ECON 21110 – Applied Microeconometrics

Juanna Schrøter Joensen
Department of Economics
University of Chicago
jjoensen@uchicago.edu

February 10, 2022

Monday **February 21**, 2022, by 09:30am

The assignment gives a maximum of 100 points which amounts to a maximum of 6% of the total exam score. All questions within each of the three problems are equally weighted in the evaluation. Remember to hand in (i) a write-up with answers to posed questions and (ii) a clearly commented do file that shows how you have solved each of the problems. Remember to present parameter estimates in tables or figures rather than as raw output. See the Syllabus and the course Canvas page for further instructions regarding how the solutions are to be structured. The key to a good score is to be clear, precise, and concise in your answers.

Assignment 5

Problem 1 (Regression Discontinuity Design (RDD) – in Practice) **70 points**

In this problem you will implement part of the analysis from the paper "Politics, Markets, and Schools: Quasi-Experimental Evidence on the Impact of Autonomy and Competition from a Truly Revolutionary UK Reform" by Damon Clark (2004) later published as "The Performance and Competitive Effects of School Autonomy" in 2009 in the *Journal of Political Economy*. The basic idea of the paper can be described very simply. Traditionally schools in the UK have been funded and managed by Local Education Authorities (LEA). In London, this would be a borough (e.g. Camden or Westminster) with rather little in the way of autonomy given to individual schools. But the 1988 Education Act allowed schools to opt out of LEA control and become funded by central not local government with much more autonomy — this was called "Grant-Maintained" (GM). Schools could become GM if a simple majority of parents chose that option in a ballot. So if 51% of parents voted for GM status that school would become a GM-school while if 49% voted for it, it would remain under LEA control. This is the basis of the regression discontinuity design. The paper can be thought of as contributing more generally to the debate about how public institutions like schools or hospitals should be run: should they be given a budget and left to spend it how they want or should they be more tightly controlled? In the case of GM schools, becoming GM resulted not just in more autonomy but also more resources, which were justified as the school now had to deal with some issues that had previously been handled by the LEA and by some people perceived to be

bribes as the government wanted to encourage the growth of GM schools. Thus the change to GM resulted in both more autonomy and possibly more resources.

Reading the paper(s) is not required, though it always helps to better understand the context and potential threats to the identifying assumptions. All references below are to the 2004 job market version of the paper, while the published 2009 version gives you an idea of how a paper can change through the publication process – even if the basic idea still is the same.

The `damonclark.dta` data consists of only three variables:

- `passrate0`: the pass rate of pupils in the school in the year immediately prior to the vote
- `passrate2`: the pass rate of pupils in the school two years after the vote
- `vote`: the percentage vote in favor of the GM status

Preliminaries:

- Generate a dummy variable for a winning vote (`win`) and one for a losing vote (`lose`) in the GM election. 50% is the critical threshold.
- Generate a variable called `margin`, for the margin (difference from the critical threshold) of victory in the vote (negative for a losing vote).
- Generate interactions of `win` with `margin` and margin squared.
- Generate interactions of `lose` with `margin` and margin squared.

Now you have all the necessary variables to perform the basic RD analysis, even though no additional control variables have been provided in the data.

First, you will assess the "Nuts and Bolts" of this Regression Discontinuity Design (RDD).

(a) What are the assumptions underlying the RDD in this paper? What type of RDD is the study using, fuzzy or sharp?

(b) Present a graph similar to the one in Figure 8 of the paper, where you show averages of the change in % Pass over bins of the vote share (set the bins to 2, and add a line at 50 for the discontinuity, no need to also have the lines through the data points). Comment on the graph. *Hint: collapse the dataset in order to easily create the averages you need.*

(c) In Table 3a Clark restricts his sample to those schools with votes in favor of GM status between 15% and 85%. What are the reasons to restrict the sample? Why do subsequent columns of Table 3a include functions of the vote share, both on their own and interacted with the win/lose variable?

(d) Produce a table in the spirit of Table 3a in the paper, that is present results using the specifications in columns 1, 2, 3, 5, of course without weighting or conditioning on other controls, since we do not have them in our data. If the RDD is valid in this case, to what extent is internal validity of the results affected by not conditioning on controls? Use the same sample restriction as in the paper and comment on the results comparing the different specifications.

(e) Experiment with smaller thresholds for sample inclusion w.r.t. the [15,85] chosen by Clark and comment on the difference in the results. What are the trade-offs to be considered here?

(f) Is there a reason for using `passrate0` as the dependent variable? Regress this on the win variable and interpret the result. Does this invalidate the RDD? Explain why (or why not).

(g) In relation to points (e) and (f) what is the only additional relevant evidence that we can provide about the fundamental assumption of RDD given our very limited data? Provide it in graphical form and comment.

(h) RDD can be also construed in an IV framework. The instrument is the forcing variable (here the variable `vote`) which sharply increases the probability of being treated (or switches it from 0 to 1 in a sharp RDD). In this light, discuss what are the implications for the effect that is estimated and comment on internal and external validity of RD designs in general.

Second, you will assess the "Bells and Whistles" of this RDD:

(i) Install the Stata command `rdrobust` and use it to run a local linear regression for RDD with optimal bandwidth choice instead of a polynomial specification that you have used so far. Report the results and comment on the optimal bandwidth choice. Try varying the bandwidth choice for the local linear regression. Report the results and comment on how they compare.

(j) Produce several RDD graphs using the `rdplot` command for different bandwidth choices. Compare to the graph you produced in (b) and comment on the similarities and differences.

Problem 2 (Regression Discontinuity Design (RDD) – in Theory) **28 points**

For each of the the following three papers:

(i) "Van Der Klaauw (2002)

(ii) "DiNardo & Lee (2004)

(iii) "Chay & Greenstone (2005)

also mentioned on slides 5-6 in the Lecture 5 notes. Answer the following questions:

(a) What is the outcome(s) of interest?

(b) What is the running variable, X?

(c) What is the treatment variable, D, and how is it determined by X? Is this a sharp or a fuzzy RDD?

(d) Argue why (or why not) an RDD has more credible identifying assumptions than a multiple linear regression model to answer the causal question(s) of interest?

Problem 3 (THEORY and Final Project) **2 points**

This problem helps you take the next step for the final project. Now that you are (hopefully!) getting familiar with the most important variables in the data, you should be ready to start thinking more carefully about the research question(s) addressed in this paper. Answer the following questions:

(a) What is the main hypothesis tested in this paper?

(b) Is this hypothesis clearly motivated by economic theory? For example, is the paper trying to test an economic theory or distinguish between competing economic theories?

In Assignment 2, Problem 3, you stated the "ideal" randomized experiment to answer the causal question of interest in the paper you are replicating. Now assume that this "ideal" randomized experiment is not feasible in reality.

(c) Propose a RD design that would allow you to estimate the causal effect of interest. Is this a sharp or a fuzzy RDD?