ECNM10112: Applied Labour Economics

Regression Discontinuity Design (RDD) notes

As the name suggests, Regression Discontinuity Design is an empirical strategy that exploits discontinuities in the data to credibly establish causality and estimate treatment effects.

Rules, policies, or traditions often introduce arbitrary cutoffs that affect the probability of receiving some treatment of interest. For example:

- In many countries, you can start receiving your pension only after you turn 65 years old.
- Tax systems often have cutoffs that introduce discontinuities on your tax rate. For example, currently in the UK you pay 0% over all income below £12,750 but you pay 20% over income between £12,571 and £50,270.
- Income thresholds often determine your ability to access government aid. For example, suppose only students whose parents make £15,000 or less can get access to school meals.
- Schools in Israel often split classes with more than 40 students (Angrist and Lavy, 1999). That is, if there are 40 students in fourth grade they stay as a single class. However, if a new student arrives then the school splits fourth grade into two classes.

RDD exploits the arbitrariness of these cutoffs to estimate treatment effects.

1 A simple example

To fix ideas, suppose that you are interested in determining the effect of free school meals on student achievement. Your first idea is to run the following regression:

$$score_i = \alpha + \tau school\_meal_i + u_i \tag{1}$$

where $score_i$ is some measure of student achievement and $school\_meal_i$ meal is a dummy equal to one if the pupil receives a free school meal. Here, you are interested in estimating τ, the treatment effect of the school meal.

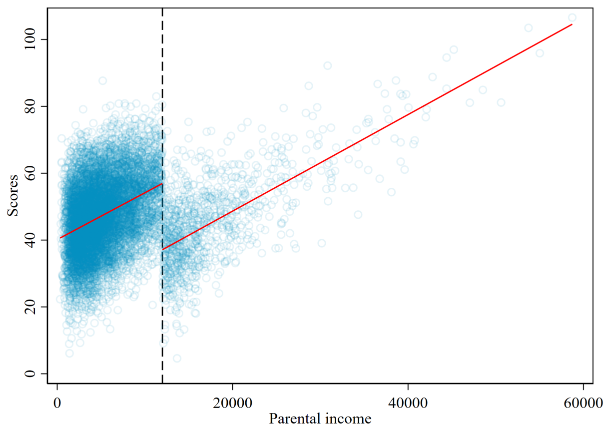
Regression (1) is unlikely to give us an unbiased estimate of the effect of school meals because chances are that eligibility for school meals is correlated with a host of unobserved factors that determine student achievement, that is $E(school\_meal\_i, u_i) \neq 0$. For example, kids receiving subsidised school meals might be attending worse schools, live in worse neighbourhoods, etc.

However, suppose that after extensive research of the government rules, you discover the following:[1]

- All students whose household income is £12,000 or less get free school meals.
- All students whose household income is greater than £12,000 must pay meals in full.

The basic intuition of RDD is that this rule introduces differences in treatment for people that are similar in all other respects. Chances are that students whose parental income is £12,000 are very similar to those whose parental income is £12,001. However, the former get the free meals, while the latter do not. Therefore, if the £12,000 students have higher scores than the £12,001 students, we can credibly argue that these differences are caused by the free school meals.

Figure 1: RDD in a graph



Notes: Vertical line denotes £12,000. Figure generated on 23 Sep 2024 at 18:26:33.

Figure 1 illustrates this argument graphically. The figure plot the relationship between students' scores and parental income in some simulated data. Unsurprisingly, there is a positive relationship between parental income and student achievement. However, note that there is a discontinuity at £12,000. Students with incomes just below £12,000 have higher scores than those with income just above £12,000. Because, all students to the left of this point get a free meal, while those to the right don't, the discontinuity in achievement is –presumably– caused by the free meal.

[1] The eligibility rule in this example is very extreme. In practice, rules are often less stringent. For example, rather than going from 100% subsidy to zero, there can be more bands of income and subsidy that generate a more gradual decline in the levels of subsidy.

