Extrapolation of Treatment Effects in Sharp RD Design: A Double Machine Learning Approach

Nan Ge

June 27, 2019

Abstract

Sharp rgression discontinuity (RD) design identifies the causal effect of the treatment at the treatment threshold. The treatment effect on inframarginal populations is of interest, but the identification of such treatment effect needs stronger assumptions. This paper first proves an implicit behavioral assumption in the conditional independence assumption (CIA) in Angrist and Rokkanen (2015), then leverages this assumption to develop a new double machine learning method which selects covariates to extrapolate the average treatment effect in sharp RD design. The new method improves the computational performance of extrapolation. The simulation results confirm that the new extrapolation method performs better. Two extensions of the method are presented in the final part.

1 Introduction

Regression discontinuity (RD) design is one of the most credible identification strategy in the absence of a randomized experiment (Cattaneo et al. 2019). It has been widely used in economics, political science, and public policy. Under relatively mild identification assumptions, i.e., the conditional expectation functions of potential outcomes are continuous around the treatment threshold and the existence of first moment of potential outcome functions, Hahn et al. (2001) shows that the treatment effect is identifiable under mild functional form assumptions. Therefore, in a non-compliant setting, the canonical RD parameter identifies the local average treatment effect (LATE). It has limited external validity because the sub-population near the threshold is likely to be very different from the general population that researchers are interested in. Therefore, a natural question is if we can identify the treatment effect away from the cutoff in a RD design.

The raw version of the extrapolation method used in this paper comes from Angrist and Rokkanen (2015). They extrapolate the treatment effect away from the cutoff in RD design by invoking the Conditional Independence Assumption (CIA): $E[Y_i(d)|R_i, x_i] = E[Y_i(d)|x_i]$, where $d \in \{0, 1\}$ and the Common Suport Assumption. In other words, we assume that potential outcomes are mean independent of running variable once we condition on the correct set of control covariates, x_i. They also prove that if the assumptions hold, the RD estimate should not be different from properly weighed matching estimates. Therefore, to implement the methodology in Angrist and Rokkanen (2015), it boils down to select the correct covariates x_i that make CIA works.

However, the variable selection task can be computationally demanding if we have a large dimension of covariates in the data set but no precise knowledge of the sources of omitted variable bias. Consider the following situation. Suppose that we have 100 covariates, and 3 of them generate both potential outcomes and running variable. Further, suppose that we are not aware of the data generating process of running variable. Then, a potential variable selection algorithm would be: given treatment or control group, first, we try to regress outcome variables on each covariate, which amounts to 100 regressions; second, we try to regress the outcome variable on two covariates, which amounts to $C_{100}^2 = 4950$; third, we try to regress the outcome variable on three covariates, which amounts to $C_{100}^3 = 485100$ regressions. This is not to include the regression tests we might need to examine CIA. As we could see, the dimension of controls is not rare in empirical studies, but the naive method is too computationally demanding to implement.

In this paper, I first justify an implicit behavioral assumption in CIA in Angrist and Rokkanen (2015). I'll show that sufficient conditions for CIA to work are: 1) we can model running variable as: $R_i = g(x_i, \varepsilon_{ri})$; 2) $\varepsilon_{ri} \perp\!\!\!\perp (x_i, Y_i(d))$, where $d \in \{0, 1\}$. Note that in Angrist and Rokkanen (2015), they briefly mentioned the first assumption but not the second one. And they did not discuss the consequences of allowing endogeneity of ε_{ri} exiplicitly in their paper. Considering that the second assumption is very strong, I also study the consequences of its violation by different sets of simulations. The simulation results show that if the second assumption is violated, even when we select the correct set of covariates, we cannot wipe out the effect of running variable on potential outcomes, i.e., the coefficient of running variable on outcome variables is statistically significant even if we have the right set of covaraites in our model. Therefore, the variable selection method in Angrist and Rokkanen (2015) will in general not work if there is endogeneity in DGP.

This paper is also built on Chernozhukov et al. (2018), as we will use the double machine learning (DML, also know as double lasso) method to select covariates in the extrapolation of treatment effect. In Chernozhukov et al. (2018), they used DML to identify LATE in an instrument variable setting, and proved that the method has very good inference properties. Specifically, the method works as follows: first, split the data into train and test set; second, use train set data to run following two regressions, regress Y_i on x_i by LASSO given treatment and control groups, and regress R_i on x_i by LASSO;

third, union the selected covariates in the second step to generate conditional set; fourth, use the conditional set to extrapolate the Average Treatment Effect (ATE) in sharp RD design. After getting the point estimates of extrapolated ATE, compare the estimates from the DML method to those in Angrist and Rokkanen method, and the single LASSO method.

Very often in RD designs, units would be assigned to treatment based on different running variables (Wong et al. 2013). Such cases are very common in education policies and labor policies. For example, Cattaneo et al. (2016) studies the ACESS program which provides tuition credits to underprivileged populations. Treatment is the same across the country, but the cutoff that determines treatment assignment varies widely by department and changes each year. Multiple-cutoff settings are also applicable in political science, when there are more than two candidates, and the margin differs across states (Cattaneo et al. 2016).

