tag{71}$$

where T follows the t distribution with $n_w - 4$ degrees of freedom. Finally, we reject the null hypothesis H_0 if the p value is smaller than the given significance level α.

[7] $\hat{\sigma}_Y^2 = \frac{1}{|\mathcal{D}_w| - 4} \sum_{u \in \mathcal{D}_w} (Y_u - \hat{Y}_u)^2$

[8] The reason we do not set the alternative hypothesis to be non-zero effect, i.e., $H_1 : \tau_{\text{SRD}} \neq 0$, is because we are assuming that the effect is non-negative.

Confidence Interval

The confidence interval for τ_{SRD} is given in a similar manner using (67). For example, a $(1-\alpha)$-confidence interval for τ_{SRD} can be constructed from the fact that

$$\Pr(t_{\alpha/2} \leq \frac{\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}}{\hat{\sigma}_Y^2 (X_c^T X_c)_{33}^{-1}} \leq -t_{\alpha/2}) = 1 - \alpha, \tag{72}$$

where $t_{\alpha/2}$ is the $\alpha/2$-quantile of the t distribution with $n_w - 4$ degrees of freedom. Simple manipulations of the equation yields the following $(1-\alpha)$-confidence interval.

$$\text{CI}(1-\alpha) = \big[\hat{\tau}_{\text{SRD}} + t_{\alpha/2} \cdot \hat{\sigma}_Y^2 (X_c^T X_c)_{33}^{-1}, \hat{\tau}_{\text{SRD}} - t_{\alpha/2} \cdot \hat{\sigma}_Y^2 (X_c^T X_c)_{33}^{-1}\big]. \tag{73}$$

Chapter 5

Sensitivity Analysis

In this chapter, we analyze the effects of confounding without the core assumption of the RD design. As listed in the Chapter 3, the unconfoundedness assumption is a powerful and key assumption in the RD design. However, it is not always true and needs to be justified in practice. For the time being, we will assume that the difference of confounders between those that are below and above a threshold is not ignorable in any small neighborhood. Since the unconfoundedness assumption is no more valid, the confounding should be considered. There are many ways how confounding can arise in the RD design model. Among those confoundings, we focus on three different models: XY-confounding, XT-confounding, and TY-confounding models. As names imply, for example, the XY-confounding model is a model that a confounder affects X and Y, and the others are defined similarly.

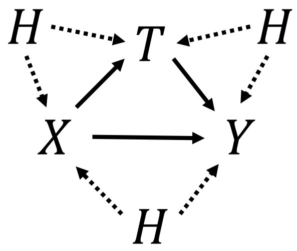


Figure 5.1: Three Confounding Models

For the simplicity of the analysis, we will introduce a single confounder H. There are three possible influences of H, which are on a covariate X, a treatment T, and an outcome Y. Let us denote the strength of each influence as α, β and γ respectively. In principle, we may express each influence of H, if it has, as

$$X = \xi_\alpha(H) + \epsilon_X, \tag{74}$$

$$T = \psi_\beta(X, H) + \epsilon_T, \tag{75}$$

$$Y = \phi_\gamma(X, T, H) + \epsilon_Y, \tag{76}$$

where ϵ_X, ϵ_T, and ϵ_Y are errors. However, for the clarity of the violation of the unconfoundedness assumption and the simplicity of analysis, we do not consider the influence of H on X (i.e., ξ_α), and set

$$H = \begin{cases} h_0 + \epsilon_H, & \text{if } X < c, \\ h_1 + \epsilon_H, & \text{if } X \geq c, \end{cases} \tag{77}$$

with $h_0 \neq h_1$ constant, and ϵ_H being a gaussian error. Then, there exists no small neighborhood of a threshold with similar confounder values. Furthermore, we set

$$\psi_\beta(X, H) = \mathbb{1}\{X + \beta H \geq c\}, \tag{78}$$

$$\phi_\gamma(X, T, H) = \mu(X, T) + \gamma H. \tag{79}$$

Note that the confounded treatment assignment, ψ_β, is “sharp” in this setting.

