FLS 6415: Class 7 Homework

October 19, 2017

Remember to answer all the questions in R markdown and produce a PDF. Email your completed homework (R markdown file and PDF) to jonnyphillips@gmail.com by midnight the night before class. Remember to refer to the example code from this week and the last couple of weeks for coding guidance.

Dube and Vargas (2008) use a complex difference-in-differences approach. We will take a simpler approach to the same question using their dataset. The variables we will use are listed below:

Variable	Description
year	Year
origmun_name	Municipality Name
paratt	Number of paramilitary attacks
lop	Domestic Price of Oil
oilprod88	Oil production in the municipality (in 1988)
linternalp	Domestic Price of Coffee
cofint	Area of coffee-growing in the municipality (in hectares)

1. Load the Dube and Vargas (2008) data. Our two main dimensions of variation are across time (*year*) and across geographic municipalities (*origmun_name*) so report the number of distinct (unique) values in the dataset for each of these variables.

2. Our treatment will focus first on oil, and specifically on an increase in the oil price. To identify 'before' treatment and 'after' treatment points first produce a graph of how the international oil price (*lop*) changes over time. *Hint:* In the data, the international oil price varies over time but is constant across municipalities, so the value is just repeated for every municipality in the same year.

3. Your graph in Q2 should show a sharp rise in the oil price from 1998 to 2005. So we are going to define 1998 as our 'before' treatment point in time (when oil prices are low), and 2005 as our 'after' treatment point in time (when oil prices are high). Create a separate dataset which is filtered so that it only includes data for 1998 and 2005 (we are ignoring all the years between for our simplified analysis). Create a new variable that defines units for 2005 as 'After' and units for 1998 as 'Before'.

4. We can now define treatment and control units by whether they would be affected by a change in the price of oil or not. What percentage of the units in your data produce oil (*oilprod88*)? Define a new variable so each unit that produced oil is 'treated' and any unit that did not produce oil is 'control'.

5. We now have a suitable dataset for a differences-in-differences analysis. Our outcome is the number of paramilitary attacks (*paratt*). But first let's consider the naive 'observational' study that we might conduct if we had never heard of diff-in-diff. Use a regression to compare the average number of paramilitary attacks in treated units with the average number of paramilitary attacks in control units. Interpret the results and what it suggests for the effect of oil income on violence. *Hint:* You can ignore differences over time so no need to do anything with the *year* variable; we can just imagine it provides multiple measures of the outcome.

6. Provide one reason why your answer to Q5 may not be an accurate estimate of the causal effect of oil income on violence.

7. Another naive methodology would be to simply compare the 'before' and 'after' values of the outcome variable for the treated units alone. In this approach, we are using each

municipality as its own control, but at an earlier point in time when oil prices were lower. Conduct a regression to compare paramilitary attacks before and after the oil price rise for the treated units alone. Interpret your findings an compare them to your findings in Q5.

8. Provide one reason why your answer to Q7 may not be an accurate estimate of the causal effect of oil income on violence.

9. A difference-in-differences methodology simply combines the two sources of variation you explored in Q5 and Q7: variation in whether the unit can be affected by treatment (treated and control units) and variation over time (before and after treatment). Create a 2x2 table summarising the average number of paramilitary attacks by both treatment status (treated/control) and time (before/after treatment). Use the table to calculate a difference-in-differences estimate and interpret that estimate.

10. The best way to understand differences-in-differences results is often with a chart. Plot the values from your table in Q9 in a chart, where time (before/after) is on the x-axis, the outcome is on the y-axis and there are two lines, one for the control group and one for the treated group. *Hint:* The graph may be easier to plot and interpret if you use *year* as the x-axis variable, rather than a character variable such as 'after'/'before'.

11. The logic of the difference-in-differences methodology is that the change over time in the outcome for the control units is an appropriate counterfactual for what would have happened to the treatment units in the absence of treatment. Add an additional line to your chart that shows the counterfactual outcome for the treated units. Use this line to estimate the difference in the outcome between the treatment units and the predicted counterfactual in 2005. *Hint:* This line starts at the 'before' value for the treated units and runs parallel to the line for the control units. Use `add_row` to add new rows to your table and define their treatment status as 'Counterfactual'.

12. We can implement exactly the same analysis using a regression which interacts time (before/after) with treatment status (treatment/control). Interpret this regression. *Hint:* To make the regression easier to interpret, define your "Before/after" variable is a factor variable with the baseline level as 'Before', eg. `factor(time,levels=c("Before","After"))`. Also remember any regression with an interaction needs three terms; each variable on its own and the interaction.

13. The main assumption in our regression is that paramilitary attacks would have changed at the same rate in both treatment and control units in the absence of treatment (i.e. if oil prices did not change). This requires that the pre-treatment *trends* (not levels) of the outcome variable are comparable in treatment and control units. Go back to the original dataset and create a chart that compares paramilitary attacks over time for all years for treated and control units. Add a vertical line to show the start of treatment in 1998. What does the chart suggest about the plausibility of the parallel trends assumption?

14. Now we will conduct the same analysis for coffee. Create a chart of local coffee prices (*linternalp*) over time.

15. The chart in Q14 shows coffee prices fell from around 1998 to 2003. Define Before/After status for the units where before is 1998 and after is 2003 (and get rid of the other years). Next define treatment and control units depending on whether the municipality grows any coffee (*cofint* measures the number of hectares of coffee grown in the municipality).

16. Create the difference-in-differences chart showing the change in the outcome over time for the treated and control groups separately (the equivalent to the chart in Q10 for the coffee prices treatment). Interpret the chart.

17. Implement the difference-in-differences regression (same as in Q12) for the effect of a fall in coffee prices on paramilitary attacks. Interpret the results.

18. Assess the pre-treatment parallel trends assumption for the differences-in-differences regression in Q16. *Hint:* You'll need to go back to the original data, plot the outcomes for the treated and control groups as in Q13, but this time with treated and control defined by coffee production, not oil production.