Computer assignment on DiD

The data for this computer assignment are random samples from those generated by Blundell and Costa Dias (2009) to perform Monte Carlo analysis used to illustrate the methods explained in their survey paper. We consider the data were generated for the case that the unobservables u and v are *negatively correlated*. We assume that in region 0 (*x=0*) there is no subsidy for participating in advanced education, while in region 1 (*x=1*) there is no subsidy in period 0 (*t=0*), but in period 1 (*t=0*) students who exceed the cutoff value of 4 (*s>4*) are eligible for a subsidy if they enrol in higher education. The indicator *sub=1* if *s>4* and *sub=0* otherwise, and this independently of time or region, i.e. for both *t=0* or *t=1* and *x=0 or x=1.* It is therefore an indicator of the *eligibility group* to which the individual belongs. The indicator *sub1=1* only in case *t=1* and defines therefore the *eligibility status*. In region 1 this is the true eligibility status: the individual is eligible for a subsidy if she participates in advanced education. In region 0 this is a *placebo* eligibility status, since there is no subsidy assigned in region 0. The indicator *d=1* in case the individual effectively participates in higher education and *d=0* otherwise: *d* is the *treatment indicator*. We therefore have an "encouragement design" (also called a "fuzzy" DiD): the subsidy encourages the individuals to enrol in advanced education. We have seen in the lectures that such a design can identify the Local Average Treatment Effect (LATE), i.e. the return to advanced education, of students who are induced to enrol in advanced education by the subsidy (see also Blundell and Costa Dias (2009, p. 588-593).

We consider two cases. In Part 1 we consider the case that the subsidy is not anticipated from birth, while in Part 2 we assume that students who live in region 1 and have an age to enrol in advanced education in period *t=1* knew already from birth that they would be eligible to a subsidy is case their test score would cross the aforementioned threshold: they can therefore exert effort to make it more likely to pass this cutoff and, hence, to obtain the subsidy in case of enrolment in advanced education. In both cases we aim at estimating the return of advanced education on labour market earnings. "ly" defines the logarithm of the observed labour market earnings after graduating from basic or advanced education, depending on whether *d=0* or *d=1*.

Notice that Table 2 of Blundell and Costa Dias (2009, p. 592) reports the true treatment effects. In the assignment we will use in Parts 1 and 2 the same 4 random cross section samples of the population in the two regions and periods to estimate the LATE.

Finally, please report throughout this assignment standard errors of parameters that are robust to heteroscedasticity, i.e. using the option r in the estimation commands.

Part 1: An unexpected subsidy in region 1 and period 1

Use the dataset "assignment_DiD1.dta" for this part of the assignment.

QUESTION 1

Perform an OLS regression of ly on a constant term, the treatment indicator *d* and the time indicator *t* using the data of region 1 (*x=1*) only. Does the coefficient of d consistently estimate the return to advanced education? Briefly justify your response.

QUESTION 2

2.1 Calculate the *aggregate effect of the subsidy on the treatment group* (also called the "Intention to Treat" (ITT) effect) in region 1 (***x=1***) by first taking the differences over time of the mean outcomes "ly" for the groups eligible and not eligible for a subsidy (i.e. for *sub=1 or sub=0*) and then by taking the difference of these differences. How does this compare to the true outcome displayed in Table 2 of Blundell and Costa Dias (2009)?

Would the outcome differ if you would first take the differences between the groups of the mean outcomes for time *t=1* and *t=0* and then take the difference of these differences? Why (not)?

2.2 Calculate this point estimate of the ITT effect by performing OLS on an appropriate linear regression. Don't forget (as mentioned earlier) to report the heteroscedasticity robust standard errors. Why would you prefer this method to the previous one?

QUESTION 3

3.1 Check by a linear regression the assumption that in the absence of a subsidy there is no time trend in the participation in advanced education in region 1 (*x=1*). What is the P-value of the test that there is no such time trend?

