Université Paris-Saclay
Master of Economics – 1st year
Advanced Econometrics – Tutorial #8 *

April 6th 2021

Context of the session and objectives

This session illustrates the analysis of regression discontinuity based on the paper by Jens Ludwig and Douglas L. Miller "*Does Head Start Improve Childrens Life Chances? Evidence from a Regression Discontinuity Design*, Quarterly Journal of Economics (2007).

The U.S. "Head Start" program has been providing preschool, health and other social services to poor children age three to five and their families since 1965. At the time of the research article, 900,000 children benefit from the program each year. Evaluating whether the program improves childrens life chances is challenging because participation is likely correlated with outcomes.

Ludwig and Miller (2007) exploit the fact that in 1965, only the poorest 300 counties in the U.S. received federal support in applying for the program. This leads to participation and funding rates to be 50–100 percent higher in counties with poverty rates above the cutoff compared with those just below. This discontinuity in the implementation generated a "treatment" group of counties above the poverty rate cutoff and a "control" group made of counties below. This funding difference appears to have persisted through the late 1970s. This discontinuity is then used by the authors to identify impacts by comparing outcomes for people in treatment and control counties "near" the cutoff.

In this session we replicate the parametric (OLS) estimates they get on mortality reduction, one of the main results of their study.

1 Presentation of the estimation strategy and the data

1.1 The regression discontinuity design estimation strategy

1. We first come back on some the regression discontinuity design estimation strategy and the required assumptions to appropriately identify an effect of the policy.
   (a) What is the principle of the regression discontinuity design estimation strategy?

> **Solution:** Regression discontinuity designs exploit precise knowledge of the rules determining treatment. RD identification is based on the idea that in a highly rule-based world, some rules are arbitrary and therefore provide good

*Contact: thibault.richard(at)ens-paris-saclay.fr

experiments. Sharp regression discontinuity is used when treatment status is a deterministic and discontinuous function of a covariate, x^f, which is also called the forcing variable. Suppose, for example, that

$$D_i = \begin{cases} 0 & \text{if } x_i^f < c \\ 1 & \text{if } x_i^f \geq c \end{cases} \tag{1}$$

where c is a known threshold or cutoff. This assignment mechanism is a deterministic function of x_i^f because once we know x_i^f we know D_i. Its a discontinuous function because no matter how close x_i^f gets to c , treatment is unchanged until $x_i^f = c$.

Then, for estimation of the average treatment effect, since it is defined as x^f approaches c from above and below and holds in the limit, one restricts attention to observations around the threshold. We estimate the effect using linear regressions **locally**. We here need to determine **bandwidth** above and below $c \Rightarrow$ running OLS on the subset of observations with poverty rate near the thresholds.

(b) What assumption on the potential outcomes needs to be made to identify an effect at the cutoff?

Solution: Continuity of the Conditional Regression Function: (Assumption 1 in your class) continuity of the outcome in a neighborhood of the threshold. $\mathbb{E}[y \mid x, D = 0]$, and $\mathbb{E}[y \mid x, D = 1]$ are continuous in x^f, where x^f is the forcing variable. In our setting, approaching the threshold on x^f, denoted c, from below covers control observation and approaching c from above covers treated observations.

This smooth / continuous outcomes around the cutoff allows to derive the average treatment effect as follows:

$$\delta_{ATE} = \lim_{x^f \to c^+} \mathbb{E}[y \mid x] - \lim_{x^f \to c^-} \mathbb{E}[y \mid x]$$

By looking at observations on both sides of the threshold, but sufficiently close to it, we eliminate differences in the treated and control populations in terms of unobservables. Relatedly, this assumption implies that there is no potential for distorting the threshold. Here, the cutoff is based on a predetermined variable $\Rightarrow$ no strategic behavior to distord the cutoff which would be concerning in the regression discontinuity design.