2 Estimating treatment effects using RDD

We can use regressions to estimate the treatment effect illustrated in Figure 1. To do so, define the free-meal eligibility rule dummy:

$$D_i = \begin{cases} 1, & parental\_income_i \leq 12,000 \\ 0, & parental\_income_i > 12,000 \end{cases}$$

In RDD parental income is called the *the running variable*, i.e. the variable that generates the discontinuity in treatment status. We can estimate the treatment effect of the free meal with the regression:

$$y_i = \alpha_0 + \tau D_i + \beta_1 parental\_income_i + u_i \tag{2}$$

here τ provides an estimate of the size of the discontinuity at £12,000 in Figure 1. Intuitively, τ is estimated by comparing the scores of students just below and just above the 12,000 cutoff. Therefore, τ will be an unbiased estimate of the (local) treatment effect whenever students that are just below and just above the cutoff are similar in all other respects. **This is the key identifying assumption in RDD: outcomes for students on each side of the cutoff would have been the same in the absence of the free meal.**

2.1 Allowing for more flexibility between the dependent and the running variables

Regression (2) is very simple, but imposes a lot of restrictions on the relationship between scores are parental income: it is linear and it is the same on both sides of the discontinuity. However, the relationship between these variables could be non-linear or change at the discontinuity. Figure 2 Panel (a) gives an example where there is a non-linear relationship between the variables. Panel (b) gives an example of a linear relationship but with different slopes (different functional form) on each side of the discontinuity.

These situations are very easy to accommodate. Flexible functional forms can be easily incorporated by using polynomials of the running variable. For instance, below is a RDD specification using a quadratic polynomial:

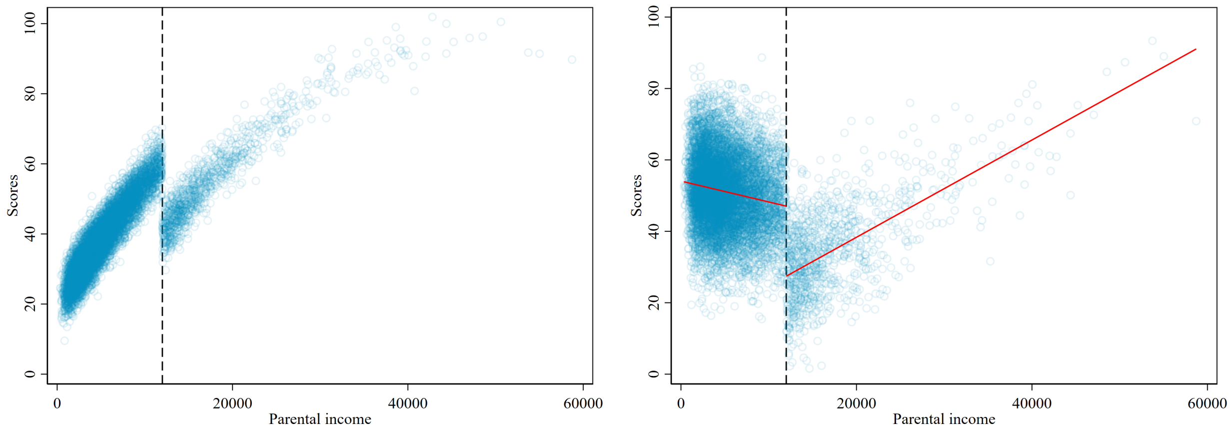
$$y_i = \alpha_0 + \tau D_i + \beta_1 parental\_income\_i + \beta_2 parental\_income_i^2 + u_i \tag{3}$$

here, τ still gives us the treatment effect estimate.

The specification above still imposes the same functional form on each side of the discontinuity.

Figure 2: RDD and functional form

(a) Non-linear relationship with the running variable (b) Different function on each side of the discontinuity



Notes: Vertical line denotes £12,000.

A simple modification relaxes this restriction:

$$\begin{aligned} y_i &= \alpha_0 + \tau D_i + \beta_1 parental\_income\_i + \beta_2 parental\_income_i^2 + \\ &\quad \beta_3 D_i \times (parental\_income\_i - 12,000) + \beta_4 D_i \times (parental\_income\_i - 12,000)^2 + u_i \quad (4) \end{aligned}$$

the two additional terms allow the relationship between scores and income to switch at 12,000.

References

Angrist, J. D. and Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2):533–575.