For the following simulations, the covariate X is generated in the range $[0, 10]$, the threshold $c = 5$, $h_0 = 1$, $h_1 = 3$, and

$$\mu(x, t) = \begin{cases} 0.2x + 3 & \text{if } t = 1, \\ 0.4x + 1 & \text{if } t = 0. \end{cases} \tag{80}$$

With these in mind, we will observe the results of the RD design which is conducted as if the unconfoundedness assumption still holds is some bandwidth, and analyze how the introduction of a confounder H affects the results of the design with different values of β and γ. In each simulation, the alternative bandwidth selection method with $\lambda = 5.5$ was used, and a bandwidth was selected from values ranging from 0.2 to 1 with a gap 0.05.

5.1 XY-Confounding Model

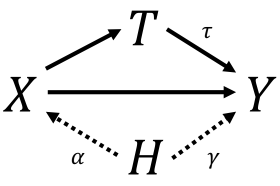


Figure 5.2: XY-Confounding Model

The first model is the XY-confounding model where H influences X and Y. As mentioned earlier, we will fix the relation between X and H, and only examine the results of the design with varying values of γ. From (76) and (79), it is clear to see that

$$Y = \mu(X, T) + \gamma H + \epsilon_Y. \tag{81}$$

It is simply the sharp RD setting when γ is 0, and in this case τ_{SRD} (19) is the same as the true local effect τ of the treatment T, i.e.,

$$\tau_{\text{SRD}} = \mu(c, 1) - \mu(c, 0) = \tau. \tag{82}$$

However, if $\gamma \neq 0$, then τ_{SRD} is not the same as τ, but

$$\begin{aligned}\tau_{\text{SRD}} &= \mu(c,1) - \mu(c,0) + \gamma(h_1 - h_0), && (83)\\ &= \tau + \gamma(h_1 - h_0). && (84)\end{aligned}$$

Therefore, if there is an effect of a confounder, i.e, $\gamma \neq 0$, it is obvious that the estimator of the local causal effect from the sharp RD design will be biased. The degree of the bias, or the confounding will increase as the absolute value of γ increases. Indeed, the square of the bias is

$$(\tau_{\text{SRD}} - \tau)^2 = \gamma^2(h_1 - h_0)^2. \tag{85}$$

With the true values of h_0 and h_1, the square of the bias is a quadratic function of γ, or $(\tau_{\text{SRD}} - \tau)^2 = 4\gamma^2$.

Figure 5.3 is the plot of bias squared with different values of γ ranging from -1 to 1 with a gap 0.1. Each black dot is the actual average value from 100 data sets and the red curve is the fitted function with quadratic model, $y = a\gamma^2$, with $\hat{a} = 4.3747$, which backs up the previous analysis.

