Regression Discontinuity Design

Contents

Introduction

We know that, to do treatment effect, we need make convincing assumption that inviduals chooses gets into the treatment based on some observable variables. A speical case is that, invidiuals get treatment when x is larger than a threhold value c. A big problem with this situation is that common support assumption is violated: there is no overlapping x in treatment group and control group.

Sharp RDD

A strategy is focusing on the local area around c. We can think of that for people with x very close to c, they are actually very close to each other; whether they get treatment or not is nearly random.

To be more specific, consider $(c-\epsilon, c+\epsilon)$, where ϵ is very small. For those individuals at $(c-\epsilon, 0)$ and at $(0, c+\epsilon)$, we can regard them as having identical x, and whether they get treatment or not is purely random.

Therefore, we can impose the following CIA assumption: given $x = c$, z, whether getting treatment or not is random . This allows us to estimate the ATE at $x = c$. Of course, in other values of x, the CIA is violated, so we cannot estimate ATE.

$$LATE(z) = E(y_1 - y_0 | D = 1, x = c, z) = E(y_1 | D = 1, x = c, z) - E(y_0 | D = 1, x = c, z)$$

with the above CIA assumption, s.t. given $x = c$, z, we easily have

$$LATE(z) = E(y_1 | D = 1, x = c, z) - E(y_0 | D = 0, x = c, z)$$

Next we need to estimate $E(y_1 | D = 1, x = c, z)$ and \$ E(y_0 |D =0, x= c, z)\$

Estimation Method

Non-parametric

How to estimate $E(y_1 | D = 1, x = c, z)$? Only those inviduals with $x < c$ get the treatment, so we can estimate this term using

$$\lim_{x \to c^+} E(y | x, z)$$

Similarly, we can estimate $E(y_0|D = 0, x = c, z)$ using $\lim_{x\to c^-} E(y|x, z)$.

$$LATE(x = c, z) = \lim_{x\to c^+} E(y|x, z) - \lim_{x\to c^-} E(y|x, z)$$

In which z is other exogeneous covariates. This expression gives the local effect of treatment at $x = c$ and z. Of course we can average the $LATE$ over z:

$$LATE(x = c) \equiv \int \lim_{x\to c^+} E(y|x, z)pr(z|x)dz - \int \lim_{x\to c^-} E(y|x, z)pr(z|x)dz$$

This expression gives more details we need pay attention to. From this expression, $LATE(x = c)$ comes from two part: one is the effect of treatment, the other is difference of z's distribution between $pr(z|x, x \to c^+)$ and $pr(z|x, x \to c^-)$. If these two distributions are different, then we cannot estimate the real effect of the treatment. Therefore, before estimation, confirm that the these two distributions are identical !

Regression:

We can also consider the following regression when $x \in (c - h, c + h)$:

$$y = \alpha + \beta(x - c) + \delta * 1\{x > c\} + \gamma * 1\{x > c\} * (x - c) + \theta z + \epsilon$$

- By estimating δ we can get the effect the net effect of treatment on y. The $\gamma * 1\{x > c\} * (x - c)$ is to allow for the differnet in slopes on $x < c$ and $x > c$. This term is not necessary, though.
- Again, be sure that z's distributions are identical on the two sides of $x = c$ (i.e, check that z is independent of $1\{x > c\}$). If not, the estimated δ is in general not the net effect of the treatment.
- One obvious problem with this regression is that it only use the information of those data at $x \in (c - h, c + h)$. This is not good. To make use of more data points, we can instead use local linear regression, by estimating:

$$\min_{\alpha,\beta,\delta,\gamma} \sum_{i=1}^{n} K\left(\frac{x_i - c}{h}\right)(y_i - \alpha - \beta(x_i - c) - \delta * 1\{x_i > c\} - \gamma * 1\{x_i > c\} * (x_i - c))^2$$

Remarks

- Check This Assumption :Individuals do not know the value c ex-ante, i.e., they cannot adjust their own x ex-ante according to c. If, otherwise, individuals knows the c (and also, the benefit of getting treatment) beforehand, they could change their x to affect whether they get treatment or not. In this sense, treatment is no longer random.
  - Consider a situation, where we study the effect of whether going to university on subsequent income. Suppose people know that they can get into university if their score is above 500. if I am a hard-working person and I know I can get into the university if my exam score is abot 500, I will work hard to make my score, x, be above 500. Since I am a hard-working person, it is also possible that I earn high income. Self-selection problem happens again! Therefore, RDD is good for the situation where x is predetermined (determined before people know the c)
- So in general, we need to check the distributions of x on both sides near the c and confirm if the distribution is continuous at $x = c$. If not, then randomness is doubtful.
- As is stressed above, make sure the other covariates z has the same distribution at the two sides of $x = c$
- Although the RDD seems limited in the sense that we can infer the effect of the treatment around the threshold. But it is still useful, since in reality we do encounter similar cases:
  - Customer can be prime member when their purchase x is above a threshold value.

