Simulation work of RDD

Qianyi Wang

August 22, 2024

Abstract

Under the setting of sharp RDD, we want to estimate the treatment effect by using parametric model or non-parametric model. In this project, we will test the behaviors of 3 models–parametric model, local linear regression, and double robust model–under computer generated data and mortgage-CPI data. In the research, we will read papers on the regression discontinuity design and mechanisms that have been made to analyze the data, and try to develop an algorithm of our own.

1 Introduction

1.1 Background

Average treatment effect estimation is a crucial problem in causal inference, and has been the topic of a considerable amount of recent literature (Bradic, Wager, and Zhu, 2019, Imbens and Rubin, 2015). Regression discontinuity designs are a popular approach to causal inference that rely on known, discontinuous treatment assignment mechanisms to identify causal effects. (Hahn,Todd, and van der Klaauw, 2001, Imbens and Lemieux, 2008, Thistlethwaite and Campbell,1960, Eckles, Ignatiadis, Wager, and Wu, 2023). The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor (the covariate xi) being on either side of a fixed threshold (Imbens and Lemieux, 2008). In RDD, we assume there is a running variable and a cutoff(threshold), such that if the running variable is above the cutoff, we regard it as "assigned treatment", vice versa (Villamizar-Villegas, Pinzon-Puerto, and Ruiz-Sanchez, 2021). Since we do not take the treatment assignment to be random, approaches in random controlled trials, such as IPW, do not apply to RDD (Rubin, 08, Wager, 2022). This means we have to develop new algorithm to do the estimation.

1.2 Settings and Features

- We are interested in the effect of a binary treatment T_i on a real-valued outcome Y_i, and posit potential outcomes $\{Y_i(0), Y_i(1)\}$ such that $Y_i = Y_i(T_i)$.
- Unlike randomized trial, we do not take the treatment assignment T_i to be random. We assume there is a continuous variable $Z_i \in \mathbb{R}$ and know the cutoff or threshold c. Z_i determines who gets treatment, denoted by $T_i = 1$. By convention, Z is called the running variable.
- In **sharp RDD**, a unit is treated ($T_i = 1$) if $Z_i \geq c$ and not treated ($T_i = 0$) if $Z_i < c$. Than is, T_i is a deterministic function of Z_i: $T_i = f(Z_i)$. The running variable completely determines who gets treatment. We mainly focus on **sharp RDD** in this study.
- In fuzzy RDD, $f(Z_i)$ is not a deterministic function of Z_i; other variables affect treatment assignment, some of them could be unobserved.

1.3 Identification

One assumption of RDD is that it requires the continuity of Z_i for identification. If $\mathbb{E}[Y_i(1)|Z_i]$ and $\mathbb{E}[Y_i(0)|Z_i]$ are both continuous, we can identify the conditional average treatment effect at $z = c$, i.e., $\theta_c = \mathbb{E}[Y_i(1)|c] - \mathbb{E}[Y_i(0)|c]$ via $\theta_c = \lim_{z \to c^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to c^-} \mathbb{E}[Y_i|Z_i = z]$, provided that the

running variable Z_i has support around the cutoff c. To be more clear, we identify θ_c as the difference between the end points of two different regression curves.

1.4 Outline of the study

The rest of the study is organized as follows. First, we will introduce the parametric models and a strategy to select model. Second, we will introduce the local linear regression model and a strategy to find an optimal "neighborhood" of the local linear regression. Third, we will define a double machine learning model. Finally, we will do simulation on computer generated data and real world data and make some comments.

2 Global Linear Regression

2.1 Definition

This strategy uses every observation in the sample to model the outcome as a function of the rating variable and treatment status. This method "borrows strength" from observations far from the cut-point score to estimate the average outcome for observations near the cut-point score. To minimize bias, different functional forms for the rating variable — including the simplest linear form, quadratic, cubic, as well as its interactions with treatment — are tested by conducting F-tests on higher-order interaction terms and inspecting the residuals. This approach conceptualizes the estimation of treatment effects as a "discontinuity at the cut-point."