In practice, it will be challenging to determine the proper neighborhood of the cutoff to identify an effect $\Rightarrow$ the choice of the bandwidth. Indeed, too narrow will lead to imprecise estimates (too little data). Too wide leads to bias (control observations too far from treated observations for this assumption to ensure

unobservables will not matter). Here, we chose "reasonable" bandwidth, and then (if time) checks the robustness of this choice.

1.2 Analyzing the discontinuity setting with the data

We have county-level observations regarding several variables, and in particular federal spending data at the county-level. These measures of expenditures at the county level are only available for years 1968 and 1972. The Head Start program includes public spending in several dimensions. Around 40 percent of the program's budget is for early childhood schooling. Other services included nutritional program, social services, mental health services and health services. This bundle of services in the Head Start program may affect health (our focus in this tutorial) and schooling (the other focus of the research article) through a variety of channels. Concerning health, most of these services increased the chances children saw a doctor, and improve ealy detection and treatment of some conditions. We agregate all spendings at the county level and divide them by the number of kids in the county at the time of the program and obtain the variables `hsspend_per_kid_68` and `hsspend_per_kid_72`.

Data on child mortality are also at the county level. `age5_9_sum2` measures the mortality of children in the age group five to nine on the 1973 and 1983 period, as all children five to nine at this time period would have been of Head Start age after the program was launched. `age5_9_injury_rate` measures the children mortality caused by injuries (around 55 percent of deats to all children five to nine in 1973–1983). `age25plus_sum2` measures the mortality. For `rate_5964`, it gives the children mortality in the age group five to nine on the 1959 and 1964 period, **before the implementation of the program**.

2. We are here interested in assessing whether the regression discontinuity design can be an appropriate strategy given our data. To do so, we first implement some preliminary graphical and statistical analysis.

   (a) First, generate an identifying variable, denoted `id`, which assigns for every county its rank in the distribution of the poverty rate in 1960 from the poorest to the richest. Second, generate a dummy variable, denominated `g`, which identifies counties benefiting from the provision of grant-writing assistance. Recall that those are the 300 counties with the highest level of poverty rate in 1960. Third, generate a variable, denominated `povrate`, which is the `povrate60` variable rescaled so that the value of `povrate` at the discontinuity corresponds to zero.

   (b) Produce an histogram for the poverty rate variable before the implementation of "Head Start" program (in 1960). What can you conclude regarding this distribution and our estimation strategy?

Solution: See STATA do-file for the density of the forcing variable (poverty rate in 1960).

We need to have enough observations around the threshold in poverty rate. If there is a jump in the density at the threshold, the forcing variable may have been manipulated, which could invalidate the analysis (hypothesis 1 could be violated). This is not the case here.

(c) We chose to compare counties closed to the cutoff. We arbitrarily decide to only use counties with either a bandwidth in poverty rate of 8% below and above the cutoff, or a bandwith in poverty rate of 16% below and above the cutoff. Generate the two variables, denominated respectively `bandwidth1` and `bandwidth2` which identify counties within these bandwidths.

Solution: See STATA do-file for replication

(d) We now investigate the similarities or differences in covariates and in children mortality between the groups of counties benefiting from the program and those excluded **within** the second bandwidth and **before** the implementation of the program in 1965. Do you observe significant differences in the counties characteristics between treated counties and non-treated counties? Given the presence or the absence of differences, what do you conclude on the identification of any effect using the RDD estimation strategy?

Solution: If another variable jumps when the forcing variable crosses the threshold, we cannot attribute changes in the outcome to the program.

Here, we run t-test of differences in means between countries treated and non-treated, for several covariates.

See STATA do-file.

We do observe some clear statistically different means, especially for the urban rate, and the percentage of black population. Treated counties are in average more composed of black people and more rural than non-treated counties. This suggestq that when observation crosses the threshold, variations/jumps in outcomes cannot be fully attributed to other variables than the treatment cutoff, but also to differences in the structure of the counties. By including these variables as control in our regressions we will somewhat control for these differences, although they could be too strong and these differences will not be entirely controlled for. Also, to have a better look, we should maybe conduct these statistical tests on bandwith with a higher threshold to have more precision?

