The Regression Discontinuity Designs*

Yingyu Wu, Yushin Chen, Zihan Yang

February 10, 2023

Abstract

In recent years, as one of the most important quasi-experimental methods, the importance of regression discontinuity design (RDD) in economics research is not only reflected in the rapid development of its application in practice, but also in a series of frontier theoretical breakthroughs and improvements. In this paper, we review some of the practical issues in implementation of regression discontinuity methods, including the basic setup of RDD, identification of sharp RD and fuzzy RD, kernel selection, estimator selection, and bandwidth selection. Furthermore, we also did Monte Carlo simulations to validate some general conclusions.

1 Introduction

In economics research, researchers are interested in causal relationships between variables in order to explain the patterns behind economic phenomena or to assess the impact of specific event shocks and economic policies. Among the many quasi-experimental methods, the Regression Discontinuity Design (RDD) is a very important method for causal identification and has some unique advantages. In general, regression discontinuity designs are closer to randomized experiments than other methods and can yield estimates similar to those of RCTs (Lee and Lemieux, 2010[1]), allowing for the reduction of causal effects from experimental benchmarks (Green et al., 2009[2],), have stronger causal inference, can avoid the endogenous problem of causal estimation, and reflect the true causal relationship between variables (Lee, 2008[3]). Thus, RDD is one of the most credible quasi-experimental methods for causal inference and policy evaluation (Cattaneo and Titiunik, 2022[4]). In addition, RDD is able to identify causal effects

*Yingyu Wu, Bonn University. Matrikel Nr.: 3497426. Email: ywuu1@uni-bonn.de.
Yushin Chen, Bonn University. Matrikel Nr.: 3461265. Email: s6ynche2@uni-bonn.de.
Zihan Yang, Bonn University. Matrikel Nr.: 3504373. Email: s6ziyang@uni-bonn.de.

under weaker assumptions, and the assumptions are easily tested (Valentim et al., 2021[5]). It also has the flexibility to estimate, infer and robustly test local average treatment effects by using different estimation methods, in both parametric and non-parametric ways, and adjusting the bandwidth, to enhance the reliability of RDD estimation results (Cattaneo et al., 2023[6]). Therefore, it is necessary to have a good understanding of the way RDD works.

Guido and Thomas (2008)[7] provided a summary guide of steps to be followed when implementing RDD. This paper reviews the basic setup and concepts of regression discontinuity designs in section 2, explains the way to identify SRD and FRD, and also discusses necessary assumptions. This portion is mainly done by Yingyu Wu. In section 3, different types of kernels, different estimators, and different bandwidth selection methods are introduced. This portion is mainly done by Yuhsin-Chen. In section 4, we present our Monte Carlo simulation proceeds and results. This portion is mainly done by Zihan Yang.

2 Regression Discontinuity Designs

To start with, let us consider an example at first. Assume that students with a National College Entrance Exam (NCEE) score larger or equal than 100 can attend university, otherwise can not attend university. The difference between the salaries of those whose score is slightly smaller than 100 and those whose score is slightly larger than 100 after they graduate is actually what regression discontinuity design want to estimate.

We adopt the Robin (1974)[8] potential outcomes framework in the context of a binary treatment. Let $Y_i(0)$ denote the outcome without treatment, and $Y_i(1)$ denote the outcome with treatment. Since we can not observe the pair $Y_i(0)$ and $Y_i(1)$ at the same time, we therefore focus on the average effects of treatment, that is, averages of $Y_i(1) - Y_i(0)$. For treatment status, RDD use $W_i \in \{0,1\}$ to denote treatment received or not. Then, the outcome observed can be written as this:

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

The regression representation of observed outcome is:

$$Y_i = \alpha + \beta_i W_i + u_i,$$

which implies

$$\beta_i = Y_i(1) - Y_i(0)$$

Except the assignment W_i and the outcome Y_i, we may observe a vector of covariates or pretreatment variables denoted by (X_i, Z_i), where X_i is a scalar and Z_i is an M-vector, and both X_i, Z_i are covariates. Note that both X_i and Z_i are known not to have been affected by the treatment.

RDD exploits the facts that, Some rules are arbitrary and generate a discontinuity in treatment assignment. For example, NCEE score equals to 100 is the rule that give students different treatment assignment. The assignment of treatments is determined based on the value of a predictor (the covariate X_i) located on either side of a fixed threshold. Assuming that the association between the predictor and the potential outcome is smooth, which means other factors do not change abruptly at threshold. Then any change in outcome of interest can be attributed to the assigned treatment.

In general, depending on enforcement of treatment assignment, RDD can be categorized into two types:

1. Sharp Regression Discontinuity (SRD): the units below threshold do not get treatment, and the units above threshold do receive treatment. Everyone follows treatment assignment rule.

2. Fuzzy Regression Discontinuity (FRD): some units above threshold do not get treatment and some units below threshold do receive treatment. The probability of getting the treatment jumps discontinuously at the cutoff point.

2.1 Sharp Regression Discontinuity

2.1.1 Identification for SRD

In Sharp Regression Discontinuity (SRD) design, units assigned to or selected for treatment solely on the basis of a cut-off point value for an observed continuous variable, called the assignment (a.k.a., forcing, selection, running) variable. It can be a single variable, like credit score, income, accounting variable. It can also be a function of a single variable, or a function of several variables mapping into $\mathbb{R}$, like average quarterly debt-to-ebitda ratio and sum of all household expenditures.

Treatment assignment W_i is a deterministic function of the forcing variable X_i and the cutoff point c in SRD:

$$W_i = \begin{cases} 1 & X_i \geq c \\ 0 & X_i < c \end{cases}$$

For $W_i = 1$, the units are assigned to the treatment group; for $W_i = 0$, the units are assigned to the control group. The conditional expectation function of the observed outcome is:

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x] = \begin{cases} \mu_0(x) = \mathbb{E}[Y_i(0)|X_i = x], & \text{if } x < c \\ \mu_1(x) = \mathbb{E}[Y_i(1)|X_i = x], & \text{if } x \geq c. \end{cases}$$

Denote the average treatment effect at $X_i = x$ as $\tau(x)$, then

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mu_1(x) - \mu_0(x) \tag{1}$$

The average casual effect of the treatment at the cut-off point is :

$$\tau_{SRD} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c] \tag{2}$$

It should be clear that there are difference between equation (1) and equation (2), equation (2) is only the average treatment effect for individuals at $x = c$, i.e., the estimation is actually about the **Local Average Treatment Effect (LATE)**, which limits the external validity of the regression discontinuity estimate to some extent.

2.1.2 Assumptions

In SRD, we look at the discontinuity in the conditional expectation of the outcome to uncover the average treatment effect, which is :

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

There is a question. The assignment variable may be correlated with the outcome variable. When comparing averages of treatment and control, effect of W on Y will be confounded by X. The strong ignorability conditions (Rosenbaum and Rubin, 1983[9]) requires that Y_i be independent of W_i conditional on X_i (unconfoundedness), and for all values of the covariate, there are both treated and control units (overlap). But neither of these conditions can be satisfied, since conditional on the covariates there is no variation in the treatment, and in SRD, for all

values of x the probability of assignment is either 0 or 1. This implies we have to extrapolate. To avoid excessive extrapolation, we assume potential outcomes do not change at the threshold.

Key Assumption: (Local Continuity)

$$\mathbb{E}[Y(0)|X=x] \text{ and } \mathbb{E}[Y(1)|X=x], \text{ are continuous in } x \text{ at } c.$$

This means that except treatment assignment, all other unobserved determinants of Y_i are continuous at cutoff point c, which implies no other confounding factor affects outcomes at c. Thus, any observed discontinuity in the outcome can be attributed to treatment assignment, ensuring that units near the threshold are comparable.

Since we assumed the conditional expectations were continuous only at the cut-off point, we make some stronger assumptions to have a credible estimation.

Stronger Assumption 1: (Continuity of Conditional Regression Functions)

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

Stronger Assumption 2: (Continuity of Conditional Distribution Functions)

$$F_{Y(0)|X}(y|x) \text{ and } F_{Y(1)|X}(y|x), \text{ are continuous in } x \text{ for all } y.$$

The key difference is that these assumptions require continuity for all x, as opposed to only at the point of discontinuity, it is rare to assume continuity for one value of x and not others.

Under either assumption, we can know that equation (2) and equation(3) are equivalent. Thus, τ_{SRD} satisfies:

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

2.2 Fuzzy Regression Discontinuity

The main difference between FRD and SRD is that at cutoff point $x=c$, the probability of units getting a treatment jumps from a to b rather than from 0 to 1, where $0 < a < b < 1$. Treatment assignment is no longer a deterministic function of the forcing variable X_i. There are other variables affect treatment assignment, some of them could be unobserved. This also means, in fuzzy RDD, X at cutoff point c is a predictor of who gets treatments, but not completely determines treatment assignment. The graphical comparison between SRD and FRD is in section 4.