Given the broad application of multiple-cutoff design, it's very important to test whether the new method could be applied when there are more than one cutoff. Consider the simplest case where there are two cutoffs in the sharp RD design: $C = \{c_1, c_2\}$. As noted in Cattaneo et al. (2018), we can think of the multi-cutoff as different subpopulations that are different in terms of both observables and unobservables. In the multi-cutoff case, there could exist two types of CIA: $E[Y_{di}(c)|x_i, R_i, C_i = c] = E[Y_{di}(c)|x_i]$ and $E[Y_{di}(c)|x_i, R_i, C_i = c] = E[Y_{di}(c)|x_i, C_i = c]$. The new method can be easily extended to the multi-cutoffs case by implementing the method at each cutoff. The simulation results again confirm that the extrapolated ATE of DML approach has better prediction accuracy. In this aspect, this paper also extends Cattaneo et al. (2018) as their methodology can only deal with the situation where there are two cutoffs.

One critical drawback of Angrist and Rokkanen (2015) is that they did not specify the assumptions CIA imposed on DGP. However, as I will prove below, the sufficient conditions for CIA require exogeneity in DGP. This might not be a problem when we are very certain about the controls, but weakens the necessity for RD design. As an extension, I study the behavior of the new method when we allow for endogeneity. There are two findings from the simulation. First, even when we allow endogeneity in the DGP, DML approach can still select the correct covariates that generates the potential outcomes and running variables with high probability. Second, when we use selected covariates to extrapolate ATE, the extrapolated ATE tends to be over-estimated.

In addition, this paper also complements with the extrapolation method developed in Dong and Lewbel (2015), and a larger literature on RD design extrapolation without external information of the experiment units. In Dong and Lewbel (2015), their research question is to extrapolate the treatment effect when there is a marginal change in the cutoff. They managed to identify the Marginal Threshold Treatment Effect (MTTE), the derivative of the treatment effect with respect to the cutoff, with a nonparametric approach. To do the extrapolation, they make the local policy invariance assumption. However, this paper is mainly concerned about extrapolating the treatment effect when the treatment threshold does not change. Therefore, the local policy invariance assumption is not necessary.

The rest of the paper proceeds as follows. The second section proves the sufficient condition that make CIA works and lays out the method in detail. The third section develops several simulation units and presents the simulation results to compare different extrapolation methods. The fourth section extend the methods to the situation where there are multi-cutoffs. The assumption of exogeneity in the DGP is also relaxed in the fourth section. The final section concludes.

2 Method

2.1 Set-Up

Let us consider a canonical RD design. We have a running variable, R_i, and a binary treatment variable, D_i. Suppose we are interested in the outcome variable, Y_i. Then, potential outcomes can be defined accordingly: $Y_i(d)$, where $d \in \{0, 1\}$. Specifically, $Y_i(1)$ is the outcome of an individual if he or she receives the treatment $(D = 1)$, and $Y_i(0)$ is the outcome of an individual if he or she does not receive the treatment $(D = 0)$. Thus, the observed outcome of an individual is:

$$Y_i = (1 - D)Y_i(0) + DY_i(1).$$

Moreover, suppose the treatment is assigned by: $D_i = 1\{R_i \geq c\}$, i.e., individuals with running variable greater than c receive the treatment, otherwise, they do not receive the treatment.

In addition, we assume that researchers have a rich set of covariates, x_i, which is high-dimensional. This assumption is innocuous because it is very often that researchers have a rich set of predetermined covariates and include them to estimate RD effects (Cattaneo et al. 2019).

2.2 Identification and Extrapolation in Sharp RD Design

The following assumptions are required to guarantee the point identification of ATE at the treatment threshold:

Assumption 1. $f_{R_i}(r) > 0$ in a neighborhood around c.

Assumption 2. $E[Y_i(d)|R_i = r]$ is continuous in r at c, where $d \in \{0, 1\}$.

Assumption 3. $E[|Y_i(d)||R_i = c|] < \infty$, where $d \in \{0, 1\}$.

Lemma 1. (Hahn et al. 2001) Suppose Assumption 1 to 3 hold, then,

$$E[Y_i(1) - Y_i(0)|R_i = c] = \lim_{\epsilon \to 0}\{E[Y|R = c + \epsilon] - E[Y|R = c - \epsilon]\}$$

Lemma 1 shows that the ATE at the treatment threshold is identifiable given Assumption 1 to 3 hold. However, RD design does not allow researchers to learn effect of treatment that is not at the cutoff, $E[Y_i(1) - Y_i(0)|R_i = r]$ for $r \neq c$. Therefore, RD design has poor external validity.

To extrapolate the treatment effect away from the cutoff in sharp RD design, it boils down to extrapolate following two conditional expectation functions: $E[Y_i(1)|R_i = r]$, where $r < c$, and $E[Y_i(0)|R_i = r]$, where $r \geq c$.

One approach developed in Angrist and Rokkanen (2015) is to invoke Conditional Independence Assumption (CIA) and common support assumption to extrapolate the treatment effect. Suppose we have a set of covariates, x_i. And the following two assumptions hold:

Assumption 4. Conditional Independence Assumption (CIA)

$$E[Y_i(d)|R_i, x_i] = E[Y_i(d)|x_i];\ d = 0, 1.$$

Assumption 5. Common Support Assumption (CSA)

$$0 < P[D_i = 1|x_i] < 1\ a.s.$$

Assumption 4 (CIA) assumes that potential outcomes are mean independent of the running variable R_i, conditional on x_i. Assumption 5 (CSA) assumes that the treatment status varies conditional on x_i. Angrist and Rokkanen (2015) shows that Assumption 4 and Assumption 5 can assist us to identify any counterfactual expectations of potential outcomes.