We can also compare our outcome variable, the child mortality **before** the program implementation (variable `rate_5964`). We do not observe clear differences in means between treated and non-treated counties before the implementation

of the Head Start program.

(e) Compare the per-kid total federal county-level spendings in 1968 and 1972. What does this suggest?

Solution: Clear differences in per-kid federal spending, especially in 1968. Average Head Start spending per kid in 1968 is about twice as high as in the countrol counties with 1960 poverty rates within 8 points below the cutoff 300$ vs 137. These differences seem to fade out slightly in 1972, which is also confirmed by differences in means which are not statistically significant at the 5% level but at the 10%.

2 Further graphical analysis

We now implement some deeper graphical analysis to observe (i) the differences in federal spending introduced by the discontinuity, (ii) the differences in health status introduced by the discontinuity.

3. We want to check graphically that spending changed at the discontinuity. To do so, we want to reproduce their Figures II and III, which shows the differences in per-kid spendings between treated and non-treated counties. Instead of their nonparametric estimates of the regression function we can show flexible cubic fits. **We here compare counties with 16% poverty rate above and below the cutoff.**

(a) Generate five variables, first generate the squared and the cubic transformations of the variable `povrate`, and second generate the interaction of `povrate` with the fact of being a treated county (variable `g`), the interaction of the squared transformation and the cubic transformation of `povrate` with the fact of being a treated county (variable `g`).

Solution: See STATA do-file for replication code.

(b) We now estimate the federal spending per-kid in 1968 as a function of poverty rate, the dummy variable identifying treated counties and their interaction term. Write the linear model. How do you interprete the respective coefficients? We explore three specifications: (i) a linear fit , (ii) a quadratic fit, (iii) and a cubic fit. Estimate these models and store the predicted values.

Solution: The model is:

$$\texttt{hsspend\_per\_kid\_68}_i = \alpha + \beta \texttt{g}_i + \gamma \texttt{povrate}_i + \mu \texttt{g}_i \times \texttt{hsspend\_per\_kid\_68} + \nu_i,$$

where, β is the increase in spending associated to the discontinuity in program implementation, γ is the potential increase in spending as counties gets more deprivated, and μ is the change in spending for treated counties as they get more deprivated (change in slope above the cutoff).

See STATA do-file for replication code.

(c) Now group the counties into categories on each side of the cutoff. Five categories need to be constructed. We will then calculate the mean and standard error of the mean of federal spending for these five groups. To construct these categories, generate a variable denoted `bin` which is directly obtained from the following formula: $\texttt{bin} = \left\lfloor \frac{\texttt{povrate}}{4} \right\rfloor \times 4 + 2 + 59.1984$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. Keep observations with the variable `bin` above 40 and below 80.

Solution: See STATA do-file for replication

(d) Now generate two variables which give the mean value by categories of `hsspend_per_kid_68`, denominated `bin_mean` and the standard error of the mean by categories, denominated `stderror`. Compute the 95% confidence intervals for these categories, by generating two variables which gives lower and upper limit of the 95% confidence intervals of the `hsspend_per_kid_68` variable.

Solution: See STATA do-file for replication.

(e) Finally, generate a graphic to replicate figure II. To do so, concentrate on observations with poverty rate in 1960 above 40% and below 80%. Then, for each categories plot the mean value. Then add the confidence intervals, and the three lines corresponding to the three parametric estimates (linear, square and cubic).

Solution: See STATA do-file for replication.

(f) Transform your commands from question 3.(b) to question 3.(e) so that these series of commands are implemented through a program, which allows you to replicate the figure for any other variables than `hsspend_per_kid_68`.

Solution: See STATA do-file for replication.