Therefore the practical implication of FRD is that the conditional probability is discontinuous as X approaches c in the limit. The formal definition of a probabilistic treatment assignment can be written as:

$$0 < \lim_{x \downarrow c} Pr[W_i = 1 | X_i = x] - \lim_{x \uparrow c} Pr[W_i = 1 | X_i = x] < 1$$
$$\Leftrightarrow \lim_{x \downarrow c} Pr[W_i = 1 | X_i = x] \neq \lim_{x \uparrow c} Pr[W_i = 1 | X_i = x]$$

In this design we interpret the ratio of the jump in the regression of the outcome on the covariate to the jump in the regression of the treatment indicator on the covariate as an average causal effect of the treatment, which can be written as equation (4). For the detailed process of identifying ATE of FRD, please check the appendix A.1.

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

Hahn, Todd and van der Klaauw (2001)[10] pointed out that one must assume "monotonicity" (i.e., X crossing the cutoff cannot simultaneously cause some units to take up and others to reject the treatment) and "excludability" (i.e., X crossing the cutoff cannot impact Y except through impacting receipt of treatment), then equation (4) can be identifies as LATE (equation (5)). The LATE represents the average treatment effect of the "compliers", which are units that would receive the treatment when they satisfy the cutoff rule ($X_i \geq c$), but otherwise would not receive it. Define compliance status:

$$\lim_{x \downarrow X_i} W_i(x) = 0 \quad \textit{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

The nevertakers are units that whatever forcing variable value is, they do not receive treatment:

$$\lim_{x \downarrow X_i} W_i(x) = 0 \quad \textit{and} \quad \lim_{x \uparrow X_i} W_i(x) = 0.$$

The alwaystakers are units that whatever forcing variable value is, they all receive treatment:

$$\lim_{x \downarrow X_i} W_i(x) = 1 \quad \textit{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

Then,

$$\tau_{FRD} = \mathbb{E}[Y_i(1) - Y_i(0) | \textit{unit } i \textit{ is a complier and } X_i = c] \tag{5}$$

Since FRD restricts subpopulation even further to that of the compliers with x close to c,

and also both SRD and FRD estimate the average effect of the subpopulation with x close to c, only with strong assumptions (e.g., homogeneous treatment effect) can we estimate the overall average treatment effect. Thus, RDD have strong internal validity but weak external validity.

3 Non-parametric Estimation

Before knowing how to estimate the treatment effect (SRD or FRD) by the local estimator, we need to know some concepts in advance, including the kernel, local constant estimator, and local linear estimator. Therefore, these topics would be introduced in the following part.

3.1 Kernel Function

$$K\left(\frac{Y_i - y}{h}\right)$$

Firstly, we can understand the kernel function as a bounded, symmetric, probability density function and people often take it as a weighted method. The variable Y_i in the kernel function would be weighted. In most cases, when Y_i is closer to y, the kernel function would return a larger number. There are many different kernels, which means there are many weighted ways. Like what we list following. However, the difference between those kernels in the local estimator is quite small. Additionally, in some situations, we only have the dataset Y_i, but we do not have their original distribution of y. Then we need to estimate them piece by piece, like a histogram with a very thin bin. Each time we only estimate the observations around y. Here "h", bandwidth, denotes the range around y. If the observation is outside of the range, then the kernel function would just return zero. Even when Y_i tends to infinity or minus infinity. Finally, compose all of them together and we would get our estimation result for y distribution. For the detail of kernel assumption and kernel density estimation derivation process please check appendix B.1.1 B.1.2

Table 1: **Normalized Kernels**

Kernel	Density Function	$R(k)$
$Uniform$	K(u) = $\frac{1}{2\sqrt{3}}$ if $\lvert u\rvert < \sqrt{3}$	$\frac{1}{2\sqrt{3}}$
$Gussian$	$K(u) = \frac{1}{\sqrt{2\pi}} * e^{\frac{-1}{2}*u^2}$	$\frac{1}{2\sqrt{\pi}}$
$Epanechnikov$	$K(u) = \frac{3}{4\sqrt{5}} * \left(1 - \frac{u^2}{5}\right)$ if $\lvert u\rvert < \sqrt{5}$	$\frac{3\sqrt{5}}{25}$
$Triangular$	$K(u) = \frac{1}{\sqrt{6}} * \left(1 - \frac{\lvert u\rvert}{\sqrt{6}}\right)$ if $\lvert u\rvert < \sqrt{6}$	$\frac{\sqrt{6}}{9}$

3.2 Local Polynomial Estimator

After knowing different kernels, we can use **weighted least squared method**, to derive the polynomial estimator. In order to save space, the derivation process please check B.1.3. Generally, it is similar to the least squared method but also uses Taylor expansion and puts the weight in the objective function. We have the objective function like the following.

$$\min \sum K\left(\frac{X_i - X}{h}\right)(Y_i - \hat{m}(x))^2$$

And the result is showing like following.

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

$$x = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - X & \cdots & (X_n - x)^p \end{bmatrix}$$

$$W = diag(K\left(\frac{X_1 - X}{h}\right), K\left(\frac{X_2 - X}{h}\right), \cdots, K\left(\frac{X_n - X}{h}\right) \text{ and } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Now if p=0, we can obtain the local constant estimator, and the estimator is also called the **Nadaraya-Waston estimator**, which is from Nadaraya, E.A. (1964) [11] and Watson, G.S. (1964) [12]. It is like following.

$$\beta_0 = \frac{\sum K\left(\frac{X_i - X}{h}\right) * Y_i}{\sum K\left(\frac{X_i - X}{h}\right)}$$

If p = 1 we can obtain a local linear estimator like the following.

$$\sum \frac{1}{n} * \frac{\hat{s_2}(x;h) - \hat{s_1}(x;h)(X_i - x)}{\hat{s_2}(x;h)\hat{s_0}(x;h) - \hat{s_1}(x;h)^2} * K\left(\frac{X_i - x}{h}\right) * Y_i$$
$$where\ \hat{s_r}(x;h) = \frac{1}{n}\sum (X_i - x)^r K\left(\frac{X_i - x}{h}\right)$$

3.2.1 Asymptotic Bias and Variance

Taking a glance at both estimators, we can find that except for the kernel, the only item we can control is "bandwidth". Therefore, we can imagine that bandwidth is very important for our

estimation. In fact, both estimators face a tradeoff between bias and variance when selecting the optimal bandwidth. If we can choose the bandwidth properly, we can estimate precisely. We would see this point clearly when we have asymptotical bias and variance later on.

Before deriving the bias of the estimator, we need the following assumption.

$$h \to 0 \; and \; nh \to \infty \tag{6}$$

$$m(x), f(x) \; and \; \sigma_X^2 \; is \; continuous \; in \; some \; neighborhood \; N \; of \; x \tag{7}$$

$$f(x) > 0 \tag{8}$$

Point (6) is an interesting point of view for these estimators because this is the only way that we can let the bias and variance shrink to zero at the same time. Additionally, the set-up for point (7) is for smoothness. For point (8), where f(x) is the marginal density function of x. We need f(x)>0 because our estimation is conditionally at point x, and we need to have observations X_i around x for estimation. For the derivation process of the bias and variance of the local constant estimator please check B.1.5 B.1.6. For the process for linear estimator please check Fan, J. and Gijbels, I. (1996)[13]

$$Asymptotic \; Bias = \begin{cases} \frac{h^2\mu_2(K)}{2}\left(\frac{2m'(x)f'(x)}{f(x)} + m''(x)\right) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right) \\ = B_0h^2 + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)...NW \\ \frac{h^2\mu_2(K)}{2}(m''(x)) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right) \\ = B_1h^2 +_{op} (h^2) + o_p(h^2) + O_p\left(\sqrt{\frac{h}{n}}\right)...LL \end{cases}$$

$$Asymptotic \; Variance = \begin{cases} \frac{R(k)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right)...NW \\ \frac{R(k)\sigma^2(x)}{nhf(x)} + o\left(\frac{1}{nh}\right)...LL \end{cases}$$

$$where \; \mu_2 = \int t^2K(t)dt \; R(k) = \int K(x)^2dx$$

Here NW denotes the Naradaya-Waston(local constant) estimator and LL denotes the local linear estimator. We can observe that the bias of local constant estimators is related to h, $m'(x)$, $f'(x)$, and $m''(x)$. If h tends to zero, then bias tends to zero. Secondly, the curvature $m''(x)$ also, play a role when calculating the bias. If $m''(x) < 0$, usually at the peak, we are more likely to have a negative bias, and when $m''(x) > 0$, usually at the valley we are more likely to have a positive bias, but it also depends on other terms. Thirdly, when $m'(x)$ is not equal to zero, because f(x) is the density function of x, so if $f'(x) < 0$, it means that there are more

observations at the left than those at the right of x and it could also cause biased. Intuitively, we can also understand as that when we are estimating a constant estimator, we are taking the mean conditionally on x, so the unbalanced data point matter for biased. In contrast, for the local linear estimator, we can find that the only term left over in the bias is h and $m''(x)$, so it means that now only the curvature and bandwidth would affect the bias and it would not be affected by the slope of f(x) anymore.

