A Basic Tutorial for Regression Discontinuity

Mingyang Yan Jingyan Jiang

November 10, 2019

1 Why do we need RD?

In the studies of social science, Randomized Controlled Trial (RCT) are extremely limited due to high economic costs and ethical reasons. Then we consider the quasi-randomized experiment method, which uses statistical control to simulate experimental control to test the hypothesis on the interested casual effects. One of the representatives is Regression Discontinuity (RD).

Recall that the underlying causal relationship we want to study is essentially a counterfactual problem. More specifically, let us see an example of new drug treatment. A group of patients were assigned to the experiment group for new drug treatment. The effect of this group of patients after treatment is the "facts" we can observe. "Counter-fact" means that if the same group of patients are assigned to the control group, then what are their symptoms? Statistically, the causal effect of the new drug refers to the difference between the same subjects in the experiment group and in the control group. Define a dummy variable D, D=1 for experiment group and the corresponding outcome is Y1; D=0 with outcome Y0 for control group. p is the probability of taking the new drug. Then the true treatment effect is

$$T = p*[E(Y_1|D=1) - E(Y_0|D=1)] + (1-p)*[E(Y_1|D=0) - E(Y_0|D=0)]$$

However, we cannot observe the performance of the experiment group when there is no experiment, either can we know the experimental effect of control group.

2 What is RD?

From previous example, we can see that $E(Y_0|D=1)$ and $E(Y_1|D=0)$ are not possible to be observed. Then statisticians proposed the unconfoundedness assumption, which assumes that $E(Y_1|D=1) = E(Y_1|D=0)$ and

$E(Y_0|D=0) = E(Y_0|D=1)$. That requires D and Y are independent with each other.

The main idea of RD is that the samples are approximated to be randomly distributed around the critical value. Samples smaller than the critical value are used as control groups, and samples larger than the critical value are used as experimental groups. The causal relationship between the intervention variables and the outcome variables is studied by comparing their differences. In formal definition, RD can be divided into two types.

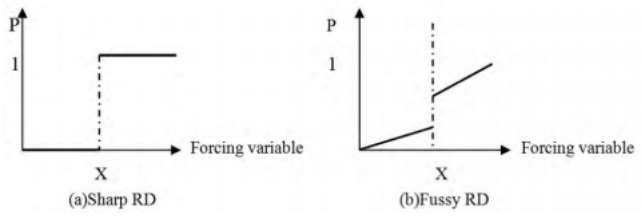

Figure 1: Sharp RD and Fuzzy RD

2.1 Sharp RD

The key to the RD design is that we have a deep understanding of the mechanism which underlies the assignment of treatment D_i; In this case, assignment to treatment depends on a single variable X_i. In the sharp RD design this variable fully determines the treatment according to the cutoff rule: $D_i = 1, if \quad X_i \geq X_0; D_i = 0\, if \quad X_i < X_0$. X_i is called the running variable.

$E[Y_{0i}|X_i]$ is a function of X_i, so $Y_{0i} = f(X_i) + e_i, \quad Y_{1i} = Y_{0i} + \beta$

$$Y_i = f(X_i) + \beta D_i + e_i = f(X_i) + \beta I(X_i \geq X_0) + e_i$$

Notice that the function f must be continuous atX_0. In practice, we always assume some flexible form for f, such as a polynomial.

2.2 Fuzzy RD

Instead of a deterministic assignment rule there may only be a change in the probability of treatment at the cutoff.

$$Pr(D_i = 1) = p(X_i)$$

$$\lim_{X_i \to X_0} p(X_i) \neq \lim_{X_0 \to X_i} p(X_i)$$

The probability of treatment $p(X_i)$ is also a continuous function, except at X_0. There are more than one regression now.

$$\text{structural model} \quad Y_i = f_1(X_i) + \beta D_i + e_i$$

$$\text{reduced form} \quad Y_i = f_2(X_i) + \pi_2 I(X_i \geq X_0) + \xi_{2i}$$

$$\text{first stage} \quad D_i = g(X_i) + \pi_1 I(X_i \geq X_0) + \xi_{1i}$$

The cutoff induces a change in the probability of treatment. If treatment matters, this induces a change in the outcome. Since the treatment does not affect all units, the jump at the cutoff in the outcome needs to be rescaled by the jump at the cutoffs in the probability of treatment. According to the standard IV model, $\beta = \frac{\pi_2}{\pi_1}$

3 How can we apply RD?

RD is a very useful tool for policy evaluation and assessment of education program. In this part, we use Stata to illustrate how we can run RD in a software.

3.1 Generate a simulated dataset

```
set obs 4000
set seed 123
gen x = runiform()
gen T=0
gen e = rnormal()/5 // noise
replace T=3 if x>0.5 // set the cut point = 0.5
gen y = T + 3*x + e
scatter y x, msize(*0.5) ytitle("y") xtitle("x") // plot the data points in terms of y and x
```