2.2 Setup

Consider the following model:

$$Y_i = \beta_0 + \theta_c \cdot T_i + f(Z_i, X_i) + \epsilon_i \tag{1}$$

where:

β_0 = the average value of the outcome for those in the treatment group after controlling for the running variable;

Y_i = potential outcome of observation i;

$T_i = 1$ if observation i is assigned to the treatment group, 0 otherwise;

Z_i = running variable, centered at the cutoff c;

X_i = covariates;

θ_c = treatment effect parameter;

ϵ_i = a random error term for observation i, which is assumed to be i.i.d.

2.3 Choices of $f(Z_i, X_i)$

- linear $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \gamma \cdot X_i + \epsilon_i$
- linear interaction $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i \cdot T_i + \gamma \cdot X_i + \epsilon_i$
- quadratic $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i^2 + \gamma \cdot X_i + \epsilon_i$
- quadratic interaction $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i^2 + \beta_3 \cdot Z_i \cdot T_i + \beta_4 \cdot Z_i^2 \cdot T_i + \gamma \cdot X_i + \epsilon_i$
- cubic $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i^2 + \beta_3 \cdot Z_i^3 + \gamma \cdot X_i + \epsilon_i$
- cubic interaction $Y_i = \beta_0 + \theta_c \cdot T_i + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i^2 + \beta_3 \cdot Z_i^3 + \beta_4 \cdot Z_i \cdot T_i + \beta_5 \cdot Z_i^2 \cdot T_i + \beta_6 \cdot Z_i^3 \cdot T_i + \gamma \cdot X_i + \epsilon_i$

2.4 Challenges and Solution

Since we have many functional forms, making an optimal choice is one of the greatest challenges for this approach to estimation. Several strategies have been proposed. In this study, we mainly focus on F-test Approach.

F-Test Approach Lee and Lemieux (2010) suggested testing the set of the models above against the data that underlie the initial plot of the rating versus the outcomes. The algorithms are following steps:

1. Create k indicator variables for $k-2$ of the bins used to graphically depict the data. Exclude any two of the bins to avoid having a model that is colinear.
2. Run regression 1 using one of the model above using $k-2$ bins of data.
3. Run regression 2, which use the same model as regression 1, but also includes the bin indicator variables created in step 1.
4. Obtain R-squared values from each of the two regressions: R_u^2 from regression 2, and R_r^2 from regression 1.
5. Calculate an F statistic using the following formula:

$$Fstatistic = \frac{\frac{R_u^2 - R_r^2}{k}}{\frac{1-R_u^2}{n-k-1}}$$

   where n is the sample size.

6. A p-value corresponding to this F statistic can be obtained using the degrees of freedom k and $n-k-1$. If the resulting F statistic is not statistically significant, the data from each of the bins are not adding any additional information to the model. This indicates that the model being tested is not underspecified and therefore is not oversmoothing the data.

In the study, we start with the linear model. If the F-test for the linear model versus a model with the bin indicators is not statistically significant, it implies that the linear functional form adequately depicts the relationship between the outcome and the running variables and therefore can serve as an appropriate choice for the RD estimation model. If, however, the F-test indicates oversmoothing of the data, a higher-order term (and its interaction with treatment indicator) needs to be added to the functional form and a new F-test carried out on this higher-order polynomial model. The idea is to keep adding higher-order terms to the polynomial until the F-test is no longer statistically significant.

Principle of F test F-test is testing whether or not there is unexplained variability in the relationship between the outcome and rating that the specified model is not capturing; in other words, whether something is missed from the model.

3 Local Linear Regression

3.1 Definition

This strategy views the estimation of treatment effects as local randomization and limits the analysis to observations that lie within the close vicinity of the cut-point (bandwidth), where the functional form is more likely to be close to linear. Once the bandwidth is selected, a linear regression is estimated, using observations within one bandwidth on either side of the threshold.

3.2 Setup

We pick a small bandwidth $h_n \to 0$ and a symmetric weighting function $K(\cdot)$, and the fit $\mathbb{E}[Y_i(t)|Z_i]$ via weighted linear regression on each side of the boundary,

$$\theta_c = argmin\{\sum_{i=1}^{n} K(\frac{|Z_i - c|}{h_n}) \times (Y_i - \beta_0 - \beta_1 \cdot (Z_i - c)_- - \beta_2 \cdot (Z_i - c)_+ - \gamma \cdot X_i)^2\} \quad (2)$$

where the β_0, β_1, and β_2 are nuisance parameters. In this study, we choose the triangular kernel $K(x) = (1 - |x|)_+$ for weighting function $K(\cdot)$.

