

# EC406 Economic Policy Analysis

## Problem Sets, 2007

**John Van Reenen and Rajeev Dhejia**

During Michaelmas term, you are expected to complete two sets of written work (the first is to be handed in by week 7, your class teacher will tell you when the second should be handed in). Choose two pieces from the five questions below. You are also expected to participate in the class presentations (see below for topics).

### **(1) Structured essay (3-4 sides of A4)**

- a. What is sample selection bias and why is it a potential concern in evaluating economic policies?
- b. Critically evaluate the analysis of the relationship between earnings and college education in Willis and Rosen (1979)

### **(2) Practical Econometrics Exercise: Wages and Education**

Use the Stata econometrics package and the dataset `card.dta` available from <http://econ.lse.ac.uk/staff/sredding/EC406teach/CARD.DTA> to run the following regressions.

See the brief Stata tutorial in the Appendix and further references therein.

Record the output of your regressions in a log file (see the Appendix at the end of this problem set). Open the log file using Notepad or Microsoft Word and bring a print out of the log file to the relevant class.

- a. Run a simple OLS regression of  $\log(\text{wage})$  on education. What do you find? If education were strictly exogenous, how would you interpret these results?
- b. Run a multivariate regression of  $\log(\text{wage})$  on education and the other controls specified on pages 497-499 of Wooldridge. What do you find? What do we learn from comparing these results with those in the specification in (a) above?
- c. Run a 2SLS regression of  $\log(\text{wage})$  on education and the other controls specified on pages 497-499 of Wooldridge. Education is assumed to be the only endogenous variable and is instrumented with college proximity. What do you find? What do we learn from comparing the results with those in the specification in (b) above?
- d. Estimate the first-stage regression and the reduced form regression. What do you find? What do we learn from these estimates?

### **(3) Practical econometrics exercise: Crime**

Use the Stata econometrics package and the dataset `crime4.dta` available from <http://econ.lse.ac.uk/staff/sredding/EC406teach/CRIME4.DTA> to run the following regressions.

See the brief Stata tutorial in the Appendix and further references therein.

Record the output of your regressions in a log file (see the Appendix at the end of this problem set). Open the log file using Notepad or Microsoft Word and bring a print out of the log file to the relevant class.

- a. Use the commands `tab year` and `tab county`
  - i. How many years and counties are included in the dataset?
- b. Run a pooled cross-section OLS regression of the log crime rate on a constant and time dummies for the years 1982-87:
  - i. How should the constant be interpreted in this regression?
  - ii. How should the estimated coefficient on the year dummy for 1987 be interpreted?
- c. Run a pooled cross-section OLS regression of the log crime rate on the log estimated probability of arrest and time dummies for the years 1982-87:
  - i. What do you find?
  - ii. How should the constant be interpreted in this regression?
- d. Run a pooled first-differenced cross-section OLS regression of the change in the log crime rate on the change in the log estimated probability of arrest and time dummies for the years 1982-87:
  - i. Why does Stata drop one of the time dummies from the estimated equation?
  - ii. How should the estimated coefficients on the year dummies in this specification be interpreted?
  - iii. What do we learn from comparing the estimated coefficient on the change in the log estimated probability of arrest in this specification to the estimated coefficient on the log estimated probability of arrest in the previous specification?
- e. Run a fixed effects (within groups) regression of the log crime rate on the log estimated probability of arrest and time dummies for the years 1982-87, where the fixed effects included in the specification are for counties.
  - i. What do you find? What do we learn from comparing the estimated coefficient on the log estimated probability of arrest in this specification with the estimated coefficients in specifications (c) and (d) above?
- f. Re-estimate the specifications in (c)-(e) including the full set of controls specified on pages 451-2 of Wooldridge.
  - i. What do you find? What do we learn from comparing the estimated coefficients to those without including the full set of controls?

#### **(4) Randomized Experiment Computer Exercise**

The data for this exercise comes from Project STAR: A Randomized Experiment of the Impact of Class Size Reductions on Pupil Achievement. This was a 4-year experiment in Tennessee designed to evaluate the effect of class size on learning. Each participating school had at least one 'control' group class and one 'treatment' group.

Some details of this experimental design can be found in Stock and Watson, pp390-391 or at [http://wps.aw.com/aw\\_stockwatsn\\_economtrcs\\_1/0,7018,291498-,00.html](http://wps.aw.com/aw_stockwatsn_economtrcs_1/0,7018,291498-,00.html)

The dataset for this exercise (a reduced version of the full data set) can be downloaded from the course website.

1. Run a regression of the test score in kindergarten on whether you were in a small class (we are going to ignore the other treatment – being in a small class with aide). This can be written as:

```
. reg tscorek sck
```

Show that the OLS estimate of the intercept in this regression will be the mean of the test score for those who are in the control group and that the coefficient on sck will be the difference between the mean test score for those in treatment and control groups.