For asymptotic variance, it is identical for both estimators. We can also find that when n multiplies h tends to infinity, then the variance tends to zero. Therefore, only when assumption (6) holds, our asymptotic bias and asymptotic variance would tend to zero. Besides n and h, the variance is also decided by R(k), f(x), and sigma squared, which is the squared of the difference between m(x) and y(x). When introducing the kernel, we found that the Epanechnikov kernel has the lowest R(k), but in fact, the difference of R(k) for each kernel is not large. Additionally, if f(x) is relatively small, which means the density is low, we are more likely to have a higher variance. It is also why we need assumption (8) to hold, otherwise, the variance may diverge. Besides, the sigma squared is $E(e(x)^2|X=x)$ and $e(x)=Y(x)-m(x)$. If sigma squared is large, also means that is not precise, and the variance of the estimator also goes large.

3.3 Comparing Local Constant Estimator with Local Linear Estimator

Firstly, Yu, K., & Jones, M. C. (1997)[14] mentioned that the difference between the local constant estimator and the local linear estimator is small in the interior. However, the main demerit of the local constant estimator is that it has boundary biased, which is different from the local linear estimator. To understand it intuitively, assume that we have a distribution $X \sim U[0,5]$, and $Y \sim [X,1]$. In this case, we would have m(x) = x. If we want to estimate the m(0) at low boundary x = 0, we would always have an upward biased, because we do not have any observations x smaller than 0. In contrast, if we want to know the value m(5) at x = 5, we would have a downward biased because all of our x are smaller or equal to 5. On the other hand, if we apply a local linear estimator, then because there is a slope not equal to zero, it can be unbiased. Additionally, from a math perspective, BRUCE E. H. (2021) [15] and Imbens, G. W.(2007) [7] mentioned that at the boundary point, the convergence rate of bias of the local constant estimator is h, notice that it is not h squared, which is slower than the convergence rate of the local linear estimator, which is h squared. Moreover, Langat, R. C. (2020) [16] also prove it further from the Monte Carlo simulation perspective. Langat, R. C. (2020) [16] apply five different kinds of

data generation functions, including cubic, bump, quadratic, linear, and exponential functions in Monte Carlo simulation. For detail please check B.1.7. Also, experiment with different sample sizes. Unsurprisingly, LLR usually has lower AMISE than the local constant estimator in every scenario.

Additionally, from the RDD perspective, the boundary point is important. Because we are actually running the nonparametric regression at the boundary point in order to get the treatment effect. Therefore, using a local linear estimator is a better option.

3.4 Bandwidth Selection

3.4.1 Asymptotic Mean Integrated Squared Error and Bandwidth selection

After knowing asymptotic bias and variance, we can start to derive the optimal bandwidth. First, We need to calculate the mean integrated square error (MISE). It can be written in the following form.

$$MISE_n[m(\hat{x};h)] = \int E[(m(\hat{x};h) - m(x))^2]dx$$

$$h_{MISE} = \arg\min_{h>0} MISE[m(\hat{x};h)]$$

Here, we are calculating the MSE of function m(x) at each x point, therefore, we need to integrate them together. Following, the same as what we know, MSE is equal to the squared of bias plus variance. So we plug in our asymptotic squared bias and variance. Then using the first-order condition to derive the best h makes bias and variance small simultaneously. Here we can directly plug in the asymptotic squared bias and the variance for both estimators. Note that, in order to focus on the place where the density is relatively high, we multiplied f(x) inside. The effect is like the weight. Besides that, w(x) is also multiplied inside. w(x) is an integrable function. We adding it to prevent it from x being unbounded. In other words, if x is bounded, we can get rid of it. If x is unbounded, we may need to set up our interest region, such as the following.

$$w(x) = 1[\phi_1 \leq x \leq \phi_2]$$

$$AMISE[m(\hat{x};h)] = \int B_p^2 h^4 f(x)w(x)dx + \frac{R(k)}{nh}\int \sigma^2(x)w(x)dx, p = 0 or 1$$

Then, calculate the derivative on h and make a first-order condition. Then finally we get the

bandwidth.

$$h_{AMISE} = [\frac{R(K)\int \sigma^2(x)w(x)dx}{4n\int B_p^2 w(x)f(x)dx}]^{\frac{1}{5}} \tag{9}$$

Simplified to the following form

$$\begin{aligned} \bar{B} &= \int B_p(x)^2 f(x)w(x)dx \\ \bar{\sigma}^2 &= \int \sigma(x)^2 w(x)dx \\ h_{AMISE} &= [\frac{R_k\bar{\sigma}^2}{4n\bar{B}}]^{\frac{1}{5}} \end{aligned} \tag{10}$$

If we plug optimal h back into AMISE equation, with some algebra, we can have the following result.

$$AMISE \approx 1.65(R_k^4\bar{B}\bar{\sigma}^8)^{\frac{1}{5}}n^{\frac{-4}{5}}$$

Notice that, although R_k would affect the AMISE, if we actually plug in the R_k for each kernel that we list before, we can find that the difference is little. In the rule of thumb, we would only assume $(\frac{R(k)}{4})^{\frac{1}{5}}$ as 0.58, the detail derivation of rule of thumb please check B.1.4. The optimal bandwidth for the rule of thumb is like the following. It is derived from equation (10).

$$\begin{aligned} h_{rule\ of\ thumb} &= 0.58[\frac{\hat{\sigma}(\phi_2-\phi_1)}{n\hat{B}}]^{\frac{1}{5}} \\ \hat{B} &= \frac{1}{n}\sum_{i=1}^{n}[\hat{B_p}(x_i)^2]1[\phi_1 \le x_i \le \phi_2] \end{aligned} \tag{11}$$

3.5 Applying LLR in RDD

3.5.1 Estimation For SRD

$$\begin{aligned} (\hat{\alpha}_{yl}, \hat{\beta}_{yl}) &= \arg\min_{\alpha_{yl}\beta_{yl}} \sum_{i:c-h<X_i<c} (Y_i - \alpha_i - \beta_i(X_i - c))^2 \\ (\hat{\alpha}_{yr}, \hat{\beta}_{yr}) &= \arg\min_{\alpha_{yr}\beta_{yr}} \sum_{i:c<X_i<c+h} (Y_i - \alpha_i - \beta_i(X_i - c))^2 \end{aligned}$$

Firstly, using a local linear estimator to find the alpha hat and the beta hat. Notice that the range of X_i is from c-h to c and c to c+h. After that, we plug in "c", our cutoff point, to both estimation functions.

$$\hat{\mu}_l(c) = \hat{\alpha}_{yl} + \hat{\beta}_{yl}(c-c) = \hat{\alpha}_{yl} \; and \; \hat{\mu}_r(c) = \hat{\alpha}_{yr} + \hat{\beta}_{yr}(c-c) = \hat{\alpha}_{yr}$$

$$\hat{\tau}_{SRD} = \hat{\alpha}_{yr} - \hat{\alpha}_{yl}$$

3.5.2 Estimation for FRD

For FRD, again, we use the local linear estimator four times (y and w) to derive the $\hat{\alpha}$ and $\hat{\beta}$, then plug in c again to get the following result.

$$(\hat{\alpha}_{yl}, \hat{\beta}_{yl}) = arg \min_{\alpha_{yl}\beta_{yl}} \sum_{c-h<X_i<c} (Y_i - \alpha_{yl} - \beta_{yl}(X_i - c))^2$$

$$(\hat{\alpha}_{yr}, \hat{\beta}_{yr}) = \arg \min_{\alpha_{yr}\beta_{yr}} \sum_{i:c<X_i<c+h} (Y_i - \alpha_{yr} - \beta_{yr}(X_i - c))^2$$

$$\hat{\tau_y} = \hat{\alpha_{yr}} - \hat{\alpha_{yl}}$$

$$(\hat{\alpha}_{wl}, \hat{\beta}_{wl}) = arg \min_{\alpha_{wl}\beta_{wl}} \sum_{c-h<X_i<c} (W_i - \alpha_{wl} - \beta_{wl}(X_i - c))^2$$

$$(\hat{\alpha}_{wr}, \hat{\beta}_{wr}) = arg \min_{\alpha_{wr}\beta_{wr}} \sum_{c<X_i<c+h} (W_i - \alpha_{wr} - \beta_{wr}(X_i - c))^2$$

$$\hat{\tau_w} = \hat{\alpha_{wl}} - \hat{\alpha_{wr}}$$

$$\frac{\hat{\alpha_{yr}} - \hat{\alpha_{yl}}}{\hat{\alpha_{wr}} - \hat{\alpha_{wl}}} = \frac{\hat{\tau}_y}{\hat{\tau_w}} = \hat{\tau}_{FRD}$$