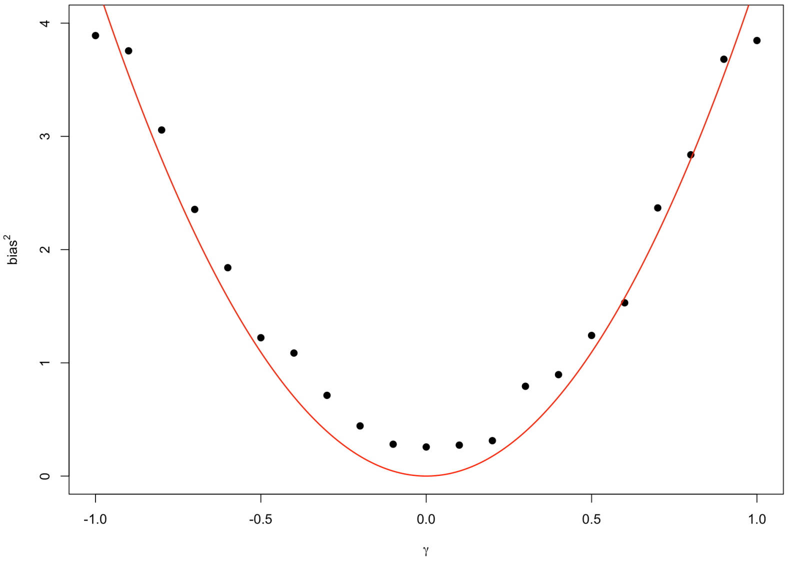


Figure 5.3: γ – Bias squared plot in the XY-confounding model

5.2 XT-Confounding Model

The second model is the XT-confounding model where H now influences on the treatment assignment T. The relationship with X is fixed as before. In this model, the treatment is not solely dependent on X, but also on H with β describing the strength of the dependence.

$$T = \mathbb{1}\{X + \beta H \geq c\}. \tag{86}$$

Let us assume that $\beta \geq 0$. If we ignore the gaussian error ϵ_H on H, then units with $X_u < c$ should satisfy $X_u + \beta h_0 \geq c$ to get the treatment. On the other hand, units with

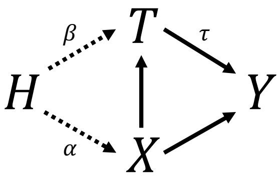


Figure 5.4: XT-Confounding Model

$X_u \geq c$ always get the treatment because βH term is positive. To sum up, the treatment assignment rule is changed with "adjusted" cutoff $c_{\text{adj}}(\beta) = c - \beta h_0$, and

$$T = \mathbb{1}\{X \geq c - \beta h_0\}. \tag{87}$$

The change of the assignment rule can be interpreted as a shift of the cutoff to the left. In case of $\beta < 0$, it can easily be deduced that the cutoff is shifted to the right with $c_{\text{adj}}(\beta) = c - \beta h_1$. In either case, if we naively implement the sharp RD design with the cutoff c, then the estimated local effect is biased.

As the cutoff is shifted with the influence of H on T, the true local effect τ is not the same as before. If the treatment is determined only by whether $X \geq c$ or not, τ is $\mu(c, 1) - \mu(c, 0)$ as (82). However, in this case, it is defined as

$$\tau = \tau(\beta) = \mu(c_{\text{adj}}(\beta), 1) - \mu(c_{\text{adj}}(\beta), 0), \tag{88}$$

which is the same as τ_{SRD} at $x = c_{\text{adj}}(\beta)$. Note that the true local effect depends on the value of β. Now let us denote τ_{SRD} at $x = c_{\text{adj}}(\beta)$ by $\tau_{\text{SRD}}^{\text{adj}}(\beta)$, then the estimator $\hat{\tau}_{\text{SRD}}^{\text{adj}}(\beta)$ is an unbiased estimator of $\tau(\beta)$.

Figure 5.5 is the plot of bias squared, $(\hat{\tau}_{\text{SRD}}^{\text{adj}}(\beta) - \tau(\beta))^2$, with different values of β ranging from 0 to 1 with a gap 0.05. As assumed earlier, the gaussian error of H, ϵ_H, is set to be 0 in the simulation because of the unstable estimation with a positive error. One can proceed with a positive error, and in that case, the fuzzy RD estimation should be considered. Each black dot in the figure is again the average from 100 data sets, and the red solid line ($y = 0.1794$) is the mean value of black dot results. This stable and relatively small average error is an evidence of the fact that cutoff shifts in the XT-confounding model and that $\hat{\tau}_{\text{SRD}}^{\text{adj}}(\beta)$ is indeed an unbiased estimator of $\tau(\beta)$.