3.2 Under the Assumption DID1 and the additional assumptions of monotonicity and the absence of a time trend in the counterfactual of no subsidy, Blundell and Costa Dias (2009) show that one can identify the local average treatment effect (LATE) of the individuals who are induced to enrol in advanced education by dividing the ITT effect by the proportion of individuals who enrol in higher education as a consequence of the subsidy. Does the subsidy induce a significant increase in the proportion of individuals enrolling in advanced education? Calculate the point estimate of this LATE using the sampled individuals in region 1.

3.3 We now show that we can estimate in region 1 the point estimate of the LATE reported in 3.2 by a 2SLS regression. To do this form first for each individual in *t=1* and *x=1* the time difference of the outcome "*ly*" between the individual outcome in time *t=1* (denoted by "*ly_1*") and the estimate of the *conditional* average of the outcome in time *t=0*, where the conditioning is on the eligibility group (*sub*): *dly=ly_1-E[ly|t=0, sub, x=1]*.[1] Next, form for *t=1* and *x=1* the corresponding difference for the treatment indicator *d* for individuals belonging to the eligible group *sub=1*: *dd=d_1-E[d|t=0, sub, x=1].* For individuals belonging to the group that is not eligible for the subsidy, i.e. for *sub=0*, set *dd=0.* The latter comes from the assumption of no time trend, as tested in 3.1. Estimate a linear regression of *dly* on a constant term and *dd* using *sub* as an instrumental variable (IV). Why would you prefer estimating the 2SLS procedure to the procedure proposed in 3.2?

QUESTION 4

4.1 Redo the OLS regression that estimates the ITT effect of the subsidy in region 1 of question 2.2 by adding *z* as an explanatory variable. Why would you want to do this?

4.2 Estimate an ITT effect of the subsidy in region 1 on *z,* i.e. taking *z* as the dependent variable. Why would you be interested in doing this?

[1] If we would have access to panel data instead of to repeated cross section data, then one could just form the individual within differences of the outcome and participation indicator and apply the 2SLS estimator on these differences instead.

QUESTION 5

5.1 Use the control region *x=0* to estimate by OLS the placebo ITT effect in the control region. What do you conclude?

5.2 Use the 2SLS to estimate the placebo LATE in region 0 as in question 3.3 for region 1. Comment your findings.

5.3 The identification of the DiD estimator relies on the assumption that the error term of the regression model is additively separable. We have chosen to express the outcome as the logarithm of earnings, because we know that the true model is generated by a such a log-linear model with an additive error term. In general we do not know, however, how the data are generated. We might therefore want to check for some alternative specification choices (or estimate a nonlinear CiC model). One potential alternative choice is a model that is linear instead of log-linear in earnings. A check therefore is how this linear model performs in the placebo region. Estimate therefore the ITT effect for the linear model as applied to region 0. Comment your finding.

Part 2: An expected subsidy in region 1 and period 1

Use the dataset "assignment_DiD2.dta" for this part of the assignment.

QUESTION 6:

6.1 Estimate (as in 2.2) the ITT effect in region 1 by estimating an appropriate linear regression model that contrasts the evolution over time of the eligible group (*sub=1*) to the ineligible group (*sub=0*). Contrast this to the true effect. Explain your findings.

6.2 Estimate (as in 4.2) an ITT effect of the subsidy in region 1 on *z,* i.e. taking *z* as the dependent variable. Why would you be interested in doing this?

QUESTION 7:

7.1 Estimate by linear regression the ITT effect of the subsidy by contrasting the average earnings between region 1 and region 0 (rather than contrasting the eligible group with the ineligible group within region 1). Briefly discuss your findings.

7.2 Estimate as in 6.2 the ITT effect of the subsidy in region 1 based like in 7.1 on the contrast between regions (instead of within region 1). Briefly discuss your findings.

7.3 Estimate by 2SLS the LATE corresponding to the ITT effect found in 7.1. Briefly discuss your findings.