The significance of CIA is that it has a testable implication that can assist us selecting useful covariates, x_i (Angrist and Rokkanen 2015). Specifically, if CIA works, we have:

$$\begin{aligned} E[Y_i(1)|R_i, x_i, R_i \geq c] &= E[Y_i(1)|x_i] \\ &= E[Y_i(1)|x_i, R_i \geq c] \end{aligned}$$

Therefore, we would expect that if x_i are the covariates make CIA works, then, we would observe following holds for the treated groups that are right of the cutoff:

$$E[Y_i|R_i, x_i, D_i = 1] = E[Y_i|x_i, D = 1] \qquad (1)$$

Similarly, we have the same type of implications for untreated groups that are left of the cutoff:

$$E[Y_i|R_i, x_i, D_i = 0] = E[Y_i|x_i, D = 0] \qquad (2)$$

Therefore, a simple variable selection procedure can be developed based on (1) and (2): we can regress outcome variable on x_i and R_i on either side of the cutoff, if we observe that the coefficient of R_i becomes insignificant when conditional on x_i, we have the evidence that x_i are the covariates that make CIA works, otherwise, we try another set of covariates and do the same test again. However, for reasons stated above, this procedure is too computationally demanding to implement.

2.3 Sufficient Condition that Makes CIA Works

As suggested by Lee and Card (2008), one interpretation of CIA in RD context is that we can model the running variable as a function of observables, x_i, and unobservables, ε_{ri}. Therefore, we can think of this is a selection equation. Formally, $R_i = g(x_i, \varepsilon_{ri})$, where g is an unknown function.

In addition, CIA implies that we can also model the potential outcomes as a function of observables, x_i, and unobservables, ε_{di}, where $d \in \{0, 1\}$. Formally, we have: $Y_i(d) = f_d(x_i, \varepsilon_{di})$, where $d \in \{0, 1\}$ and f_d is an unoknown function.

To make CIA hold, a sufficient condition will be: the unobservables in the selection stage, ε_{ri} is independent of x_i and potential outcomes $Y_i(d)$, where $d \in \{0, 1\}$. The condition has two implications. First, it can be interpreted as we are using exogeneous covariates, x_i as instruments for R_i. Second, we also assume that the unobservables in the selection stage, ε_{ri}, are independent of potential outcomes. Therefore, one implication from $\varepsilon_{ri} \perp\!\!\!\perp Y_i(d)$ is that ε_{ri} is independent of ε_{di}, where $d \in \{0, 1\}$. The economic interpretation of the assumption is that the unobservables that affect potential outcomes and running variables are independent of each other. This is also a strong behavioral assumption about the unobservables in the selection stage. Following lemma formalizes the idea we just discuss:

Lemma 2. Assume that the running variable can be modeled as $R_i = g(x_i, \varepsilon_{ri})$, where g is a measurable function. Also, assume that conditional on x_i, R_i and ε_{ri} are one-to-one function. Further, assume that $(Y_i(d), x_i) \perp\!\!\!\perp \varepsilon_{ri}$, then, CIA holds, i.e., $E[Y_i(d)|R_i, x_i] = E[Y_i(d)|x_i]$.

Proof.

$$\begin{aligned}
E[Y_i(d)|R_i, x_i] &= E[Y_i(d)|g(x_i, \varepsilon_{ri}), x_i] \\
&= E[Y_i(d)|x_i, \varepsilon_i] \\
&= E[Y_i(d)|x_i]
\end{aligned}$$

where the first equality uses the fact that $R_i = g(x_i, \varepsilon_i)$, the second equality uses the fact that R_i and ε_{ri} are one-to-one functions of each other given x_i, the final equality uses the assumption that $(Y_i(d), x_i) \perp\!\!\!\perp \varepsilon_{ri}$. ■

2.4 Consequences of the Violation of Lemma 2

Given the significance of the CIA in the variable selection procedure, a natural question to ask is that if $(Y_i(d), x_i) \perp\!\!\!\perp \varepsilon_{ri}$ fails to hold, then, what are the consequences for the procedure of selecting useful covariates in Angrist and Rokanen (2015). I assess the consequence by doing the following four sets of simulations.

In all of the simulations, I assume that the same set of covariates, i.e., x_{i1}, x_{i2}, and x_{i3}, generates potential outcomes and running variables, but allow different forms of endogeneities. Specifically, I assume that the DGPs are:

$$Y_i(0) = 0.8 + 2.3x_{i1} - 6.3x_{i2} - 15.6x_{i3} + \varepsilon_{0i}$$

$$Y_i(1) = 1.5 + 23.5x_{i1} + 1.6x_{i2} + 2.6x_{i3} + \varepsilon_{1i}$$

$$R_i = 1.6 + 6.8x_{i1} + 8.3x_{i2} + 7.6x_{i3} + \varepsilon_{ri}$$

Where the treatment is assigned as: $D_i = 1\{R_i > 0\}$. In all simulations, x_{i1}, x_{i2}, and x_{i3} have normal distributions with mean 0 and variance 1.

For each simulation, I repeat the following exercise for 1000 times. In each repetition, regress Y_i on R_i, x_{i1}, x_{i2}, and x_{i3} given $D_i = 0$. Then record the p-value of the coefficient of R_i. Similarly, repeat the practice when $D_i = 1$. Then draw a density plot and a histogram of the p-values of the coefficient of R_i for treatment and control groups.