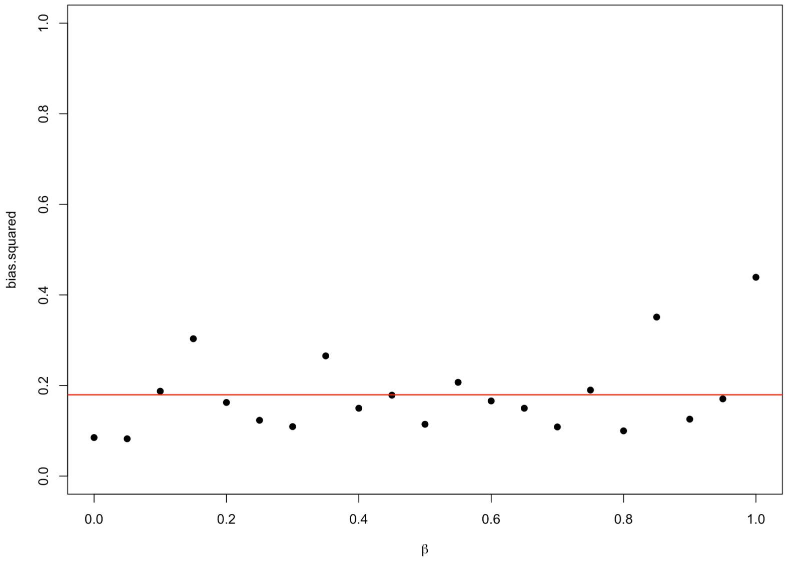


Figure 5.5: β – Bias squared plot in the XT-confounding model

5.3 TY-Confounding Model

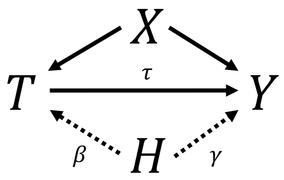


Figure 5.6: TY-Confounding Model

The third model is the TY-confounding where H influences influences both the treatment assignment T and the outcome Y. This model is more complicated than the previous two because we have to consider two different confounding effects of H. As we have already seen, the effect on T shifts the cutoff, and the effect on Y introduces the bias at a discontinuity point of H. The two confounding effects may cancel each other, or may be amplified depending on the combination of a pair (β, γ) and the discontinuity point of H. We skip the simulation since it is just a simple mixture of the previous models and contains no novel aspect.

Before we conclude the chapter, we note that the most realistic model is the XY-confounding model. In many cases, the limitations of the RD design come from the existence of unobservable confounders, which are likely to influence the outcome. On the other hand, the treatment is very unlikely to be affected by the unobservable confounders because, usually, the treatment assignment rule is determined only with observable factors, even if it is stochastic. If the confounding on T is suspected, one should instead try to inspect whether the treatment is determined by more than one variable or not.

Chapter 6

Summary

For the first half of this paper, we covered the basic concepts and subject of the causal inference, and the fundamental idea and structure of the RD design. We defined what the causal effect is, and introduced the Rubin causal model which is the grounds for the discussions that follow. Assuming the Rubin causal model, the randomized experiment is the most preferred design to infer about the causal effect. However, the randomized experiment is not always feasible for different reasons. The observational study, which uses already gathered data, is one option when a random experiment is not feasible. The RD design is one branch of the observational study and has many attractive features that locally resemble the randomized experiment design. Two sub designs, shard RD and fuzzy RD designs, are explained in details, and then formulated the assumptions of the RD design which have been used in slightly different forms in different fields.

The second half of this paper focused on the estimation and inference of the local causal effect, and some results that could arise when the unconfoundedness assumption is violated. The estimation is aided by a feature called the bandwidth for the sake of the unconfoundedness assumption. The linear regression is implemented on data which are within the bandwidth away form the threshold. We introduced two existing bandwidth selection procedures, and proposed an alternative approach to determine the bandwidth. Then, the standard inference, for example, the hypothesis testing on the parameter of interest, was covered. At the end, we introduced a single hypothetical confounder, intentionally violated the unconfoundedness assumption, and considered three possible confounding models. The aspects of confounding effects in these models were analyzed.