3.5.3 Bandwidth Selection for SRD and FRD

There are some different methods for estimating the optimal bandwidth selection. Firstly, we can use the rule of thumb of bandwidth, which is exactly the same as equation (11), when we use the normalized kernel, definition please check B.1.8. But for unnormalize kernels, for instance, the rectangular kernel, BRUCE E. H. (2021) [15] suggest that we can change from 0.58 to 1.00 and for the triangular kernel, we can change from 0.58 to 1.42. Secondly, we can also apply a leave-one-out prediction error. In the beginning, we define CV estimator like the following.

$$CV_Y(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{\mu}_{-i}(X_i)]^2 w(x) \; and \; \hat{\mu}_{-i}(X_i) = \begin{cases} \hat{\mu}_{-il}(X_i) \; if \; X_i < c \\ \hat{\mu}_{-ir}(X_i) \; if \; X_i \geq c \end{cases}$$

Here $-i$ denote that we would not take into account ith observation when training the mu hat. Therefore, in this case, Yi would not be in the estimator $\hat{\mu}_{-i}$, but only the observation excludes Yi would be used for the estimator. To have a clear view, please check page 7, the

formula of Nadaraya-Waston estimator, there is Yi originally in the numerator, but now we would skip that one observation *i*th. Then afterward we obtain the optimal h like the following. We did this because we want to calculate the RSS but also want to avoid over-fitting problems.

$$h_{cv} = \arg\min_{h} CV(h)$$

Additionally, Imbens, G. W., & Lemieux, T. (2008) [7] suggest another method that would discard part of observations far away from the cut-off point. It can be shown like the following.

$$CV_Y^{\delta}(h) = \frac{1}{n} \sum_{i:q_{X,\delta,l} \le x_i \le q_{X,1-\delta,r}}^{n} [Y_i - \hat{\mu}(X_i)]^2 \text{ and } \hat{\mu}(X_i) = \begin{cases} \hat{\mu}_l(X_i) \ if \ X_i < c \\ \hat{\mu}_r(X_i) \ if \ X_i \ge c \end{cases}$$

$$h_{cv}^{\delta} = \arg\min_{h} CV_Y^{\delta}(h)$$

Here, q is the quantile of X. The observation outside the quantile would be discarded. The expected value can be shown in the following. Derivation of the expected value of CV estimator please check B.1.9

$$E[CV_Y(h) | q_{X,\delta,l} < X < q_{X,1-\delta,r}] = [\bar{\sigma} + \int Q(x,h) f_x | q_{X,\delta,l} < X < q_{X,1-\delta,r}]$$

$$where \ Q(c,h) = \frac{1}{2}(Q_l(c,h) + Q_r(c,h))$$

$$Q_l(x,h) = E[(\lim_{z \uparrow c} \mu(z) - \hat{\alpha}_l(c))^2] \ Q_r(x,h) = E[(\lim_{z \downarrow c} \mu(z) - \hat{\alpha}_r(c))^2]$$

Notice that the sigma bar, in the expected value, is independent of h. Imbens, G. W., & Lemieux, T. (2008) [7] stated that if we only interest is at the cut-off point, then the effect of f_x in the expected value equation is relatively small. In this case, Q(c,h) would be the main term we focus on. If we didn't discard those observations(without the condition) and our cut-off point is at the center of the distribution, and the tail of the distribution has only a few observations, we may get larger bandwidth than the optimal bandwidth we get from only using the observations around x=c. This is why we may need to discard them when implementing RDD. Additionally, for FRD case is similar, the only difference is that we need to estimate the bandwidth twice, for treatment(W) and outcome(Y)

$$h_{cv,Y}^{\delta} = \arg\min_{h} CV_Y^{\delta}(h) \text{ and } h_{cv,W}^{\delta} = \arg\min_{h} CV_W^{\delta}(h)$$

However, for this method, BRUCE E. H. (2021) [15] mentions that discarding part of X may let CV criteria become noise estimator, then increase variance.

4 Monte Carlo Simulation

In the Monte Carlo experiment, we have two simulations. The first one is to calculate the MSE of SRD and FRD through Monte Carlo simulation, then the second one is to explore the influence of the polynomial order on the MSE results.

For the first simulation, we design four different DGFs to generate data firstly. In the simulation, we set all cutoff point as 0 for all DGFs. Since for SRD, the value of treatment effect w is depended on the value of x, so we decide to generate all x by runif function. For y, it is different in different DGFs:

Table 2: **the generation function of** Y

DGF1	y=x+1+gap	$x \geq 0$
	y=x+1	x<0
DGF2	y=x^2+1+gap	$x \geq 0$
	y=$-x^2$+1	x<0
DGF3	y=x^2+x^3+1+gap	$x \geq 0$
	y=x^2+x^3+1	x<0
DGF4	y=x^2+x^3+x^4+1+gap	$x \geq 0$
	y=$-x^2$+x^3-x^4+1	x<0

This is the process of SRD. For FRD, it is similar, but according to the definition of FRD, we will select 10% data to be the always-takers and non-takers, and change the distribution of w. Here is the explanation:

In SRD, the data on the left side of the cutoff point are all observation objects who do not receive treatment, and the data on the right side of the cutoff point are all observation objects who receive treatment. In FRD, we randomly draw 10% total samples from the population as the sub-sample, if some of the sub-samples are originally from the left side of the cutoff point and make them accept treatment, then they are always-takers now; If these samples are from the right side of the cutoff point and they do not accept the treatment, then they are non-takers now. The data-generating process of SRD and FRD could be: the number of observations is 100; the range of x is from -10 to 10; the true treatment effect (gap) is 99. In the scatter plot of FRD, the outliers on the left side of the cutoff point are always-takers, and the outliers on the left side of the cutoff point are non-takers.

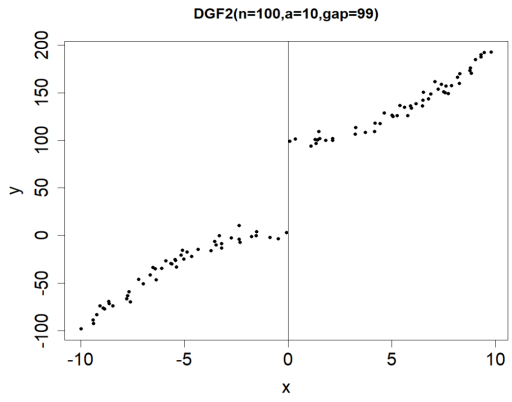

Figure 1: an example of SRD

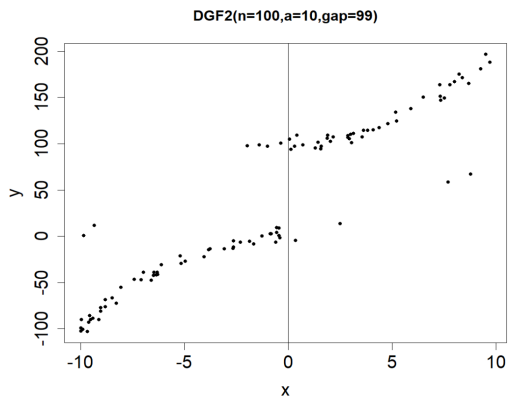

Figure 2: an example of FRD

4.1 SRD and FRD results

Local polynomial estimation consists of the following basic steps:

1. Choose a polynomial order p and a kernel function K.

2. Choose a bandwidth h.

3. For observations above the cutoff point, fit a weighted least squares regression of the outcome Y_i on a constant and (X_i-c),$(X_i-c)^2$,...,$(X_i-c)^p$, where p is the chosen polynomial order, with weight $K((X_i-c)/h)$ for each observation. The estimated intercept from this local weighted regression, $\hat{\mu}_+$, is an estimate of the point $\hat{\mu}_+$=$E[Y_i(1)|X_i=c]$:

$$\hat{\mu}_+ : \hat{Y}_+ = \hat{\mu}_+ + \hat{\mu}_{+,1}(X_i-c) + \hat{\mu}_{+,2}(X_i-c)^2 + ... + \hat{\mu}_{+,p}(X_i-c)^p \tag{12}$$

4. For observations below the cutoff point, fit a weighted least squares regression of the outcome Y_i on a constant and (X_i-c),$(X_i-c)^2$,...,$(X_i-c)^p$, the definition of p is same as before. The estimated intercept from this local weighted regression, $\hat{\mu}_-$, is an estimate of the point $\hat{\mu}_-$=$E[Y_i(0)|X_i=c]$:

$$\hat{\mu}_- : \hat{Y}_- = \hat{\mu}_- + \hat{\mu}_{-,1}(X_i-c) + \hat{\mu}_{-,2}(X_i-c)^2 + ... + \hat{\mu}_{-,p}(X_i-c)^p \tag{13}$$

5. Calculate the SRD point estimate: $\hat{\tau}_{SRD} = \hat{\mu}_+ - \hat{\mu}_-$

In this simulation, for every DGF, for sake of convenience, we set the number of loops to 1000, and the number of observations in each loop is also 1000. The gap (true treatment effect) is set to 9, because x follows the uniform distribution. We set the absolute value a of the upper and lower limits of the uniform distribution are same in the loop, and a = 10.

