Problem Set 5. Panel Data and Regression Discontinuity Designs
(due Wednesday, December 5)

1. Capital punishment and crime (iii)

In Problem Set 4, you analyzed the association between capital punishment and crime in first-differences. You created a Stata data set with the variables: crime rate, death penalty (binary), number of executions, state, and year, and estimated some preliminary regressions.

a) Now run a (state) fixed-effects regression using the full three-period panel. Estimate it first without, then with year dummies. Compare your results with your three cross-sectional regressions and the first-differences regression from Problem Set 4.

b) Can you think of any omitted variables that could still be biasing the results? Find data on at least one variable that varies both across time and state, and include it in the fixed-effects regression with year dummies. Is the variable significant? Does the main result change?

c) What would happen if you included control variables that varied across states, but not over time, in the regression (for example, average annual number of sunny days in a state)? If you included a variable that varied over time but not across states (for example, national-level policies)?

d) Run your "preferred" specification assuming homoscedastic errors, allowing for heteroscedasticity, and clustering the standard errors by state. Do your conclusions change? Which specification would you report as the main one and why?

e) What can you conclude from your results regarding whether capital punishment seems an effective tool to reduce crime?

f) Explain how you can interpret your specification where the explanatory variable of interest is binary as a difference-in-differences regression.

2. Class size and student achievement

This week we will consider the question of the effect of school inputs on student outcomes. In the course website you will find a data set with information on grades and other variables for 2,000 (3^{rd}-grade) school children in Israel. Our main question will be the effect of class size on grades.

a) Run two simple OLS regressions relating class size with math and verbal grades. Does class size appear to be related to test scores? Interpret the sign, magnitude and significance of the coefficients.

b) Do you think that the OLS assumptions hold? If not, explain the main internal validity problems.

c) Run the same regressions, adding the controls available (percentage of disadvantaged students and enrolment). [Note: Enrolment is the total number of children enrolled in that grade in the whole school, and thus proxies for school and/or cohort size.] Do the coefficients of interest change a lot? What does the direction of the change (the sign of the initial bias) suggest regarding selection into class sizes?

d) We also have school indicators, so we can include school fixed-effects in the regressions. Run the fixed-effects regressions and interpret the results. What happened with the control variables? How did the coefficient of interest change?

e) Do you think that the fixed-effects assumptions hold? Explain the extent to which the specifications that you estimated in sections c and d address the internal validity problems that you described in section b.

f) In Israel, schools tend to follow the "Maimonides rule", i.e., they try not to have class sizes over 40 students. That means that, when more than 40 students enrol in a school, they will typically be split into 2 classrooms, while enrolment under 40 will lead to only one classroom. This generates a discontinuity in class size at enrolment equal to 40 (or multiples of 40). Discuss whether this rule appears to be followed or not, according to the data, and with the help of a graph.

g) We can estimate the effect of interest using the regression discontinuity design generated by the Maimonides rule. Explain why. Create a dummy variable indicating whether enrolment in a given school is over 40, another one for over 80, and a third one for over 120. These are the relevant thresholds. Enrolment is our "running variable". Show summary statistics for the enrolment variable and the three threshold dummies. Explain whether our setup corresponds to a sharp or a fuzzy discontinuity design.

h) Before you run the RDD analysis, there are several validity checks you can run in order to validate the approach. Explain what they are, run them and report the results.

i) Run the RDD analysis for both math and verbal grades. Report the main results and whichever robustness checks you find appropriate. Provide both graphical and numerical results.

j) Compare you RDD results with your OLS and fixed-effects results. Explain the possible reasons why they are so different (or so similar).

k) Out of the three approaches (OLS, fixed-effects and RDD), which one do you think is closer to providing an unbiased and consistent estimate of the causal effect of interest? Justify your answer.

l) With your preferred approach in mind, do you think that you were able to credibly estimate the causal effect of class size on student outcomes? Explain any remaining threats to internal or external validity.