As a benchmark, in the first simulation, I assume that 1) there are no correlations between observables and unobservables; 2) the error terms are also normally distributed with mean 0 and variance 1, and they are uncorrelated:

$$\begin{pmatrix} \varepsilon_{0i} \\ \varepsilon_{1i} \\ \varepsilon_{ri} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right]$$

Since in the true model, R_i is not in the DGP that generates potential outcomes, the distribution of the p-values in both groups should be roughly uniform. The simulation results are presented in the first row of Figure 1. Overall, the simulation results confirm the claim that the p-values in both groups are distributed uniformly.

In the second simulation, I allow endogeneity on unobservables. Specifically, I assume that:

$$\begin{pmatrix}\varepsilon_{0i}\\ \varepsilon_{1i}\\ \varepsilon_{ri}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0 & -0.3\\ 0 & 1 & 0.5\\ -0.3 & 0.5 & 1\end{pmatrix}\right]$$

The results are presented in the second row of Figure 1. It shows that across all 1000 repetitions, even if we include the true covariates in the model, the p-value of the coefficient to the running variable is in general smaller than 5%. Therefore, one implication from the simulation is that given the existence of the endogeneity of unobservables, the coefficient of the running variable is significant when we include all the correct covariates. Therefore, the variale selection procedure in Angrist and Rokkanen (2015) will not work when such type of endogeneity exists.

In the third simulation, I assume that there is endogeneity on observable, x_{i1}. Specifically, x_{i1} is endogeneous, and the joint distribution is:

$$\begin{pmatrix}x_{i1}\\ \varepsilon_{0i}\\ \varepsilon_{1i}\\ \varepsilon_{ri}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.2 & -0.2 & 0.3\\ 0.2 & 1 & 0 & 0\\ -0.2 & 0 & 1 & 0\\ 0.3 & 0 & 0 & 1\end{pmatrix}\right]$$

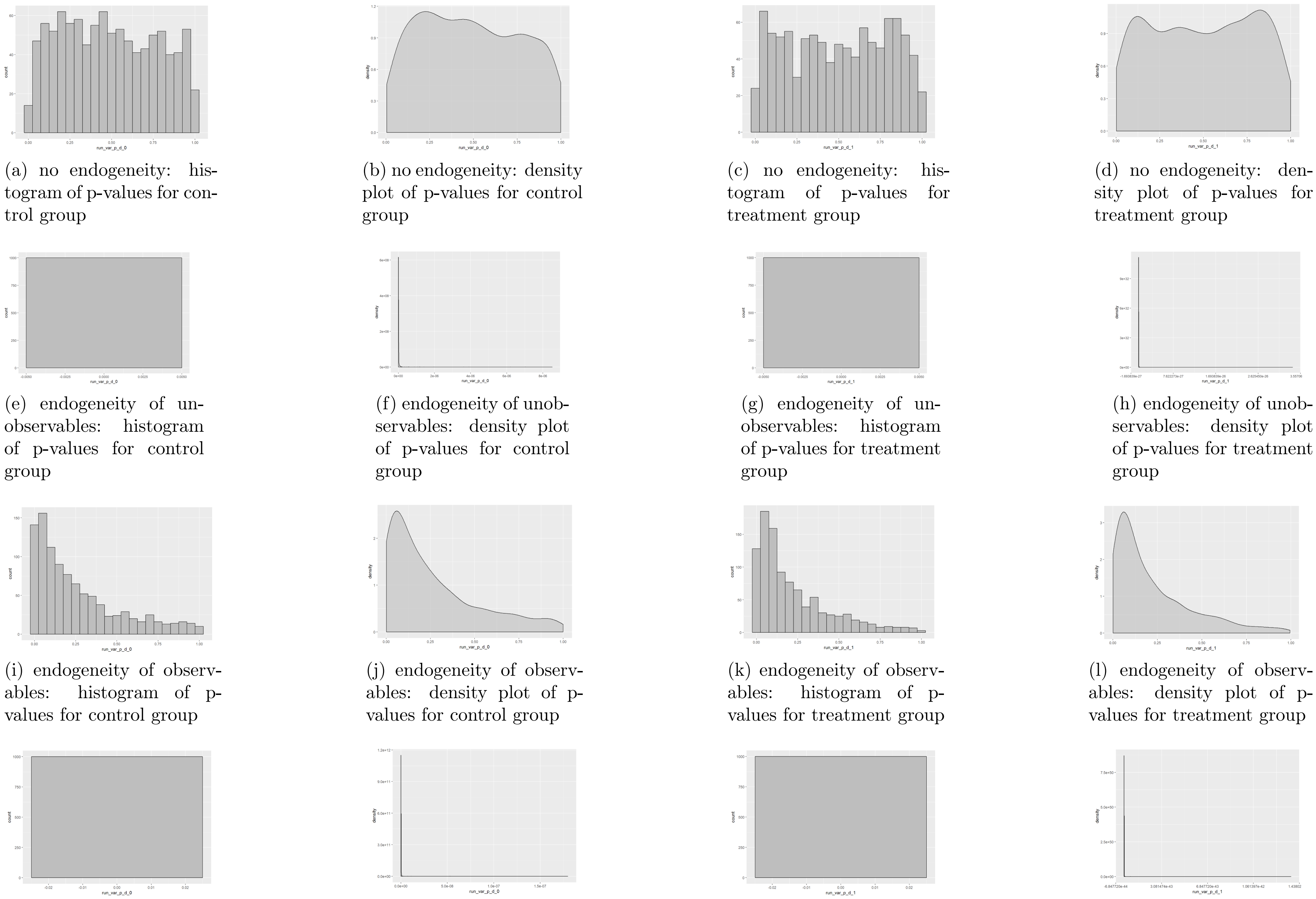
The results are presented in the third row of Figure 1. Overall, the results show that the distribution of p-value is highly skewed. Therefore, the variable selection procedure in Angrist and Rokkanen (2015) will not work well when such type of endogeneity exists.