3.3 Challenges and solution

In general, choosing a bandwidth in nonparametric estimation involves finding an optimal balance between precision and bias: using a larger bandwidth gives us more precise estimates since more data are included in the estimation; but this can also lead to overfitting and more bias. Thus, choosing an optimal bandwidth for out local linear regression model is crucial in both empirical study and real world data study. In this study, we mainly focus on cross-validation procedure to choose the bandwidth.

Cross-validation Ludwig and Miller (2005) and Imbens and Lemieux (2008) have proposed a version of the "leave-one-out" cross-validation procedure that is tailored for RDD. This cross-validation procedure can be carried out as following steps:

1. Select a bandwidth h_1.
2. start with an observation A to the left of the cutoff c, with running variable Z_A and an outcome Y_A.
3. To see how well the parametric assumption fits the data within the bandwidth h_1, run a regression of the outcome on the running variable using all of the observations that are located to the left of observation A and have a running variable that ranges from $Z_A - h_1$ to Z_A(not included).
4. Get the predicted value of the outcome variable observation A based on this regression and call this predicted value $\hat{Y}_A$
5. Shift the "band" slightly over to the left and repeat this process to obtain predicted values for observation B. Repeat this process to obtain predicted values for all observations to the left of the cutoff c.
6. Then repeat this process to obtain predicted values for all observations to the right of the cutoff; stop when there are fewer than two observations between $Z_i - h_1$ to Z_i.
7. Calculate the mean square error for bandwidth h_1 using the following formula:

$$MSE(h_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

   where n is the total number of observations in the data set and all other variables are as defined before.
8. Repeat the above steps for other bandwidth choices $h_2, h_3, ...$
9. Pick the bandwidth that produces the smallest mean square error(MSE).

4 DML Partial Linear Regression

4.1 Definition

Chernozhukov et al.(2018) combined machine learning techniques with partial linear model, introducing a robust methodology for estimating treatment parameters in high-dimensional settings. In this study, I want to use this method as a reference for global estimation method. Zhu and Jiang (2014) have delved into partial linear regression inference framework for RDD and proved the plausibility of their method. In this study, I want to choose a different $g(x)$ and do empirical study and real world data study.

4.2 Setup

The partial linear model is a semi-parametric model that combines a linear component with a non-parametric component:

$$Y_i = \beta_0 + \theta_c \cdot T_i + g(Z_i, X_i) + \epsilon_i \tag{3}$$

$$T_i = m(Z_i, X_i) + \xi_i \tag{4}$$

where running variable and covariates affect T_i via $m(Z_i, X_i)$ and the outcome via the function $g(Z_i, X_i)$ and m and g are nuisance parameters. In this study, we use l_1-penalization based method, labeled "Lasso" to report the results.

4.3 Challenges and Solution

4.3.1 Regularization Bias

Suppose an estimation of θ_c by making use of a sample split. This sample split consists of two parts, a main part with n observations, these are indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. Given that $\hat{g}$ is obtained using the auxiliary sample, the final estimate of θ_c is obtained using the main sample:

$$\hat{\theta}_c = (\frac{1}{n}\sum_{i \in I} T_i^2)^{-1} \frac{1}{n} \sum_{i \in I} T_i (Y_i - \hat{g}(Z_i, X_i)) \tag{5}$$

The estimator $\hat{\theta}_c$ will generally have a slower than $\frac{1}{\sqrt{n}}$ rate of convergence, namely

$$|\sqrt{n}(\hat{\theta}_c - \theta_c)| \to_P \infty \tag{6}$$

This is caused by the bias in learning m. To overcome this problem, Chernozhukov et al. (2018) recommended the use of orthogonalization. First, the effect of X and Z from T is partialling out. The result variable is $\hat{V} = T - m(Z, X)$. The last term is an machine learning estimator of $m(Z, X)$ and is obtained using the auxiliary sample. g is also estimated using the auxiliary sample, which is the same as the original approach. Using the main sample of observations, this results in the following estimator for θ_c:

$$\check{\theta}_c = (\frac{1}{n}\sum_{i \in I} \hat{V}_i T_i)^{-1} \frac{1}{n} \sum_{i \in I} V_i (Y_i - \hat{g}(Z_i, X_i)) \tag{7}$$