Here is the simulation results of SRD:

Table 3: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF1)	p=1	p=2	p=3	p=4
$triangular$	0.0745	0.1257	0.1770	0.2750
$epanechnikov$	0.0699	0.1194	0.1778	0.2773
$uniform$	0.0744	0.1168	0.1838	0.2580

Table 4: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF2)	p=1	p=2	p=3	p=4
$triangular$	3.4068	3.1687	4.6417	6.8762
$epanechnikov$	3.4597	3.3991	4.0950	5.9037
$uniform$	3.3799	2.8629	4.4227	6.3841

Table 5: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF3)	p=1	p=2	p=3	p=4
$triangular$	49.4130	59.8174	67.8998	99.9413
$epanechnikov$	53.7563	61.9821	74.1045	102.0846
$uniform$	54.3758	67.1499	70.5625	105.1054

Table 6: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF4)	p=1	p=2	p=3	p=4
$triangular$	498.6723	475.9539	503.0850	660.8360
$epanechnikov$	520.3914	477.5106	520.3394	608.7728
$uniform$	529.8861	500.1863	487.1633	611.0177

In each table above, we show the MSE of kernels in different p, which is the chosen polynomial order in equation (12). In each function, y is generated by different formulas, and MSE is influenced by the variance of y, so we will compare the results in four DGFs one by one. If we compare the results between different DGFs, MSE is not a scale-invariant statistic, so it is

hard to have any reasonable conclusions. By observing each table, we can draw the following three conclusions:

Firstly, in each table, we can find that there is no big difference between the use of three different kernels when p is the same. For example, in table 3, y is generated by the linear function. When $p = 1$, the MSE of triangular, epanechnikov and uniform is 0.0745, 0.0699, 0.0744 respectively. Secondly, with the increase of p, the MSE in each kernel is increasing generally. For example, for the first row of table 3, which shows the results when the kernel is triangular, we can see that with the increase of p, the MSE is bigger: 0.0745, 0.1257, 0.1770, 0.2750. Finally, the best order of the polynomial is 2 when data is generated by DGF2 and DGF4, since the MSE is the smallest when $p = 2$. And for data generated by DGF1 and DGF3, the best order of the polynomial is always 1.

We can also get similar conclusions in FRD, the results are put in the appendix C.1: In each table, with the increasing of p, the MSE is increasing generally, and there is no obvious difference of MSE by using different kernels. Also, the best order of the polynomial is 1 for DGF1, DGF3 and DGF4. For data generated by DGF2, the best order of the polynomial is 2.

a, which is the absolute value of the upper and lower limits of x, and it will influence the estimation of the MSE. When a is set to 20, the above conclusion about the best order of the polynomial will change, and we get another conclusion by Monte Carlo simulation: the best order of polynomial of different data generated by different DGFs is associated with the power of the y, which means that best order of the polynomial of every DGF is same with the power of y of the DGF. The results of the new simulation is in the appendix C.2. It can be explained: when a is bigger, the range of x is bigger and the variance of y is larger. In order to have a precise estimation, the bandwidth will be larger. Therefore, it will detect the true shape of the curve in the simulation. But when a is smaller, the perfect bandwidth is smaller, so the shape in the bandwidth is almost linear or quadratic.

When a is set to 20, even though n is becoming larger, the best order of polynomial of different DGFs is still same with the power of the y, the simulation results of FRD can be seen in the appendix C.3.

4.2 Local Polynomial Order

We get a conclusion from the above results: the local linear and quadratic estimator is always better than the 3rd and 4th polynomial estimators. Actually, in practice, the local linear esti-

mator has become the standard in the regression discontinuity design literature, there are three reasons from Cattaneo, Idrob, and Titiunik (2019)[17]: First, a polynomial of order zero (a constant fit) has undesirable theoretical properties at boundary points, which is precisely where RD estimation must occur. Second, for a given bandwidth, increasing the order of the polynomial generally improves the accuracy of the approximation but also increases the variability of the treatment effect estimator. Third, higher-order polynomials tend to produce over-fitting of the data and lead to unreliable results near boundary points. Combined, these factors have led researchers to prefer the local linear RD estimator, which by now is the default point estimator in most applications.

But there are also objections: David and Andrea (2014) [18] proved that the local linear estimator does not always have the lowest MSE based on two empirical examples, and the MSE of the p-th order local estimator depends on the sample size and the intrinsic properties of the data generating process.

Our group is very interested in this question, so we run the discontinuity regression simulation again with the data generated by the four DGFs to see if similar conclusions are drawn. In the above paper, they used two different DGFs to generate data and found when the number of observations is 6558, the best order of the polynomial is 4. So in this simulation, unlike previous part, the value of n are changed from 1000 to 10000. For more intuitive, we have made the results into line charts. Since there is no big difference among three kernels, just epanechnikov kernel are chosen to do the following simulation.

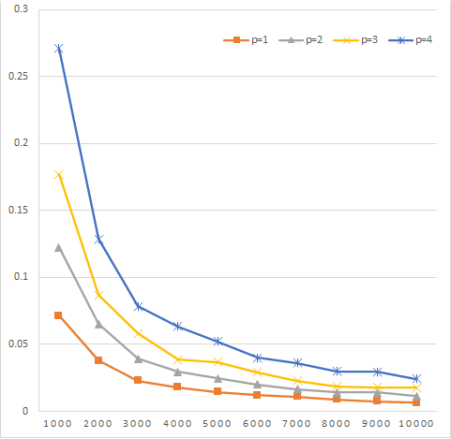

Figure 3: the MSE of DGF1 in SRD

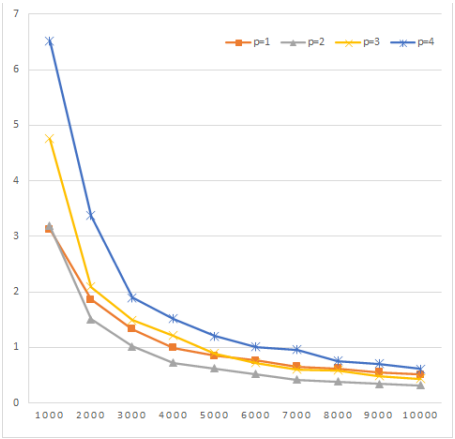

Figure 4: the MSE of DGF2 in SRD

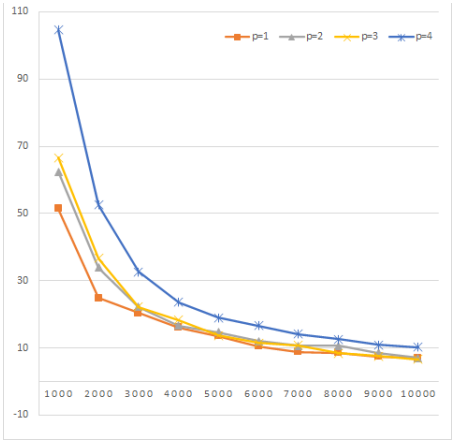

Figure 5: the MSE of DGF3 in SRD

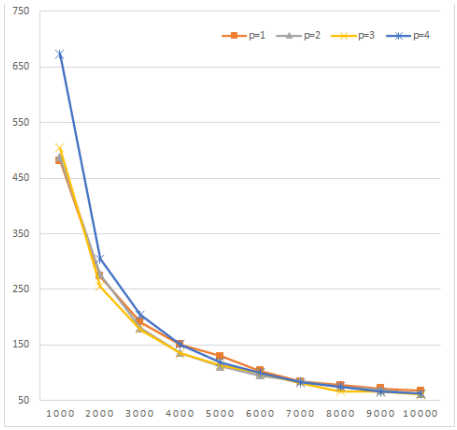

Figure 6: the MSE of DGF4 in SRD

The above line charts are MSE results of four DGFs in SRD. These figures show that as n gradually increases, the values of MSE are getting smaller and smaller. The reason is as follows:

The general form of the approximate (conditional) MSE for the RD treatment effect is:

$$MSE(\hat{\tau}_{SRD}) = Bias^2(\hat{\tau}_{SRD}) + Variance(\hat{\tau}_{SRD}) = h^{2(p+1)} * \beta + (1/nh) * \nu \tag{14}$$

Then the MSE-optimal bandwidth choice:

$$h_{MSE} = (\nu/(2(p+1)\beta^2))^{1/(2p+3)} * n^{-1/(2p+3)} \tag{15}$$

The quantities β and ν represent the (leading) bias and variance of the RD point estimator $\hat{\tau}_{SRD}$ respectively, not includ the rates controlled by the sample size and bandwidth choice. When n is getting bigger, the number of observations near the cutoff point increases, and the contribution of the variance term to the MSE decreases. And when the bandwidth is fixed, the variance and bias are decreasing with the increases of the number of observations near the cutoff point, so the MSE is decreased when n is bigger. On the other hand, when the number of observations is increasing, the best bandwidth is decreasing, which can be seen from the equation (15). But the effect of smaller bias and variance due to the increase in number, is greater than the effect of larger variance due to smaller bandwidth. Therefore, the MSE becomes smaller and smaller if n increases, which is reflected in the figures above.