The fourth simulation assumes there is endogeneity on both observables and unobservables. Specifically, the joint distribution is:

$$\begin{pmatrix}x_{i1}\\ \varepsilon_{0i}\\ \varepsilon_{1i}\\ \varepsilon_{ri}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.1 & -0.2 & 0.3\\ 0.1 & 1 & 0 & -0.3\\ -0.2 & 0 & 1 & 0.5\\ 0.3 & -0.3 & 0.5 & 1\end{pmatrix}\right]$$

The results are presented in the fourth row of Figure 1. Overall, the results show that the across 1000 simulations, the coefficient of the running variable is significant at 5% level.

Overall, the results show that when there is endogeneity, the CIA condition generally fails when we include the correct covariates. In addition, the variable selection procedure developed in Angrist and Rokkanen (2015) cannot select the right covariates. Even when the right covariates are selected, the coefficient of the running variable is likely to be significant.



(a) no endogeneity: histogram of p-values for control group

(b) no endogeneity: density plot of p-values for control group

(c) no endogeneity: histogram of p-values for treatment group

(d) no endogeneity: density plot of p-values for treatment group

(e) endogeneity of unobservables: histogram of p-values for control group

(f) endogeneity of unobservables: density plot of p-values for control group

(g) endogeneity of unobservables: histogram of p-values for treatment group

(h) endogeneity of unobservables: density plot of p-values for treatment group

(i) endogeneity of observables: histogram of p-values for control group

(j) endogeneity of observables: density plot of p-values for control group

(k) endogeneity of observables: histogram of p-values for treatment group

(l) endogeneity of observables: density plot of p-values for treatment group

(m) endogeneity of observables and unobservables: histogram of p-values for

(n) endogeneity of observables and unobservables: density plot of p-values for

(o) endogeneity of observables and unobservables: histogram of p-values for

(p) endogeneity of observables and unobservables: density plot of p-values for

2.5 DML-based Extrapolation Procedures

Recall that Lemma 2 gives a sufficient condition that makes CIA holds. One implication follows from Lemma 2 is that the right covariates determine both potential outcomes and running variable. Based on Lemma 2, our proposed methodology is:

- Step 1. Split the sample into two folds with equal size: train set and test set.
- Step 2. First, use train set data, regress Y_i on x_i given $D_i = d$, where $d \in \{0, 1\}$, by LASSO to select covariates with coefficients that are not zero, say, x_{1di}. Second, use train set data, regress R_i on x_i by LASSO to select covariates with coefficients that are not zero, say, x_{ri}.
- Step 3. Combine the selected covariates from Step 2 to generate the conditional set: z_{di}, for control and treatment groups. Specifically, $z_{di} = x_{1di} \cup x_{ri}$.
- Step 4. Use the test set to estimate $E[Y_i(d)|z_{id}]$ by estimating $E[Y_i|z_{id}, D_i = d]$, where $d \in \{0, 1\}$.
- Step 5. Use test set to compute extrapolated ATE by: $E[Y_i(1)-Y_i(0)] = E[E[Y_i(1)-Y_i(0)|z_i]]$.

Before proceeding to the simulation result, I want to make two remarks on this algorithm. First, to make the proposed new extrapolation method work, we need to make the assumption that the unobservables in selection stage and potential outcomes are independent from each other. In addition, we assume that the observables in the DGP are exogeneous. Second, the main advantage of this new method is in terms of computational performance.

3 Data and Results

3.1 Simulation Results on Double Machine Learning Approach

This section provides simulation studies on the finite sample behavior of the extrapolation method in Angrist and Rokkanen (2015) and the double machine learning method proposed in section 2.5. In addition, we also compare the performance of the following single machine learning method: select covariates by LASSO by regressing Y_i on x_i given $D_i = 0, 1$, then use the selected covariates to compute extrapolated ATE.

Specifically, I repeat the following exercises for 1000 times. In each repetition, set up sample size, N, and use the following model to generate potential outcomes and running variable:

$$Y_i(1) = x_i'\beta + \varepsilon_{1i}$$

$$Y_i(0) = x_i'\theta + \varepsilon_{0i}$$

$$R_i = x_i'\gamma + \varepsilon_{ri}$$

where x_i is a four dimensional column vector with the first entry equal to 1. In other words, we have X_i in our data set, and the first three covariates are useful for extrapolation. In addition, I assume that all observables are distributed normally with mean 0 and variance 1. All unobservables are distributed normally with mean 0 and variance 1. After extrapolation, we will compare performance based on three measures: first, the mean of the extrapolation error; second, the mean of the absolute extrapolation error; third, the mean of the squared extrapolation error.

In the first simulation study, I specify the data generating process (DGP 1) as follows:

$$y_i(0) = 0.8 + 18.3x_{1i} - 13.3x_{2i} - 15.6x_{3i} + \varepsilon_{0i}$$
$$y_i(1) = 1.5 + 23.5x_{1i} + 21.6 * x_{2i} + 22.6 * x_{3i} + \varepsilon_{1i}$$
$$r_i = 1.6 + 6.8 * x_{1i} + 8.3 * x_{2i} + 7.6 * x_{3i} + u_i$$

where the treatment is generated by: $D_i = 1\{R_i > 0\}$. In addition, we have 97 redundant covariates, x_{4i} up until x_{100i}. The simulation results are presented in Table 1.

