ECON3360 Causal Inference for Microeconometrics

Tutorial 8: Hausman test and regression discontinuity design

Instructor: Julie Moschion

Problem I: construction of panel data and Hausman test

Background The dataset for this exercise comes from the paper by Baltagi and Khanti-Akom (1990) "On Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators", published in Journal of Applied Econometrics, vol. 5, p. 401-406. They demonstrate efficient estimation of returns to schooling, using a yearly panel of 595 individuals observed over the period 1976-1982 drawn from the Panel Study of Income Dynamics (PSID). Use the data in *psidw.csv* for the following questions.

(1) Load *psidw.csv* data to Stata and describe it.

(2) The loaded data is in a wide form. Convert the wide form to a long form with *id* the variable for the cross-section identifier and *year* the variable for the time dimension. Sort the observations by *id* and *year* and have a look at the dataset.

(3) We have four consecutive years of data for each individual. Define the dataset as panel data. Our collected data start from 1976. Recode the variable year to reflect the actual year.

(4) Estimate the following equation using a pooled OLS (with cluster robust standard errors). Compare the results with those obtained from OLS cross-section regressions (i.e. for each year separately).

$$lwage_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 union_{it} + \beta_3 married_{it} + \beta_4 exper_{it} + \beta_5 exper_{it}^2$$

$$+\beta_6 black_{it} + \beta_7 female_{it} + f_t + c_i + u_{it}$$

(5) Estimate the same equation, now using: (i) random effects and (ii) fixed effects estimators (with cluster robust standard errors). do they look different? Run those first without time fixed effects, then add time fixed effects. Does adding time fixed effects change the results?

(6) Perform a Hausman test and choose between the random effects and fixed effects estimators.

(7) Why can't we estimate the coefficients of *educ*, *black* and *female* variables with the fixed effects estimators?

Problem II: regression discontinuity design (sharp RD)

Background This exercise is based on the study: "The performance and competitive effects of school autonomy", D. Clark (2009), *Journal of Political Economy* Vol 117, No 4.
The study contributes 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? Traditionally schools in the UK have been funded and managed by the Local Education Authorities (in London, this would be a borough e.g. Camden, Westminster) with little autonomy given to each school. 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). In addition to more autonomy, GM schools were also given more resources as they now had to deal with issues that were previously handled by the LEA.
Schools could become GM if the majority of parents chose that option in a ballot. In other words, if 51% of parents voted for GM status that school would become a GM-school, while if only 49% voted for it, it would remain under LEA control. This allows an estimation of the impact of the policy change on student achievement using a regression discontinuity design.
In 1992, school performance tables (league tables) were published for the first time providing information on student achievement: the fraction of grade 11 students who pass five or more GCSE examinations (referred to in the paper as the "school pass rate"). Use the data in schautonomy1.dta for the following questions.

(1) Describe and summarize the data.

(2) Estimate the following model using an OLS regression (with robust standard errors). Interpret the coefficient. Is it causal?

$$dpass_i = \beta_0 + \beta_1 GM_i + u_i$$

where $dpass_i$ is the change in school i's pass rate after the introduction of GMs (versus before) and GM_i is a dummy variable which takes on a value of one if the school is a GM school and zero otherwise.

(3) The dataset also includes *vote*, the exact percentage of votes in favour of converting to a GM school. Represent the relationship between *vote* and *dpass* using a scatter-plot.

(4) Plot the relationship between *vote* and *dpass* adding a linear fit and a quadratic fit before and after the threshold. Can we expect the RD estimate to differ depending on whether a quadratic term is included in the regression?

(5) How would you exploit the additional information provided by *vote* to estimate the causal effect of GM status on a school's pass rate? Provide the appropriate regression specification using robust standard errors. Are quadratic terms necessary?

(6) What assumptions need to hold for your approach to estimate the causal effect in 5?

(7) Instead of using *dpass* as the outcome variable, re-run your preferred specification using *passrate*2 as the outcome variable. Does the coefficient on "win" changes sign?

(8) Someone critical of the results suggests using *passrate*0 as the dependent variable. Suppose that the coefficient of the win variable was *significant*. What would this suggest?