6.1 Future Works

Although some bandwidth selection procedures (Ludwig and Miller, 2005; Imbens and Kalyanaraman, 2011) are introduced in this paper, there is still no wide agreement on how large the bandwidth should be for the unconfoundedness to be valid, and for the accurate estimation of the local causal effect. Therefore, we proposed an alternative approach that can properly balance the two issues. However, as noted earlier, the new approach contains a tuning parameter λ, and there is still no systematic way to determine optimal λ. For the simulation from Chapter 4, we chose λ such that the alternative bandwidth selection yields the known optimal bandwidth with this λ using a single data set. To complete the newly proposed bandwidth selection, one can try to devise a systematic algorithm to find an optimal λ.

In Chapter 5, we conducted some sensitivity analyses to see how confounding can affect the design when the unconfoundedness assumption is violated. For the simulation, we assumed a simple distribution of the confounder H, sharp nature of the treatment assignment (ψ_β), and linear confounding on the outcome (ϕ_γ). However, this is extremely simple model and can be unrealistic in practice. Therefore, one can try to give a thorough sensitivity analyses with more complicated and realistic features on ξ_α, ψ_β, and ϕ_γ, i.e., the relationships of H with X, T, and Y. One can inspect the aspects of confounding with multiple confounders as well.

The two challenges above is about completing and enriching the works of this paper. However, there are some new topics one can further study aside from those. One is the adoption of the bayesian framework in the RD design. The most works of the RD design is done under the frequentist framework, i.e., the true values of the parameters related to the distributions of the variables, and the coefficients of the regression functions are assumed to be fixed. One can, however, take the bayesian approach, which assigns prior distributions of those parameters and coefficients. We recommend Geneletti et al. (2015, 2019) for the starting point.

The other is the regression kink (RK) design, which is one variation of the RD design. While the RD design estimates the discontinuity in the conditional expectation of an outcome variable given an assignment variable, the RK design estimates the discontinuity in the first derivative (kink) of the conditional expectation. The RK design is often applied when assessing the effect of the government policy, especially economic policy (Card, Lee, Pei, and Weber, 2012, 2015, 2016).

Bibliography

Berk, R. (2010). Recent perspectives on the regression discontinuity design. In A. R. Piquero and D. Weisburd (Eds.), *Handbook of Quantitative Criminology*, pp. 563–579. Springer, New York, NY.

Berk, R. A. and D. Rauma (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association 78*, 21–27.

Bor, J., E. Moscoe, P. Mutevedzi, M.-L. Newell, and T. Bärnighausen (2014). Regression discontinuity designs in epidemiology causal inference without randomized trials. *Epidemiology 25*, 729–737.

Bühlmann, P. and M. Mächler (2016). *Computational Statistics (Lecture Note).* ETH Zürich.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist 24*, 409–429.

Cappelleri, J. C. and W. M. Trochim (2015). Regression discontinuity design. In J. Wright (Ed.), *International Encyclopedia of the Social Behavioral Sciences, 2nd Edition*, Volume 20, pp. 152–159. Elsevier.

Card, D., D. Lee, Z. Pei, and A. Weber (2012). Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design. *NBER Working Paper No. 18564.*

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015). Inference on Causal Effects in a Generalized Regression Kink Design. *Econometrica 83*, 2453–2483.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2016). Regression Kink Design: Theory and Practice. *NBER Working Paper No. 22781.*

Cattaneo, M. D., N. Idorobo, and R. Titiunik (2018). *A Practical Introduction to Regression Discontinuity Designs: Volume II. Monograph prepared for "Cambridge Elements: Quantitative and Computational Methods for Social Science" (Preliminary and Incomplete).* New York, NY: Cambridge University Press.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2019). The regression discontinuity design. Manuscript in preparation. Available on: https://arxiv.org/pdf/1906.04242.pdf.

Cook, T. D. (2008). "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics 142*, 636–654.

Fan, J. and I. Gijbels (2003). *Local Polynomical Modelling and Its Applications.* London; New York and Melbourne: Chapman and Hall.