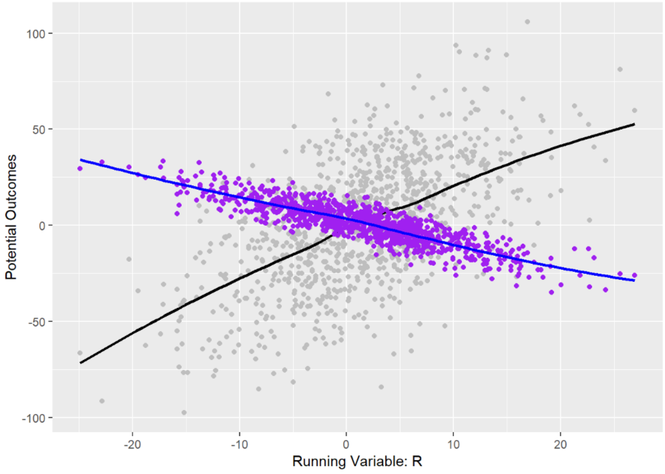


Figure 2: Potential Outcomes over Running Variable (DGP 1)

The simulation results is reported in Table 1.The three columns report results using AR's method, AR with LASSO, and the proposed double machine learning method respectively. The above panel reports results from a simulated dataset with 1000 observations, and the second panel has 2000 observations. The first row of each panel reports mean difference in extrapolation. The second row reports mean of absolute extrapolation error. Two observations could be made here. Three findings could be made here. First, when sample size equals to 1000 and 2000, among all three methods, the DML approach has the best performance in terms of three performance measures. Second, in this simulation result, simple LASSO also has better performance on three performance measures

Table 1: Simulation Result for DGP 1 with Single Cutoff

Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	0.5159	0.04276	-0.08795
Abs(Extrapolation Error)	1.905466	2.06358	1.44159
MSE of Extrapolation Error	10.30503	6.20552	3.141086
N	1000	1000	1000
Extrapolation Error	0.1624	0.164	0.1639
Abs(Extrapolation Error)	0.97363	0.972988	0.972924
MSE of Extrapolation Error	1.487618	1.483372	1.483115
N	2000	2000	2000

Table 2: Simulation Result for DGP 2 with Single Cutoff

Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	0.1665	3.7291	0.01781
Abs(Extrapolation Error)	0.799501	3.7404	0.674
MSE of Extrapolation Error	1.072785	15.16516	0.693618
N	1000	1000	1000
Extrapolation Error	0.07641	3.8978	0.07721
Abs(Extrapolation Error)	0.559227	3.991	0.560756
MSE of Extrapolation Error	0.4676639	17.1152	0.4679712
N	2000	2000	2000

than the method in Angrist and Rokkanen (2015). Third, when sample size increases from 1000 to 2000, the improvement of using DML approach is smaller.

In the second simulation study, I specify a different DGP (DGP 2):

$$y_i(0) = 0.8 + 2.3x_{1i} - 6.3x_{2i} - 15.6x_{3i} + \varepsilon_{0i}$$
$$y_i(1) = 1.5 + 23.5x_{1i} + 1.6 * x_{2i} + 2.6 * x_{3i} + \varepsilon_{1i}$$
$$r_i = 1.6 + 6.8 * x_{1i} + 3.3 * x_{2i} + 3.6 * x_{3i} + u_i$$

where the treatment is generated by: $D_i = 1\{R_i > 0\}$.The simulation result is presented in Table 2. Again, there are three findings from Table 2. First, overall, DML approach performs better in three performance measures than the methods in Angrist and Rokkanen (2015). Second, in this simulation, only use LASSO once perofrms worse than other two methods. Third, when sample size increases from 1000 to 2000, the improvement of DML approach is small.

After getting the point estimate of the extrapolated ATE, a natural task would be constructing the confidence interval of extrapolated ATE. Currently, we do not have an analytical solution to the confidence interval of the extrapolated ATE. In order to study the standard deviation of our extrapolation estimation, I generated 100 datasets, resample the simulated dataset for 100 times, and estimate the ATE for each sampling dataset. For each simulation, I constructe the confidence interval for 100 extrapolation estimates,

Table 3: Coverage Probability of 95% Bootstrapped Confidence Interval

DGP 1			
Methodology	AR + Split	AR + LASSO	AR + DML
Coverage Probability	1	1	1
DGP 2			
Methodology	AR + Split	AR + LASSO	AR + DML
Coverage Probability	1	0.99	1

and compare it with the true ATE. However, the standard deviation is too large that the confidence interval covers the true ATE with nearly 100% for all three methods. The results in Table 3 show that the bootstrapped confidence interval does not have desired coverage property.

3.2 Extension 1: Multi-Cutoff RD Design

The multiple cutoffs RD design is quite prevelant in empirical work in economics and political science (Cattaneo et al., 2016). Therefore, the first extension of the method is to do extrapolation when there are multiple cutoffs in RD design.