From these line charts, we also can see that the best order of polynomials is 1 or 2, which is consistent with our previous conclusion. However, there is an exception: in figure 6, when n is bigger than 2000, the best order of the polynomial is 3. In general, it may be due to the conditional expectation function has a large curvature, but actually, the difference is not obvious.

We also did the simulation of the FRD, and the results are in the appendix C.4. In these line

charts, for the data generated by DGF 1 and DGF 4, the best order of the polynomial is 1, and for the data of DGF2 and DGF3, the best order of the polynomial is 2.

Hence, from the whole simulation, we can see that $p = 1$ should not be the universally preferred polynomial order across all empirical applications. In fact, in a lot of situations, $p = 2$ performs better than the local linear estimator. But we do not oppose the use of a local linear estimator in RD studies, it is a convenient choice and performs quite well in many applications. Our perspective is that if the purpose is to find the optimal order of polynomial or the conditional expectation of the outcome variable Y is close to being a high power function of the assignment variable X, the choice of the order of polynomial should not be limited to local linear.

5 Conclusion

In this paper, we gave the detailed explanations of how regression discontinuity works, discussed the necessary assumptions and identified sharp regression discontinuity and fuzzy regression discontinuity respectively. After that, we talked about the way we use to do estimation, which is non-parametric. In order to know how to estimate the treatment effect by the local estimator, we introduced the kernel function at first. Then we took kernels as weight and used the weighted least squared method, to derive the polynomial estimator. To select the optimal bandwidth, we need to know whether the local constant estimator or local linear estimator is better. Since the local constant estimator can be biased at the boundary point but would not happen when using the local linear estimator, we can say that using local linear estimator is better. By asymptotic mean integrated squared error, we derived the rule of thumb optimal bandwidth. We also applied an alternative "leave-one-out method" in optimal bandwidth selection. Then we did simulations for SRD and FRD respectively by applying the local linear estimator. All of our code is published on GitHub. We got two conclusions from these results: Firstly, estimation results are typically not very sensitive to the particular choice of kernel used. Secondly, the local linear and quadratic estimators deliver a good trade-off between simplicity, precision, and stability in RD settings. But we advocate for a more flexible view towards the choice of the polynomial order, especially considering the sample size and the data generation form.

References

[1] David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, June 2010.

[2] Donald P Green, Terence Y Leong, Holger L Kern, Alan S Gerber, and Christopher W Larimer. Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis*, 17(4):400–417, 2009.

[3] David S Lee and David Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

[4] Matias D Cattaneo and Rocio Titiunik. Regression discontinuity designs. *Annual Review of Economics*, 14:821–851, 2022.

[5] Vicente Valentim. Parliamentary representation and the normalization of radical right support. *Comparative political studies*, 54(14):2475–2511, 2021.

[6] Matias D Cattaneo, Nicolás Idrobo, and Rocío Titiunik. A practical introduction to regression discontinuity designs: Extensions. *arXiv preprint arXiv:2301.08958*, 2023.

[7] Guido W. Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008. The regression discontinuity design: Theory and applications.

[8] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[9] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[10] Wilbert Klaauw, Jinyong Hahn, and Petra Todd. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69:201–09, 02 2001.

[11] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

[12] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

[13] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Routledge, 2018.

[14] Keming Yu and MC Jones. A comparison of local constant and local linear regression quantile estimators. *Computational statistics & data analysis*, 25(2):159–166, 1997.

[15] B. Hansen. *Econometrics*. Princeton University Press, 2022.

[16] Langat Reuben Cheruiyot. Local linear regression estimator on the boundary correction in nonparametric regression estimation. *J. Stat. Theory Appl*, 19(3):460–471, 2020.

[17] Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020.

[18] David E Card, David Lee, Zhuan Pei, and Andrea Weber. *Local polynomial order in regression discontinuity designs*. Brandeis Univ., Department of Economics, 2014.

A First Appendix

A.1 Identifying the ATE in FRD

Recall our regression:

$$Y_i = \alpha + \beta_i W_i + u_i,$$

which implies

$$\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x] = [\lim_{x\downarrow c}\mathbb{E}[\beta_i W_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[\beta_i W_i|X_i=x]] - [\lim_{x\downarrow c}\mathbb{E}[u_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[u_i|X_i=x]]$$

Recall local continuity assumption:

$$\mathbb{E}[Y(0)|X=x] \text{ and } \mathbb{E}[Y(1)|X=x], \text{ are continuous in } x \text{ at } c.$$

Case 1: Locally constant treatment effect

Locally constant treatment effect means $\beta_i = \beta$ in a neighborhood around c. Assume local continuity as before yields

$$\lim_{x\downarrow c}\mathbb{E}[\beta_i W_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[\beta_i W_i|X_i=x] = \beta[\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]]$$
$$= \beta[1-0] = \beta$$

Common treatment effect is identified by

$$\frac{\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x]}{\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x] - \lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]}$$

Note that the denominator is change in Pr(treatment) at cut off point, and is always non-zero because of known discontinuity of $\mathbb{E}[W_i|X_i=x]$ at c. For SRD, the denominator is just equal to one.

To nonparametrically identify a constant treatment effect at the cut off point, we need two assumptions:

1. Known discontinuity at the cut off point

$$\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x] \neq \lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]$$

We are also implicitly assuming existence of the limits, and a positive density for x in neighborhood containing c.

2. Local continuity at the cut off point

$$\lim_{x \downarrow c} \mathbb{E}[u_i | X_i = x] = \lim_{x \uparrow c} \mathbb{E}[u_i | X_i = x]$$

Since $\beta_i = \beta$ by assumption of the constant treatment effects, we do not need local continuity of β in x.

Case 2: Heterogeneous treatment effect

In addition to the assumptions from above, we also need local conditional independence. It requires that W_i to be independent of β_i conditional on x near c.

$$\begin{aligned}\lim_{x \downarrow c} \mathbb{E}[\beta_i W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[\beta_i W_i | X_i = x] &= \lim_{x \downarrow c} \mathbb{E}[\beta_i | X_i = x] \lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] \\ &\quad - \lim_{x \uparrow c} \mathbb{E}[\beta_i | X_i = x] \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x] \\ &= \beta[1 - 0] = \beta\end{aligned}$$

Average treatment effect is again identified by

$$\frac{\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x]}$$

If subjects self-select into treatment, or are selected for treatment on the basis of expected gain (i.e., as a function of the outcome variable) then conditional independence assumption may be violated.

B Second Appendix

B.1 Appendix 1: The derivation in non-parametric portion

B.1.1 Kernel Assumption

These three assumptions(16)(17)(18) are quite important especially when we derive the equation related to a kernel function

$$K(u) = K(-u) \tag{16}$$

$$0 \leq K(u) \leq \bar{K} < \infty \tag{17}$$

$$\int uk(u)du = 0 \; and \int k(u)du = 1 \; and \int |u|^r K(u)du < \infty for \; all \; positive \; r \tag{18}$$

B.1.2 Kernel density function derivation

Following is the kernel density function form and our goal is to derive it. Here we plan to derive the unnormalize uniform kernel.

$$\hat{f}_y(y) = \frac{1}{nh} * \sum k\left(\frac{Y_i - Y}{h}\right)$$

In the Equation above, n is the number of observations, and h denotes the bandwidth. In this case, we apply the kernel to estimate the probability density function of y, the idea behind it is quite similar to the histogram. Mathematically, we can derive as following.

$$\begin{aligned} f(x) &= F'(x) \\ &= \lim_{h\to 0} \frac{F(X+h) - F(X-h)}{2h} \\ &= \lim_{h\to 0} \frac{p(x-h \le X_i \le x+h)}{2h} \end{aligned}$$

$$\hat{f}_N(x;h) = \frac{1}{2nh} \sum_{i=1}^{n} 1_{x-h \le x_i \le x+h}$$

$$rewrite\ x-h \le X_i \le x+h = -1 \le \frac{X_i - x}{h} \le 1$$

$$\hat{f}_N(x;h) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{-1 \le \frac{X_i - x}{h} \le 1}$$

The result we obtain is "Uniform Kernel".