Freedman, D., R. Pisani, and R. Purves (2007). *Statistics, Fourth Edition.* New York: W.W. Norton & Company.

Geneletti, S., A. G. O'Keeffe, L. D. Sharples, S. Richardson, and G. Baio (2015). Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine 34*, 2334–2352.

Geneletti, S., F. Ricciardi, A. G. O'Keeffe, and G. Baio (2019). Bayesian modelling for binary outcomes in the regression discontinuity design. *Journal of the Royal Statistical Society: Series A 182*, 983–1002.

Groß, J. (2003). *Linear Regression.* Springer-Verlag Berlin Heidelberg.

Hahn, J., P. Todd, and W. van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*, 201–209.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Imbens, G. and K. Kalyanaraman (2011). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies 79*, 933–959.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics 142*, 615–635.

Jacob, R., P. Zhu, M.-A. Somers, and H. Bloom (2012). A practical guide to regression discontinuity. *MDRC*.

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics 142*, 675–697.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Li, K.-C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics 15*, 958–975.

Ludwig, J. and D. L. Miller (2005). Does head start improve children's life chances? Evidence from a regression discontinuity design. *NBER Working Paper No. 11702.*

Moscoe, E., J. Bor, and T. Bärnighausen (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology 68*, 132–143.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics 2*, 1–26.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology 51*, 309–317.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation.* Beverly Hills, CA: Sage Publications.

Appendix A

Tabular results for bandwidth selections in Chapter 4

Bandwidth Selection Results			
Bandwidths	Alternative Approach	Cross Validation	Quantile Cross Validation
0.20	196 (3.92%)	63 (1.26%)	70 (1.40%)
0.25	201 (4.02%)	55 (1.10%)	81 (1.62%)
0.30	270 (5.40%)	67 (1.34%)	100 (2.00%)
0.35	316 (6.32%)	89 (1.78%)	109 (2.18%)
0.40	407 (8.14%)	103 (2.06%)	117 (2.34%)
0.45	505 (10.10%)	123 (2.46%)	118 (2.36%)
0.50	588 (11.76%)	99 (1.98%)	133 (2.66%)
0.55	615 (12.30%)	97 (1.94%)	142 (2.84%)
0.60	636 (12.72%)	123 (2.46%)	139 (2.78%)
0.65	463 (9.26%)	144 (2.88%)	159 (3.18%)
0.70	307 (6.14%)	157 (3.14%)	195 (3.90%)
0.75	236 (4.72%)	164 (3.28%)	223 (4.46%)
0.80	132 (2.64%)	215 (4.30%)	277 (5.54%)
0.85	64 (1.28%)	262 (5.24%)	353 (7.06%)
0.90	40 (0.80%)	414 (8.28%)	441 (8.82%)
0.95	15 (0.30%)	742 (14.84%)	739 (14.78%)
1.00	9 (0.18%)	2083 (41.66%)	1604 (32.08%)

Appendix B

R codes for the figures

All of R codes used for this paper can be found on the following GitHub page.
https://github.com/hkk828/Regression-Discontinuity-Design.
Below are the file names for the corresponding figures in the paper.

- RandomExpSolvesConfounding.R : Figure 2.2, 2.3
- AssignmentProbabilityConditionalExpectations.R : Figure 3.1, 3.2, 3.3, 3.4
- LocallyParamApp.R : Figure 4.1
- EconCVFigure.R : Figure 4.2
- BandwidthSelectionResults.R : Figure 4.3, Tabular results in Appendix A
- EstimationFRD.R : Figure 4.4, 4.5
- XYconfounding.R : Figure 5.3
- XTconfounding.R : Figure 5.5

The following are the R codes for the bandwidth selection methods.

- EconCV.R : Cross Validation Approach
- mywindow.R : Newly Proposed Method

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

REGRESSION DISCONTINUITY DESIGN

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
KIM	HONGKYU

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics.*

Place, date:

Zürich, 16.03.2020

Signature(s):

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.