Consider the simplest case in multi-cutoff RD design, 2 cutoffs RD design. The support of cutoff is: $C = \{c_1, c_2\}$. In this paper, we consider the non-cumulative multi-cutoff RD design, i.e., the treatment at each cutoff is the same (Cattaneo, Titiunik, and Vazquez-Bare, 2018). And the treatment is assigned by the following mechanism: $D_i = 1\{R_i \geq c_i\}$, where $i \in \{0, 1\}$. We write the potential outcomes as: $Y_{di}(c)$, where $d \in \{0, 1\}$ and $c \in \{c_1, c_2\}$. We also assume that the potential outcome functions are continuous at each cutoff, i.e., $\forall c \in C$, $E[Y_{0i}(c)|R_i = r, C_i = c]$ and $E[Y_{1i}(c)|R_i = r, C_i = c]$ are continuous in r at $r = c$.

One way of thinking multiple cutoffs is to think them as different subpopulations (Cattaneo et al., 2019). Therefore, we can think of different cutoffs depend on both observed and unobserved characterstics of the units. Therefore, there are two extensions of CIA to multi-cutoff RD designs.

Assumption 6. $E[Y_{di}(c)|x_i, R_i, C_i = c] = E[Y_{di}|x_i]$

Remark 1. Assumption 6 means that covariates x_i can wipe out both the effect of cutoffs and running variable on potential outcomes. We can think of the same set of covariates x_i determine both potential outcomes and running variables at different cutoffs. ■

Assumption 7. $E[Y_{di}(c)|x_i, R_i, C_i = c] = E[Y_{di}|x_i, C_i = c]$

Remark 2. Assumption 7 means that covariates x_i can wipe out the effect of running variable on potential outcomes. We can think of different sets of covariates x_i generate potential outcomes and running variables at different cutoffs. ■

Table 4: Simulation Result 2: Two Cutoffs that are Same Types

Cutoff 1			
Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	-0.004317	3.619	0.002138
Abs(Extrapolation Error)	0.506142	3.619	0.6932
MSE of Extrapolation Error	0.3580091	13.835	0.714849
N	2000	2000	2000
Cutoff 2			
Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	0.01658	4.471	-0.1401
Abs(Extrapolation Error)	0.482994	4.471	0.81516
MSE of Extrapolation Error	0.3478324	20.966	0.986248
N	2000	2000	2000

To implement the extrapolation in multi-cutoff RD design, it boils down to the selection of the right covariates at each cutoff. A simple modification of the method in section 2.5 is that we can repeat the variable selection procedure at each cutoff to do the extrapolation of ATE at each cutoff. We conduct two sets of simulation studies to examine the performance of the method.

In the first simulation study, there are two cutoffs. In both cutoffs, the same set of covariates generate potential outcomes and running variable. In other words, we can think of different cutoffs are the same types. The DGP is:

$$y_i(0c) = 5.65 + 19.6x_{1i} - 3.6x_{2i} - 5.5x_{3i} + \varepsilon_{0i}$$

$$y_i(1c) = 11.2 + 2.5x_{1i} + 16.9 * x_{2i} + 3.5 * x_{3i} + \varepsilon_{1i}$$

$$r_i = 1.6 + 5.6 * x_{1i} + 8.3 * x_{2i} + 6.6 * x_{3i} + u_i$$

The treatment is assigned by the following scheme: $D_i = 1\{R_i \geq c\}$, where $c \in \{2, 5\}$.

The simulation result is presented in Table 4. The are two conclusions from the result in Table 4. First, DML approach does not outperform Angrist and Rokkanen's method. Second, single LASSO has the worst extrapolation precision among three extrapolation methods.

Table 5: Simulation Result 2: Two Cutoffs that are Different Types

Cutoff 1			
Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	1.0139	4.293	0.04702
Abs(Extrapolation Error)	1.248002	4.293	0.332591
MSE of Extrapolation Error	7.9508	18.58	0.157128
N	2000	2000	2000
Cutoff 2			
Methodology	AR + Split	AR + LASSO	AR + DML
Extrapolation Error	1.3691	11.978	0.029917
Abs(Extrapolation Error)	1.717539	11.978	0.495368
MSE of Extrapolation Error	10.2483	144.42	0.3658364
N	2000	2000	2000

In the second simulation study, the DGP is:

$$y_{i0}(1) = 0.65 + 9.6x_{1i} - 3.6x_{2i} - 3.5x_{3i} + \varepsilon_{0i}$$

$$y_{i1}(1) = 1.2 + 15.5x_{1i} + 3.9 * x_{2i} + 4.5 * x_{3i} + \varepsilon_{1i}$$

$$y_{i0}(2) = 0.5 + 8.6x_{4i} - 3.6x_{5i} - 5.5x_{6i} + 6.3x_{7i} + \varepsilon_{0i}$$

$$y_{i1}(2) = 1.3 + 19.9x_{4i} + 8.8x_{5i} + 3.6x_{6i} + 8.8x_{7i} + \varepsilon_{1i}$$

$$r_{i1} = 1.6 + 5.6 * x_{1i} + 8.3 * x_{2i} + 6.6 * x_{3i} + u_i$$

$$r_{i2} = 2.3 + 6.5 * x_{4i} + 8.5 * x_{5i} + 5.3 * x_{6i} + 7.2x_{7i} + u_i$$

There are two cutoffs: 0 and 5.

The simulation result is presented in Table 5. There are two conclusions from Table 5. First, DML approach performs better than the other two methods in three performance measures at both cutoffs. Second, the single machine learning method performs the worst among three methods.