  - People can enjoy a certain pension benefit when their income x is below a certain level.
  - ...

Of course, before estimation, confirm that individuals' x is predetermined before the treatment is put into effect.

Fuzzy RDD

Now we consider a case where the threshold c is 'fuzzy', in the sense that there is just of jump of the probability of getting into treatment from $x = c^-$ to $x = c^+$,i.e.,

$$a = \lim_{x \to c^-} pr(D = 1|x) \neq \lim_{x \to c^+} pr(D = 1|x) = b$$

When $\lim_{x \to c^-} pr(D = 1|x) = 0$ and $\neq \lim_{x \to c^+} pr(D = 1|x) = 1$, it turns to Sharp RDD.

The key issue here is that, even when x is above c, individuals may not get into treatment group: it may reject to get treated due to some (observable/unobservable)factors, which may also affect the outcome. Compared to the sharp RDD, fuzzy RDD is more realistic.

Notice that when we deal with some basic treatment effect problems, we always do the following procdure:

- (method 1) If the decision of whether get treatment is totally random after conditionin on observable variables x, i.e., then we can easisy do our estimation by conditioning x.
- (method 2)If the decision of whether get treatment is may depend on some ubservable factors, then we have to
  - Use some IV: find a dummy variable z that linearly affects the treatment decision, so that we have

    $$ATE = \frac{E(y|z = 1) - E(y|z = 0)}{E(D|z = 1) - E(D|z = 0)}$$

    see the chapter 'IV and LATE' for detailed information.
  - Use heckman two-step or MLE.

The same procedure can be applied to the RDD. we can use method 1, keeping the CIA; we can also use method 2, finding an IV.

Method 1: Keep CIA

$$\begin{aligned} E(y|x) &= E(y_1|D = 1, x)pr(D = 1|x) + E(y_0|D = 0, x)pr(D = 0|x) \\ &= (E(y_1|D = 1, x) - E(y_0|D = 0, x))\, pr(D = 1|x) + E(y_0|D = 0, x) \end{aligned}$$

Let's evaluate this equation at $x \to c^+$ and $x \to c^-$

$$\lim_{x \to c^+} E(y|x) = (E(y_1|D = 1, x = c) - E(y_0|D = 0, x = c)) \lim_{x \to c^+} pr(D = 1|x) + E(y_0|D = 0, x = c)$$

$$\lim_{x \to c^-} E(y|x) = (E(y_1|D = 1, x = c) - E(y_0|D = 0, x = c)) \lim_{x \to c^-} pr(D = 1|x) + E(y_0|D = 0, x = c)$$

Recall our CIA assumption that

$$E(y_0|D = 0, x = c) = E(y_0|D = 1, x = c)$$

we immedietely get our LATE estimation.

$$E(y_1 - y_0|D = 1, x = c) = \frac{\lim_{x \to c^+} E(y|x) - \lim_{x \to c^-} E(y|x)}{\lim_{x \to c^+} pr(D = 1|x) - \lim_{x \to c^-} pr(D = 1|x)}$$

Method 2: Use IV

What is a proper IV for D around $x = c$? One candidate is $Z = 1\{x > c\}$. First, Z is only determined by x, so it is not correlated with unobservable factors influencing y. Second, Z is obviously related to D. We want to estimate

$$LATE = \frac{E(y|Z=1, x=c) - E(y|Z=0, x=c)}{E(D|Z=1, x=c) - E(D|Z=0, x=c)}$$

in which

$$E(y|Z=1, x=c) = E(y|D=1, Z=1, x=c)pr(D=1|Z=1, x=c) + E(y|D=0, Z=1, x=c)pr(D=0|Z=1, x=c) =$$

$$E(y|D=1, Z=1, x=c) * b + E(y|D=0, Z=1, x=c) * (1-b)$$

in which $E(y|D=1, Z=1, x=c)$ can be replaced by $\lim_{x \to c^+} E(y|D=1, x)$, $E(y|D=0, Z=1, x=c)$ can be replaced by $\lim_{x \to c^+} E(y|D=0, x)$. Similarly, we have

$$E(y|Z=0, x=c) = E(y|D=1, Z=0, x=c)pr(D=1|Z=0, x=c) + E(y|D=0, Z=0, x=c)pr(D=0|Z=0, x=c) =$$

$$E(y|D=1, Z=0, x=c) * a + E(y|D=0, Z=0, x=c) * (1-a)$$

in which $E(y|D=1, Z=0, x=c)$ can be replaced by $\lim_{x \to c^-} E(y|D=1, x)$, $E(y|D=0, Z=0, x=c)$ can be replaced by $\lim_{x \to c^-} E(y|D=0, x)$. Finally we have

$$E(D|Z=1, x=c) = pr(D=1|Z=1, x=c) = b, \quad E(D|Z=0, x=c) = pr(D=1|Z=0, x=c) = a$$