2. What does this regression say about the impact of class size reductions on pupil performance? Estimate the regression but with robust standard errors:

```
. reg tscorek sck, robust
```

Why does the coefficient remain the same? Why does the standard error change?

3. Compute the mean level of pupil achievement for treatment and control groups. Work out the variance of these sample means. You can get all the information you need by typing the STATA command:

```
. tab sck, su(tscorek)
```

Work out the difference in means and the standard error of this difference. How do they relate to the estimates of the treatment effects derived from the regressions.

#### 4. *Other Covariates*

Put in other covariates – boy, freelunk, totexpk and a set of dummies for the school. You can generate the school dummies by typing:

```
. tab schidkn, ge(id)
```

What would you expect to happen to the estimate of the treatment effect? Is this what happens?

#### 5. *Heterogeneity in Treatment Effects*

Generate the interaction of the control group with boy, freelunk and totexpk.

Estimate the regression excluding the school dummies but including these interactions. Interpret the coefficients.

Use an F-test to test the hypothesis that there is the same treatment effect for everyone. Do you accept or reject the test of homogeneity. For what groups does the treatment effect seem to be largest?

Compute the estimated average treatment effect for everyone in the sample. You can do this by typing:

```
.      g te = _b[sck] + _b[sck_boy]*boy + _b[sck_freelunk]*freelunk +  
_b[sck_totexpk]*totexpk if e(sample)
```

after the estimation. Why is the average of this similar in size to the treatment effect we estimated when we did not include any covariates? Why is it different?

### (5) Regression Discontinuity Computer Exercise

The data from this paper is taken from Damon Clark “Politics, Markets and Schools: Quasi-Experimental Evidence on the Impact of Autonomy and Competition from a Truly Revolutionary UK Reform”.

This can be downloaded from <http://www.nber.org/~confer/2005/si2005/ch/clark.pdf>

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 (in London, this would be a borough e.g. Camden, 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’. 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 but which some people felt were bribes as the government wanted to encourage the growth of GM schools. So the change to GM resulted in both more autonomy and possibly more resources.

The data set consists of a small number of 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
- dpass: the change in the pass rate = passrate2-passrate0
- vote: the percentage vote in favour of the GM status
- win: a dummy variable if the vote was more than 50%
- win\_vote = win\*vote

- win\_vote\_2 = win\*(vote squared)
- lose\_vote = lose\*vote
- lose\_vote\_2 = lose\*(vote squared)

1. Do a scatter-plot of the change in the pass rate on the vote in favour of GM status. You can do this using the STATA command:

. scatter dpass vote

2. Reproduce the result in the first column of Table 3a of Clark

**Table 3a: Impact of GM Status on Pass Rates of Schools that become Grant-Maintained: Two Years after Base Year**

|                         | Least Squares Regression                                 |                  |                   |                   |                     |                   | Least Abs Der     |                           |
|-------------------------|--|------------------|-------------------|-------------------|---------------------|-------------------|-------------------|---------------------------|
|                         | Non-Grammar Schools with Vote Shares in [15,85] interval |                  |                   |                   |                     |                   | All Schools       | Non-Grammar Votes [15,85] |
| Win                     | 2.169<br>(0.636)   | 4.032<br>(1.367) | 3.894<br>(1.392)  | 3.297<br>(1.332)  | 2.721<br>(2.190)    | 3.454<br>(1.339)  | 3.188<br>(1.206)  | 2.698<br>(1.591)          |
| Vote                    |  | -5.3<br>(3.2)    |                   |                   |                     |                   |                   |                           |
| Vote*Loss               |  |                  | -2.693<br>(5.568) | -2.966<br>(5.324) | -16.418<br>(22.617) | -3.208<br>(5.241) | -4.251<br>(4.802) | -6.053<br>(6.735)         |
| Vote*Win                |  |                  | -6.424<br>(3.937) | -3.652<br>(3.915) | 17.945<br>(16.758)  | -5.012<br>(3.872) | -2.768<br>(2.659) | 0.831<br>(4.563)          |
| Vote <sup>2</sup> *Loss |  |                  |                   |                   | 40.750<br>(63.377)  |                   |                   |                           |
| Vote <sup>2</sup> *Win  |  |                  |                   |                   | 0.006<br>(0.004)    |                   |                   |                           |
| SES change              |  |                  |                   |                   |                     | -0.366<br>(0.094) | -0.296<br>(0.086) | -0.320<br>(0.084)         |
| Weighted Polynomial     | N<br>None  | N<br>Linear      | N<br>Linear       | Y<br>*Win         | Y<br>*Win           | Y<br>*Win         | Y<br>*Win         | Y<br>*Win                 |
| Controls                | N  | N                | N                 | N                 | N                   | Y                 | Y                 | Y                         |
| N                       | 524  | 524              | 524               | 524               | 524                 | 524               | 729               | 524                       |
| R-sq                    | 0.02   | 0.03             | 0.03              | 0.02              | 0.03                | 0.03              | 0.07              | 0.04                      |

Notes: Robust standard errors in parentheses. Vote share is divided by 100. SES change proxied by Free School Meal take-up. Additional controls are school type dummies and vote year dummies.