B.1.3 Local Linear Estimator Derivation

$$\min \sum K\left(\frac{X_i - X}{h}\right)(Y_i - \hat{m}(x))^2$$

we don't know we want to estimate m(x) by which order yet, so let's used Taylor expansion to approximate it.

$$\sum_{i=1}^{N} K\left(\frac{X_i - X}{h}\right)(Y_i - \sum_{j=0}^{P} \frac{m^{(j)}(x)}{j!}(X_i - x)^j)^2$$

Now, there is a *brilliant idea* , which taking

$$\frac{m^{(j)}(x)}{j!} = \beta_j$$

So, rewriting the (8) as follows.

$$\sum_{i=1}^{N} K\left(\frac{X_i - X}{h}\right)(Y_i - \sum_{j=0}^{P} \beta_j (X_i - x)^j)^2$$

Rewriting to matrix form.

$$x = \begin{bmatrix} 1 & X_1 - x & \cdots & (X_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - X & \cdots & (X_n - x)^p \end{bmatrix}$$

$$W = diag(K\left(\frac{X_1 - X}{h}\right), K\left(\frac{X_2 - X}{h}\right), \cdots, K\left(\frac{X_n - X}{h}\right) \text{ and } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Using the least squared method to estimate, we obtain the following result.

$$\hat{\beta} = \arg\min(Y - X\beta)W(Y - X\beta)' = (X'WX)^{-1}X'WY$$

B.1.4 Rule of Thumb derivation

For obtaining rule of thumb of optimal bandwidth, using in the Fan and Gijbels(1996), we can derive like following. First, the number of (R(k)/4)^1/5, is very close for each kernel, and here we assume it as 0.58.

Second, for equation (10), we can re-write as the following form.

$$\begin{aligned} \bar{B} &= E[B_p(x)^2 w(x)] \\ &= E[B_p(x)^2 1[\phi_1 \leq x \leq \phi_2]] \\ &= \frac{1}{n}\sum_{i=1}^{n}[B_p(x_i)^2]1[\phi_1 \leq x_i \leq \phi_2] \end{aligned}$$

$$\textit{replacing } m''(x) \textit{ to } \hat{m}''(x) \textit{ in } B_p, \textit{then obtaining } \hat{B}$$

Then assuming homoskedastic for the error term,

$$\bar{\sigma}^2 = \sigma(\phi_2 - \phi_1)$$

$$\textit{replacing } \sigma \textit{ to } \hat{\sigma}$$

Finally, the rule of thumb would be like the following form.

$$h_{rule\ of\ thumb} = 0.58[\frac{\hat{\sigma}(\phi_2 - \phi_1)}{n\hat{B}}]^{\frac{1}{5}}$$

B.1.5 Derivation of Asymptotical Bias of Local Constant Estimator

First, derive the expected value of the local constant estimator(NW).

$$\begin{aligned} E[\hat{m}_{NW}(x)|X] &= \frac{\sum K(\frac{X_i - X}{h}) * E(Y_i|x)}{\sum K(\frac{X_i - X}{h})} \\ &= \frac{\sum K(\frac{X_i - X}{h}) * m(x)}{\sum K(\frac{X_i - X}{h})} \\ &= m(x) + \frac{\hat{b(x)}}{\hat{f(x)}} \end{aligned}$$

Following, define the numerator of the bias, b hat.

$$\hat{b(x)} = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - X}{h}\right)(m(x_i) - m(x))$$

Calculate the expected value of the b hat.

$$\begin{aligned} E[\hat{b}(x)] &= \frac{1}{h}E[K\left(\frac{X - x}{h}\right)(m(X) - m(x))] \\ &= \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{v - x}{h}\right)(m(v) - m(x) f(v) dv \\ &= \int_{-\infty}^{\infty} \frac{1}{h} K(u)(m(x + hu) - m(x)) f(x + hu) du \end{aligned}$$

By Taylor Expansion, we have the following equation

$$\begin{aligned} m(x + hu) - m(x) &= m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2) \\ f(x + hu) &= f(x) + f'(x)hu + o(h) \end{aligned} \tag{19}$$

Plugging in. Note that the third moment of the kernel is zero because of the symmetric at 0 assumption, the mean of the kernel is also zero because of the same reason.

$$
\begin{aligned}
E[\hat{b}(x)] &= \int_{-\infty}^{\infty} K(u)(m'(x)hu + \frac{1}{2}m''(x)h^2u^2 + o(h^2))(f(x) + f'(x)hu + o(h))du \\
&= h(\int_{-\infty}^{\infty} uK(u)du)m'(x)(f(x) + o(h)) \\
&+ h^2(\int_{-\infty}^{\infty} u^2K(u)du)(\frac{1}{2}m''(x)f(x) + m'(x)f'(x)) \\
&+ h^3(\int_{-\infty}^{\infty} u^3K(u)du)\frac{1}{2}m''(x)f'(x) + o(h^2) \\
&= h^2\mu_2(K)\left(\frac{1}{2}m''(x)f(x) + m'(x)f'(x)\right) + o(h^2)
\end{aligned}
$$

Afterward, calculate the variance of the b hat. Firstly, use the formula of variance $Var(x) = E(x^2) - [E(x)]^2$ to derive the first equation.

$$
\begin{aligned}
\frac{1}{nh^2}var[K\left(\frac{X-x}{h}\right)(m(X) - m(x))] &\leq \frac{1}{nh^2}E[K\left(\frac{X-x}{h}\right)^2(m(X) - m(x))^2] \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} K(u)^2(m(x+hu) - m(x))^2 f(x+hu)du \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} u^2K(u)^2du(m'(x))^2f(x)(h^2 + o(1)) \qquad (20) \\
&\leq \frac{h}{n}\bar{K}(m'(x))^2f(x) + o(\frac{h}{n})
\end{aligned}
$$

In line (20), we also use the first term of the Taylor expansion result(19). By the assumption of the kernel, we know that the integral term is bounded. Also by the assumption(6) on page 9, h/n would converge to zero. So h/n is the convergence rate. Finally, we can conclude as follows

$$
\frac{\hat{b}(x)}{\hat{f}(x)} = h^2B_0(x) + o_p(h^2) + O_p(\sqrt{\frac{h}{n}})
$$

B.1.6 Derivation of asymptotical variance of local constant estimator

$$
nhvar[\hat{m}_{nw}|X] = \frac{\hat{v}(x)}{\hat{f}(x)^2}
$$

$$
\hat{v}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)^2\sigma^2(X_i)
$$

$$
\begin{aligned}
E[\hat{v}(x)] &= \int_{-\infty}^{+\infty} \frac{1}{h} K\left(\frac{v-x}{h}\right)^2 \sigma^2(v) f(v) dv \\
&= \int_{-\infty}^{+\infty} K(u)^2 \sigma^2(x+hu) f(x+hu) du \\
&= \int_{-\infty}^{+\infty} K(u)^2 du \sigma^2(x) f(x) \\
&= R_k \sigma^2(x) f(x)
\end{aligned}
$$

Again, calculating the variance of v hat. The inequality also results from the same variance formula.

$$
\begin{aligned}
nhvar[\hat{v}(x)] &= \frac{1}{h}[K\left(\frac{X-x}{h}\right)^2 \sigma^2(x)] \\
&\leq \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{v-x}{h}\right)^4 \sigma^4(v) f(v) dv \\
&= \int_{-\infty}^{\infty} K(u)^4 \sigma^4(x+hu) f(x+hu) du \\
&\leq \bar{K} \sigma^4(x) f(x)
\end{aligned}
$$

Divided nh at both sides, so Var of v hat tends to be zero. Afterward, We can simply divide the expected value of v hat by the nh and square of f(x), then we get the result.

B.1.7 Five data generating process

$Cubic : Y = 10 - X^3 + \varepsilon \; where \; \varepsilon \sim N(0, 0.5) and X \sim (1,2)$

$Bump : Y = 1 + 2(X - 0.5) + (e^{-200(X-0.5)^2}) + \varepsilon \; where \; \varepsilon \sim N(0, 0.5) and X \sim U(1,2)$

$Quadratic : Y = 1 + 2(X - 0.5)^2 + \varepsilon \; where \; \varepsilon \sim N(0, 0.05), and X \sim U(1,2)$

$Linear : Y = 1 + 2(X - 0.5) + \varepsilon \; where \; \varepsilon \sim N(0, 0.05), and X \sim U(1,2)$

$Exponential : Y = e^{-4X} + \varepsilon \; where \; \varepsilon \sim N(0, 0.0015), and X \sim U(1,2)$

B.1.8 Definition of normalized kernel function

A normalized need to satisfy the following equation

$$
\int_{-\infty}^{\infty} u^2 K(u) du = 1
$$

B.1.9 Derivation of expected value of CV estimator

$$\begin{aligned} E[CV(h)] &= E[(Y_i - \hat{\mu}(X_i))^2] \\ &= E[(\mu(X_i) - \hat{\mu}(X_i) + e_i)^2] \\ &= E[e_i^2] + E[Q(x,h)] + 2E[(\mu(X_i) - \hat{\mu}(X_i))e_i] \\ &= \bar{\sigma}^2 + E[Q(x,h)] \\ &= \bar{\sigma}^2 + \int Q(x,h) f(x) dx \end{aligned}$$

C Third Appendix

C.1 the MSE results of FRD

Table 7: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF1)	p=1	p=2	p=3	p=4
triangular	0.1083	0.2039	0.3065	0.4669
epanechnikov	0.0947	0.1983	0.2977	0 4332
uniform	0.1119	0.1751	0.2869	0.4578

Table 8: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF2)	p=1	p=2	p=3	p=4
triangular	6.2078	4.7805	8.4888	11.3297
epanechnikov	6.9701	4.9845	7.5314	11.7262
uniform	7.4143	4.8039	7.5876	10.3298