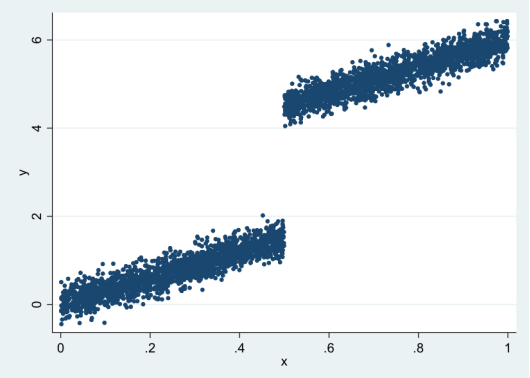

Figure 2: Scatter Plot of Simulated Data

3.2 Try simple OLS regression

reg y x
twoway (scatter y x, msymbol(+) msize(*0.4) mcolor(black*0.3)) ///
(lfit y x, lcolor(blue) msize(*0.4)), ///
ytitle("y")

```
. reg y x
```

Source	SS	df	MS	Number of obs	=	4,000
				F(1, 3998)	=	31044.60
Model	18738.0914	1	18738.0914	Prob > F	=	0.0000
Residual	2413.13728	3,998	.603586114	R-squared	=	0.8859
				Adj R-squared	=	0.8859
Total	21151.2287	3,999	5.28912946	Root MSE	=	.77691

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	7.500119	.0425672	176.19	0.000	7.416663	7.583574
_cons	-.7499307	.0247192	-30.34	0.000	-.7983941	-.7014672

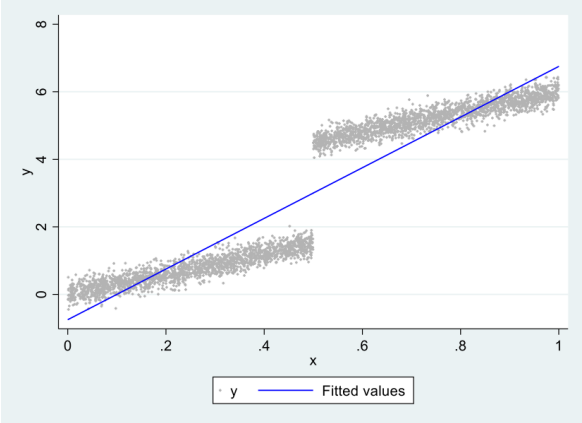

Figure 3: Scatter Plot of Simulated Data and Fitted OLS Regression

It is obvious that the fitting model cannot depict the observed data properly, as the coefficient of x is 7.5 >> 3.

3.3 Use regression discontinuity

reg y x if x>0.5
reg y x if x<0.5

```
. reg y x if x>0.5
```

Source	SS	df	MS	Number of obs	=	2,026
				F(1, 2024)	=	9507.89
Model	371.412855	1	371.412855	Prob > F	=	0.0000
Residual	79.0648199	2,024	.039063646	R-squared	=	0.8245
				Adj R-squared	=	0.8244
Total	450.477675	2,025	.222458111	Root MSE	=	.19765

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	2.970212	.0304611	97.51	0.000	2.910474	3.02995
_cons	3.020945	.023285	129.74	0.000	2.97528	3.06661

```
. reg y x if x<0.5
```

Source	SS	df	MS	Number of obs	=	1,974
				F(1, 1972)	=	9327.75
Model	369.943505	1	369.943505	Prob > F	=	0.0000
Residual	78.2105268	1,972	.039660511	R-squared	=	0.8255
				Adj R-squared	=	0.8254
Total	448.154031	1,973	.227143452	Root MSE	=	.19915

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	3.003533	.0310988	96.58	0.000	2.942543	3.064523
_cons	-.0028359	.008992	-0.32	0.753	-.0204707	.014799

In this case, the coefficient of x is 2.97 and 3.00 for the region x<0.5 and x>0.5 respectively, which are much closer to the coefficient we set for the simulated data set. The difference between continues regression model with discontinues regression model can also be observed visually.

twoway (scatter y x, msymbol(+) msize(*0.4) mcolor(black*0.3)) ///
(lfit y x if T==0, lcolor(red) msize(*0.4)) ///
(lfit y x if T==3, lcolor(red) msize(*0.4)) ///
(lfit y x, lcolor(blue) msize(*0.4)), ///
ytitle("y")

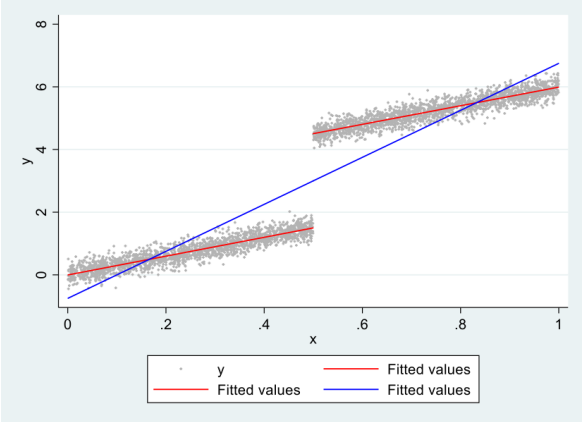

Figure 4: Fitted Regression Discontinuity and OLS Regression

4 Acknowlegement

[1] Jorn-Steffen Pischke, Regression Discontinuity Design, Lecture notes, The London School of Economics and Political Science
[2] Lemieux, L. T. .(2010). Regression discontinuity designs in economics. Journal of Economic Literature, 48(2), 281-355.
[3] Github, Jiaming Mao, 2019, https://jiamingmao.github.io/data-analysis/
[4]Wikipedia, https://en.wikipedia.org/wiki/Regression_discontinuity_design
[5]CSDN, https://blog.csdn.net/Hellolijunshy/article/details/88383040