In this Table, Clark restricts his sample to those schools with votes in favour of GM status between 15% and 85%. Why did you think he chose this sample restriction? Why do subsequent columns of Table 3a include functions of the vote share, both on their own and interacted with the win/lose variable?

Experiment with thresholds for sample inclusion that differ from the [15,85] chosen by Clark— how different are the results. What are the trade-offs to be considered here? Why is the information in the scatter-plot useful in considering what specification and sample to use?

3. Instead of using dpass as the outcome variable, repeat your analyses using passrate2 as the outcome variable. What does the theory of regression discontinuity say about the comparison of the results with this outcome variable compared to the previous set of results? How do they compare in practice? Explain this.

4. Someone critical of the results suggests using passrate0 as the dependent variable. They show that if one just regresses this on the win variable this has a significant negative coefficient. They argue this invalidates the regression discontinuity design because winning should be uncorrelated with variables prior to the treatment. Evaluate this argument.

## **Class Presentations**

The class will be divided into two groups. One group presents the paper to the other group each week. Choose among yourselves how to divide up the presentation. Even the group who are not presenting should read the paper. The first 6 weeks will be papers and the last 2 weeks will be going through the answers to the written work.

### **(1) Cost Benefit analysis of congestion charging**

Undertake a cost-benefit analysis of London's congestion charge. Do the social benefits of the congestion charge exceed the social costs? Is the current scheme optimal? If not, how could it be improved?

See reading list for details

### **(2) Labour regulation**

Consider the paper by Besley and Burgess (2004)

- a. What do we learn from comparing the effects of labour regulation in the different columns of Table 3?
- b. What potential econometric problems are addressed in Tables 4 and 5?
- c. Do you find the analysis in Besley and Burgess (2002) convincing? What (if any) are the remaining potential econometric problems that are not successfully addressed by their analysis?

### **(3) Education and wages**

- a. Explain how compulsory schooling laws can be used to estimate the rate of return to education
- b. What do we learn from a comparison of the odd and even numbered columns in Table V of Angrist and Krueger (1991)? What is the motivation for the different specifications included in this table?
- c. Leaving the potential problem of weak instruments to one side, is the approach taken by Angrist and Krueger (1991) convincing? Why or why not?

### **(4) Labour Market Programs and unemployment**

- a. The "New Deal for Young Unemployed" is a UK labour market program analyzed in Blundell et al (2004). What are its aims and what are the economic rationale behind these
- b. Discuss the identification strategies used by Blundell et al (2004) to construct "comparison groups".
- c. What other types of analyses would you perform if you had data through to 2004?
- d. If you were advising the Labour government prior to the introduction of the New Deal what would you advise them to do in order to see if the program worked?

**(5) What should be the policy of the UK (or some other OECD country) on immigration?**

You should make sure that you cover the following issues:

- how much immigration should be allowed?
- what sort of immigration should be allowed?
- Should the impact on other countries be considered?

Though there are plenty of other aspects to think about as well.

**(6) Are European-style welfare states doomed?**

You should make sure that you cover the following issues:

- What are the costs and benefits of European-style welfare states?
- What are the threats to its viability?
- What changes should or have to be made?

Though there are plenty of other aspects to think about as well.

## Appendix: Brief Stata tutorial

Detailed information is available from the Stata user's guide. See also <http://rlab.lse.ac.uk/itsupport/>. Stata is available as a networked resource from LSE public computers.

- To open a dataset in Stata, you must first clear the computer memory by typing
  - `clear`
- The command to open a dataset is
  - `open card.dta`
- When you have opened the dataset type `describe` and you will see all the variable names and their labels listed
- The command to run an OLS regression is
  - `regress [dependent variable] [independent variables]`
- The command to run an IV regression is
  - `ivreg`
- The command to run a fixed effects regressions is
  - `areg [dependent variable] [independent variables] ,  
absorb([cross section unit])`
- For further help on using these commands, type for example
  - `help regress`
- To record the results of your regressions type the following command before running your regressions (this saves the file to the `c:\` drive)
  - `log using c:\filename.log,replace`
- When you have run your regressions, type the following command to close an open log file
  - `log close`