Table 9: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF3)	p=1	p=2	p=3	p=4
triangular	79.6511	98.1870	129.2653	191.6147
epanechnikov	79.0705	90.9325	126.3930	201.4352
uniform	87.7994	98.7065	124.0248	159.7014

Table 10: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF4)	p=1	p=2	p=3	p=4
triangular	545.5236	864.1829	891.3165	1070.126
epanechnikov	546.8003	764.9351	844.1312	1197.971
uniform	663.6751	900.9801	1001.1624	1074.778

C.2 the MSE results when a = 20 for SRD and FRD

Table 11: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF1)	p=1	p=2	p=3	p=4
triangular	0.0722	0.1359	0.1771	0.2499
epanechnikov	0.0767	0.1297	0.1825	0.2429
uniform	0.0715	0.1246	0.1668	0.2355

Table 12: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF2)	p=1	p=2	p=3	p=4
triangular	5.2805	2.9026	4.8930	7.1363
epanechnikov	5.8193	3.3600	4.3482	5.9305
uniform	5.4126	3.1281	4.1048	6.1033

Table 13: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF3)	p=1	p=2	p=3	p=4
triangular	94.3993	104.6861	71.7033	114.1858
epanechnikov	97.7605	115.2860	69.5187	109.8065
uniform	100.71329	103.1097	68.8181	102.8610

Table 14: **the MSE of kernels in different p of rdrobust function(SRD)**

MSE(DGF4)	p=1	p=2	p=3	p=4
$triangular$	1213.284	956.7328	923.4224	718.4224
$epanechnikov$	1055.667	1001.7473	869.6841	777.3236
$uniform$	1181.225	1040.6569	942.3729	927.3274

Table 15: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF1)	p=1	p=2	p=3	p=4
$triangular$	0.1153	0.1991	0.2973	0.5017
$epanechnikov$	0.1085	0.1814	0.2867	0.4841
$uniform$	0.1111	0.1820	0.2783	0.4189

Table 16: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF2)	p=1	p=2	p=3	p=4
$triangular$	30.8058	5.3002	8.1453	10.5779
$epanechnikov$	29.4460	4.7243	7.2141	10.4827
$uniform$	33.8962	4.8612	7.2323	10.0473

Table 17: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF3)	p=1	p=2	p=3	p=4
$triangular$	298.2870	310.8133	126.9622	179.1860
$epanechnikov$	259.1455	298.8564	116.5592	180.1750
$uniform$	238.2336	311.8980	130.2226	179.7539

Table 18: **the MSE of kernels in different p of rdrobust function(FRD)**

mse(DGF4)	p=1	p=2	p=3	p=4
$triangular$	2704.144	8101.873	5490.575	1102.049
$epanechnikov$	2231.187	6533.730	2919.747	1218.071
$uniform$	1384.077	4129.420	2679.138	1645.742

C.3 the MSE results when a = 20 for FRD in different n

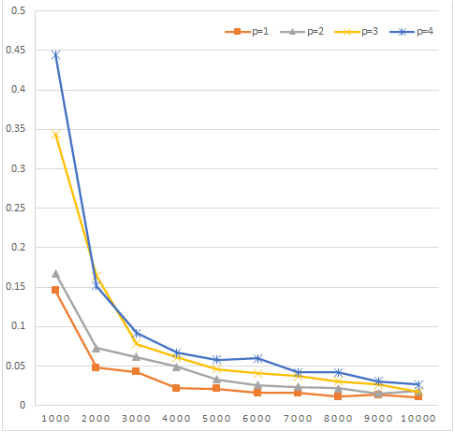

Figure 7: the MSE of DGF1 in FRD(a =20)

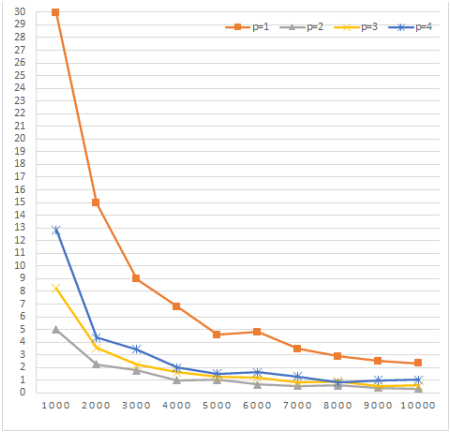

Figure 8: the MSE of DGF2 in FRD(a =20)

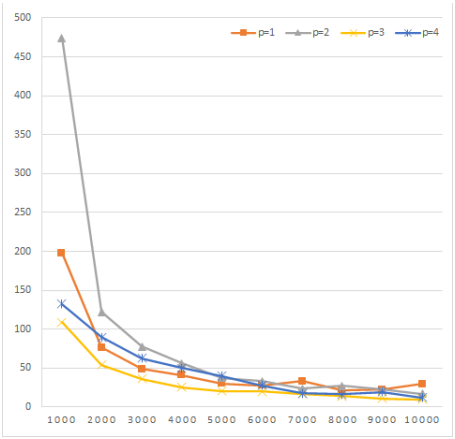

Figure 9: the MSE of DGF3 in FRD(a =20)

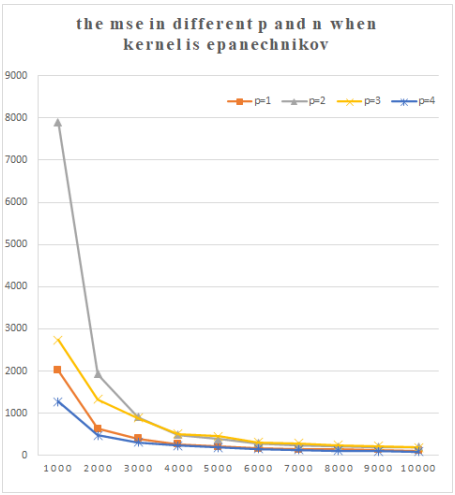

Figure 10: the MSE of DGF4 in FRD(a =20)

C.4 the best order of polynomial in FRD

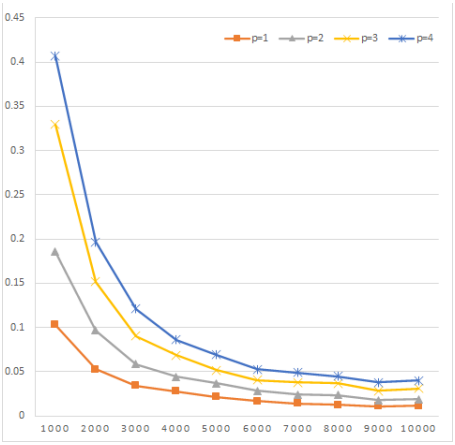

Figure 11: the MSE of DGF1 in FRD

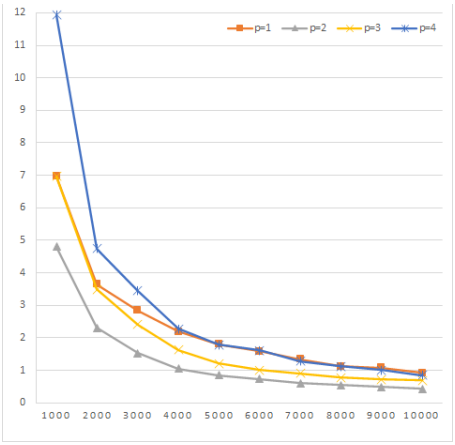

Figure 12: the MSE of DGF2 in FRD

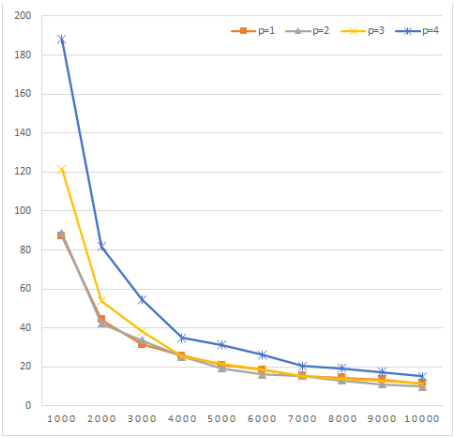

Figure 13: the MSE of DGF3 in FRD

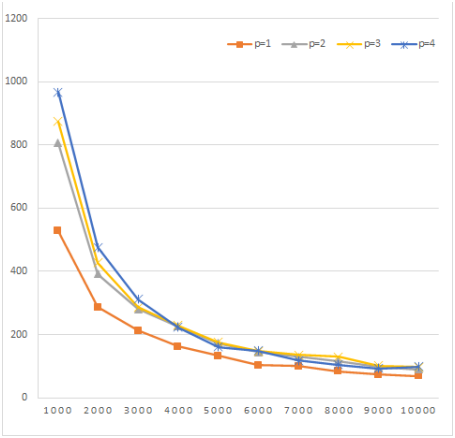

Figure 14: the MSE of DGF4 in FRD