A generalization of the orhogonalization principle is given by the Neyman orthogonality and moment conditions. The estimator $\check{\theta}_c$ can be viewed as a solution to

$$\partial_\eta \, \mathbb{E} \, \psi(W; \theta_c, \eta_0)[\eta - \eta_0] = 0 \tag{8}$$

This equation is called the Neyman orthogonality and ψ is the Neyman orthogonal moment function. This orthogonality condition means that the used momennt conditions to identify the parameter of interest are locally not sensitive to the value of the covariate. Because of this, one can plug in noisy estimations of these covariates without violating the moment conditions Chernozhukov et al. (2017).

4.3.2 Overfitting Bias

Besides the regularization bias, there is also a bias induced by overfitting. This bias can be removed by using sample splitting. Sample splitting ensures that terms like

$$\frac{1}{n} \sum_{i \in I} V_i (\hat{g}(Z_i, X_i) - g(Z_i, X_i)) \tag{9}$$

in partial linear model vanishes in probability. Although, when estimating the parameter of interest using sample splitting, only the main sample is used. This could result in an extraordinary loss of efficiency, because only a part of the available data is used. They show that this loss can be removed by flipping the role of the main- and auxiliary sample, this results in a second estimator of the treatment variable of interest. It may regain full efficiency, when averaging these two resulting estimators. They call this procedure where they flip the roles of the auxiliary- and main sample to obtain more estimations and thereafter take the average of the results cross-fitting.

4.4 Machine Learning Method

LASSO
Least Absolute Selection and Shrinkage Operator Tibshirani(1996) is a machine learning method that can be applied when having a linear regression with many regressors. It minimizes the sum of squared residuals with an additional term:

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda \cdot \|\beta\| \tag{10}$$

where $\|\beta\| = (\sum_{k=1}^{K} |\beta_k|)$. Rewrite (10) as follows to choose the penalty parameter

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 s.t. \sum_{k=1}^{K} |\beta_k| \leq t \cdot \sum_{k=1}^{K} |\beta_k^{ols}| \tag{11}$$

where t is a scalar between zero and one. When t equals zero, it is easy to see that all estimates shrinks to zero. When t is equal to one, the estimates are not shrinking and it is just OLS. The penalty parameter λ in (10) or t in (11) are chosen through cross-validation.

5 Simulation study

In this section, we evaluate the performance of the estimators in each model with respect to estimation. we fix $n = 1200$ and $c = 1.9$. We generate data for a set of variables as follows:

- $Z \sim$ Exponential(0.5)
- $X_1 \sim$ Exponential(3)
- $X_2 \sim \mathcal{N}(10, 2^2)$
- $X_3 \sim \mathcal{N}(5, 0.5^2)$

Let $\theta^c = 5$.
Define T as:

$$T = \begin{cases} 1 & \text{if } Z \geq c \\ 0 & \text{otherwise} \end{cases}$$

Define Y as:

$$\begin{aligned} Y = \theta \cdot T + 2 \cdot Z + 1.5 \cdot X_1 - 2 \cdot X_2 + 3 \cdot X_3 \\ + 1.2 \cdot Z^2 - 0.5 \cdot X_1^2 + 0.8 \cdot X_2 \cdot X_3 \\ + 0.3 \cdot T \cdot Z + 0.7 \cdot \sin(Z) + 0.4 \cdot \log(|X_1| + 1) \\ + \mathcal{N}(0, 1) \end{aligned}$$

The data frame is constructed as follows:

$$\text{data} = \{Y, T, Z, X_1, X_2, X_3\}$$

5.1 Global Linear Model

Table 1 shows linear regression estimations of the relationship between T and Y. The outcome variable is the estimated average treatment effect and the corresponding p-value of F test. We can see that all linear models are valid in estimating treatment effect. But among all the models, cubic model give the most accurate average treatment effect.

