Advanced Applied Econometrics
OECD, November 2010
Thijs van Rens

The Experimental Approach and Difference-in-Differences

Exercise 1. What is the source of identifying variation? Below you find several causal questions that can be or have been estimated using regression analysis. In each case, find the difference-in-difference (DD) estimator that corresponds to the regression. Describe the treatment and control group that are used to identify the treatment effect and argue whether we can credibly argue that individual observations have been randomly assigned to these groups. Think of one or more potential problems with the random selection. How could we test for these problems?

1. To estimate the effect of class size on students performance, we regress the SAT scores (tests taken before going to college in the US) of students i that are currently at college, on the average student-teacher ratio in the district where these students report their home address, controlling for some of the information they mentioned on their college application: gender, age, race, ethnicity, and whether the student applied for financial aid.

2. We are interested in the size of exchange rate passthrough. We have data on exchange rates X_t and detailed product-level data on retail prices P_{it}. We want to regress prices on exchange rates. Since we expect products, of which a larger fraction is imported, to be affected more by the exchange rate, we use input-output matrices to construct an index for the fraction of each product category that is imported F_i. Then, we regress P_{it} on $F_i * X_t$, controlling for industry and time dummies.

3. In their paper *Precautionary Savings and Self-Selection: Evidence from the German Reunification "Experiment",* Nicola and Matthias Schündeln (QJE, 2005) are interested "to test the theory of precautionary savings and to quantify the importance of self-selection into occupations due to differences in risk aversion." They "exploit the fact that for individuals from the former German Democratic Republic (GDR) German reunification in 1990 caused an exogenous reassignment of income risks," so that for those households self-selection in occupational choice should not be an issue. They then run the following regression:

$$\begin{aligned}\log(W) &= \beta_0 + \gamma_0 D + \beta_1 \cdot \text{risk} + \gamma_1 D * \text{risk} \\ &\quad + \beta_2 \log(P) + \gamma_2 D * \log(P) + Z'\beta_3 + (D * Z)'\gamma_3 + \varepsilon\end{aligned}$$

   where W is wealth, P is permanent income, and Z is a vector of household characteristics and year dummies. A civil servant dummy, which is equal to one if the main income earner is a civil servant, is used as the measure of risk and D is a dummy indicating whether a household came from the former GDR.

4. In *Sectoral Labour Market Effects 2006 FIFA World Cup,* Arne Feddersen and Wolfgang Maennig try to estimate the employment effects of the 2006 World Cup in Germany. They have highly disaggregated data on employment by county c and industry i over time t.

(a) Feddersen and Maennig focus their analysis on a few industries, in which the employment effects may be expected to be large (construction, hospitality) and ran the following regression (by industry):

$$\log\left(E_{ct}\right) = \beta\left(D_c^{\text{stad}} * D_t^{\text{post}}\right) + D_c'\gamma_1 + D_t'\gamma_2 + \varepsilon_{ct}$$

where E is employment, D^{stad} is a dummy indicating whether a county had a stadium, D^{post} is a dummy that takes the value 1 after 2006:q2, which is when the world cup took place. D_c and D_t represent dummies for each county and time period respectively.

(b) In my discussion of this paper (available on my website), I suggested they run instead a regression of the following form:

$$\log\left(E_{ct}\right) = \beta\left(D_i^{\text{nt}} * D_t^{\text{post}}\right) + D_i'\gamma_1 + D_t'\gamma_2 + \varepsilon_{ct}$$

where D_i^{nt} indicates non-tradable goods producing industries and D_i are industry dummies.

Excercise 2. Difference-in-difference estimation For this exercise, you may use cross-country panel data on educational attainment, GDP and inequality, which are available from my website at www.crei.cat/~vanrens/educ. Alternatively, you can use your own data, in which case you may have to slightly modify the questions.

Suppose we are interested in the causal effect of education on inequality.

1. Using the variation across countries, construct a difference estimator and estimate effect of education on inequality. What are potential problems with this estimator?

2. Using variation over time, construct another difference estimator and estimate the effect. What are potential problems with this estimator? Can you explain the differences in the findings with the previous estimator?

3. Using variation both across counties and over time, construct a DD estimator. Given the problems with the difference estimators in parts 1 and 2, can you explain the changes in the results? What are the potential problems with the new estimates?

4. Implement the DD from the previous part as a regression. Verify you get the same results.

5. Generalize the regression in the previous part using all available variation in education and inequality. How does this change the results? Why?

6. Suppose our model predicts that there is an effect of education on inequality only for poor countries. Test this prediction of the model using GDP data.

7. Consider again the same model prediction. Instead of testing the prediction, use it to construct a DDD estimate of the effect of education on inequality. Assuming the model prediction is true, why is this estimator more credible than the simple DD estimator?

8. Would it be fair to say that the results in part 6 justify the estimator in part 7?