(g) Replicate the figure for the spendings in 1972 (variable `hsspend_per_kid_72`), and also for the social spending in 1972 (variable `socspend_per_cap72`). From these three figures, what do you conclude regarding the effect of the discontinuity in the Head Start program on county-level federal spending?

Solution: See STATA do-file for replication code

The differences in means of spendings per kids observed in question 2.(e) are driven by a shar drop-off in spending at the cutoff itself. Using our bandwidth of 16% of poverty rate, the linear and quadratic parametric estimates does not fit well the data, while the cubic estimates fits better. We observe the large gap in categories just to the left and right of the cutoff. The differences in means are not statistically significant, but these differences increases as we go further into the bins (groups 3 and 4) and falls down for the fifth group.

By replicating the figure for the other social spendings, we want to identify whether the discontinuities at the cutoff is due to other forms of policies. This seems here unlikely since the decisions to for Head Start assistance in the 300 poorest counties seems to have been made arbitrarily. The figure for the other social spendings show that the discontinuity in other forms of federal social spending at the cutoff is never statistically singnificant and is very small.

4. We now want to assess how child mortality is affected at the discontinuity. To do so, we want to reproduce their Figure IV, which shows the differences in mortality outcomes between treated and non-treated counties. To do so, you can re-use your program constructed in question 3.(f) and apply it to variables `age5_9_sum2`, `age5_9_injury_rate`, `age25plus_sum2` and `rate_5964`. What can you conclude from these figures on the potential performance of the regression discontinuity design estimation strategy?

Solution: Concerning variable `age5_9_sum2`: There is clear evidence of a drop in mortality at the cutoff. According to **?**, linear specifications should be fine if we look at bandwidths of +/- 12 ppt. Beyond h=12 the linear specification yields biased discontinuity estimates.

Concerning variable `age5_9_injury_rate`: The non-linear estimates suggest that there is no discontinuity. The linear estimate in h=20 might be biased. Note that the scale on the vertical axis is different from above.

Concerning variable `age25plus_sum2`: Not much evidence of a discontinuity. Again, somewhat different scale as before.

Concerning variable `rate_5964`: This is the most problematic picture. The linear specification gives no discontinuity but is likely mis-specified. Looking at a linear estimate with h=12 for example would give similar estimates as the quadratic and cubic estimates. The magnitude of the drop relative to the comparision mean is of a similar size as for the `age5_9_sum2` cohort!

3 The implementation of the regression discontinuity estimation strategy

We now estimate the average treatment effect of Head Start program on children mortality.

5. Replicate the parametric results from Table III that correspond to the mortality outcomes in Figure IV we have just replicated (last two columns for bandwidths 8 and 16). Be careful to estimate several linear models for the four mortality outcomes (as in Table III), and by including in your linear model the interaction term between the dummy variables of treated counties and the 1960 poverty rate. What do you conclude on the effect of Head Start on children mortality?

Solution: See STATA do-file for solution.

Some support for the idea that these mortality differences are due to Head Start rather than other factors comes from the fact that we do not observe a similar discontinuity in mortality rates for children five to nine from causes that should not be affected by Head Startnamely, injuries (second panel, Table III). W do not expect the program to affect accident rates for children ages five to nine (not statistically significant for injuries).

Similarly rows 4 and 5 show that there is no discontinuity in mortality from either relevant causes or injuries for people ages 25 and older

However, concerning `rate_5964`, the estimated effect on Head Start causes for children age five to nine during the pre-Head Start years for which we could obtain data (19591964). The parametric models estimate relatively large and sometimes significant effects, but a visual inspection in Figure IV indicates that this is a spurious finding; the right limit fitted by the quadratic polynomial is implausibly low while the nonparametric estimate seems to more accurately reflect the pattern in the data.

6. To improve your estimation of the average treatment effect in question 5., now estimate impacts on the four mortality outcomes using the two different bandwidths and including polynomials in the 1960 poverty rate. How does it change the results?

References