Model	TE	p.value
Linear	-10.800502	0.5393904
Linear Interaction	-22.999497	0.1713742
Quadratic	5.262242	0.7969893
Quadratic Interaction	6.312967	0.8082864
Cubic	5.232049	0.7936150
Cubic Interaction	8.518507	0.8474581

Table 1: Performance of Linear Models

5.2 Local Linear Regression

Table 2 shows the result of the choices of bandwidth and their corresponding estimated treatment effect. We can find that if the bandwidth is greater than 0.5, the estimation would be very accurate.

Bandwidth	N	MSE	TE
0.1	1200	2.154797	4.290528
0.5	1200	1.849058	5.195666
1.0	1200	1.842228	5.460394
1.7	1200	1.873413	5.707292
2.5	1200	1.882745	5.674709

Table 2: Performance of Different Bandwidth Selection

5.3 DML Partial Linear Regression

Table 3 shows the result of the double machine learning partial linear regression. We also add two result of LASSO to compare the performance of LASSO in double machine learning model. The result of DML partial linear regression is much more accurate than using LASSO directly.

Model	TE
Lowest Mean Cross-Validated Error	-10.341107
One Standard Error	-2.931274
DML	4.996066

Table 3: Performance of Different Model Selections

5.4 Results

Table 4 shows the comparison among all models in this study. We can find that the estimator given by DML Partial Linear model is the closest to the real θ_c. Thus, we can make a conclusion that DML Partial Linear model is give the most accurate estimation in this study.

Model	TE
Linear	-10.800502
Linear Interaction	-22.999497
Quadratic	5.262242
Quadratic Interaction	6.312967
Cubic	5.232049
Cubic Interaction	8.518507
Local Linear Regression	5.460394
DML Partial Linear	4.996066

Table 4: Comparison among All Models

6 Real world data

In this section, we apply the models in this study to derive a 30-Year fixed rate mortgage average prediction model for average CPI for all urban consumers. In this study, the covariates are unemployment rate, inflation rate and annual saving rate. There are $n = 54$ observations from 1971 to 2024. We employ the same method for this prediction model.

Model	**TE**
Linear	-87.38973
Linear Interaction	-238.43901
Quadratic	-32.19375
Quadratic Interaction	-610.70933
Cubic	-33.82197
Cubic Interaction	-593.12993
Local Linear Regression	503.13774
DML Partial Linear	-81.77153

Table 5: Performance of Different Models

The estimates of average treatment effect from all method are given in Table 5.

References

Bradic, J., Wager, S., Zhu, Y. (2019). Sparsity double robust inference of average treatment effects (arXiv:1905.00744). arXiv. http://arxiv.org/abs/1905.00744

Braid, R. M. (1984). The effects of government housing policies in a vintage filtering model. Journal of Urban Economics, 16(3), 272–296. https://doi.org/10.1016/0094-1190(84)90028-7

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Eckles, D., Ignatiadis, N., Wager, S., Wu, H. (2023). Noise-induced randomization in regression discontinuity designs (arXiv:2004.09458). arXiv. http://arxiv.org/abs/2004.09458

Hahn, J., Todd, P., Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1), 201–209. https://doi.org/10.1111/1468-0262.00183

Imbens, G., Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press.

Imbens, G. W., Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Lee, D. S., Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of Economic Literature, 48(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Ludwig, J., Miller, D. L. (2006). Does head start improve children's life chances? Evidence from a regression discontinuity design. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.900828

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. The Annals of Applied Statistics, 2(3). https://doi.org/10.1214/08-AOAS187

Thistlethwaite, D. L., Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51(6), 309–317. https://doi.org/10.1037/h0044319

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Villamizar-Villegas, M., Pinzon-Puerto, F. A., Ruiz-Sanchez, M. A. (2022). A comprehensive history of regression discontinuity designs: An empirical survey of the last 60 years. Journal of Economic Surveys, 36(4), 1130–1178. https://doi.org/10.1111/joes.12461

Zhu, R. J. B., Jiang, W. (2024). Semiparametric inference for regression-discontinuity designs (arXiv:2403.05803). arXiv. http://arxiv.org/abs/2403.05803