3.3 Extension 2: Allow Endogeneity in the DGP

In the second part of extension, I allow endogeneity in the DGP and study the behavior of DML method.

The DGP is:

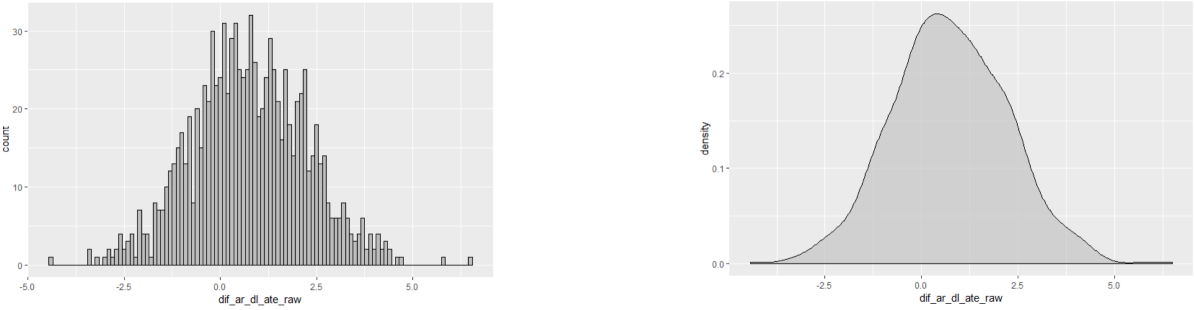
$$y_i(0) = 0.8 + 2.3x_{1i} - 6.3x_{2i} - 15.6x_{3i} + \varepsilon_{0i}$$

$$y_i(1) = 1.5 + 3.5x_{1i} + 21.6x_{2i} + 2.6x_{3i} + \varepsilon_{1i}$$

$$R_i = 1.6 + 1.8x_{1i} + 3.3x_{2i} + 7.6x_{3i} + \varepsilon_{ri}$$

Table 6: Probability of Selecting Correct Covariates in Double LASSO

D = 0	D = 1	Running Variable	D = 0 & Running Variable	D = 1 & Running Variable
0	0	0.994	0.994	0.994



(a) Distribution of extrapolation error

(b) Histogram of extrapolation error

Figure 3: Distribution of Extrapolation Error When There is Endogeneity

I also assume that:

$$\begin{pmatrix} x_{i1} \\ \varepsilon_{0i} \\ \varepsilon_{1i} \\ \varepsilon_{ri} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & -0.8 \\ 0 & 1 & 0 & -0.3 \\ 0 & 0 & 1 & 0.5 \\ -0.8 & -0.3 & 0.5 & 1 \end{pmatrix} \right]$$

I will focus on two aspects of the new method in the presence of endogeneity: 1) the probability of selecting the correct covariates; 2) the distribution of the extrapolation error. The results on the probability of selecting the correct covariates are presented in Table 6. There are three conclusions. First, using merely treatment or control group to select covariates does a poor job on variable selection. The probability that we select the correct set of covariates is 0 when only using treated and control groups. Second, the probability of selecting correct covariates is close to 1 when using selection stage regressions. Third, the selection stage covariates selection can correct the mistake of variable selecion when only using treated or control groups.

In addition, I also plot the histogram and the density plot of the extrapolation error in Figure 3. Despite high accuracy in selecting covariates that generate both potential outcomes and running variable, our method tends to overestimate the true ATE when doing extrapolation.

4 Conclusion

Sharp regression discontinuity (RD) design identifies the causal effect of the treatment at the treatment threshold. The treatment effect on inframarginal populations is of interest, but the identification of such treatment effect needs stronger assumptions. Following the methodology in Angrist and Rokkanen (2015), I leverage the CIA assumption to develop a new double machine learning method to select covariates to extrapolate the average treatment effect in sharp RD design. The main advantage of the new method is that it is computationally faster. The simulation results confirm the better performance of the new extrapolation method. I also provide two extensions of our methods. First, extend the new method to the case when there are more than 2 cutoffs. Second, study the performance of the new method when we allow for endogeneity in the DGP.

There are three directions that I could work on in the future. First, compare performances between double machine learning with other machine learning approach and matching methods. Second, extend the methods to the case where there are more than three cutoffs and try to model how the unobservables affect potential outcomes directly. Third, apply the extrapolation method to fuzzy RD design cases.

References

Angrist, Joshua D and Miikka Rokkanen, "Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff," *Journal of the American Statistical Association*, 2015, *110* (512), 1331–1344.

Cattaneo, Matias D, Luke Keele, Rocio Titiunik, and Gonzalo Vazquez-Bare, "Extrapolating treatment effects in multi-cutoff regression discontinuity designs," *arXiv preprint arXiv:1808.04416*, 2018.

_ , Rocio Titiunik, and Gonzalo Vazquez-Bare, "Analysis of Regression Discontinuity Designs with Multiple Cutoffs or Multiple Scores," 2019.

_ , Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele, "Interpreting regression discontinuity designs with multiple cutoffs," *The Journal of Politics*, 2016, *78* (4), 1229–1248.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

Dong, Yingying and Arthur Lewbel, "Identifying the effect of changing the policy threshold in regression discontinuity models," *Review of Economics and Statistics*, 2015, *97* (5), 1081–1092.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 2001, *69* (1), 201–209.

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

Wong, Vivian C, Peter M Steiner, and Thomas D Cook, "Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods," *Journal of Educational and Behavioral Statistics*, 2013, *38* (2), 107–141.