Exercise 2

Kwame Adjei-Mantey, kwamemantey@gmail.com
Yoshiki Nakajima, ynakajima146@gmail.com

2018/07/18

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

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.

> The result shows that the coefficient for higher education, d, is 1.078 with significance level of 1%, indicating that those who get higher education in region 1 earn average of 1.078 higher in log scale.

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)?

> The observations in region 1 can be split into four groups based on sub and t, and denoted as $\overline{ly}_{0,0}, \overline{ly}_{0,1}, \overline{ly}_{1,0}$ and $\overline{ly}_{0,0}$ where subscripts represents sub and t, respectively. ITT is then calculated $(\overline{ly}_{1,1} - \overline{ly}_{1,0}) - (\overline{ly}_{0,1} - \overline{ly}_{0,0}) = (\overline{ly}_{1,1} - \overline{ly}_{0,1}) - (\overline{ly}_{1,1} - \overline{ly}_{0,0})$. The result of this two approaches is identical but interpretation differs. In the first approaches, the counterfactual observation is build on the assumption that the macro effect, m_t is same toward both groups (d=0 and sub=1), while second approach assumes the difference between the groups, u_{sub} is same over time.
> The difference the difference in dataset. Original paper used all the observations while the result here restricts to region 1. Also noted is the difficulty of estimation when unobserved correlation exists between u and v.

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?

> ITT can be estimated by OLS with; $y_i = \beta_0 + \beta_1 sub_i + \beta_2 t_i + \beta_3 sub_i t_i$, where β_3 can be interpreted as ITT. This method is preferred because we can get standard error with just a few lines of code.

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?

> The coefficient of t in question 2.2 tells us the existence of time trend. According to the result, coefficient of t is -.009 with standard error of .016. The p-value of t test therefore is 0.577 which fails to reject the null hypothesis to conclude that there is no time trend in region 1.

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.

Recall that in case of incomplete compliance, DID identifies ITT and we can get LATE by dividing the fraction of participants

$$\text{LATE} = \frac{E[\hat{\alpha}^{\text{DID}}]}{(p_{11} - p_{10}) - (p_{01} - p_{00})}$$

Since we know from 3.1 that there is no time trend in region 0, we can assume $(p_{11} - p_{10}) = 0$. Therefore LATE is estimated to be .456.

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?

We transform the variable as instructed to get `dly` and `dd`. Then we implement 2SLS using the stata command `ivregress` to get the estimated LATE of .513. We prefer this method because we can exploit all the observatios while the approach in 3.2 relies on those who changed the education attainment baceuse of subsidy, hence liable to be affected by the extremes. Another upside of this method is we can estimate standard error based on which we can discuss the significance of the effect of subsidy.

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?

The coefficient of family background, z, is significantly positive, indicating that those from rich family background tend to earn more. This extra variable can be considered important because of the possible correlation between family background and the test score (eligibility, sub).

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?

The result shows that all the variables are not significant, indicating region and subsidy status is not a factor of family background.

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?

Unlike those in region 1, people in region 0 is not entitled to receive subsidy even if they get a test score higher than the threshold mark. The OLS result shows the coefficient for t*sub is not significant, while this term was significant in region 1. This indicates that the actual distribution of subsidy is in fact important to the increase of income.

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

By following the same steps in 3.3, we get LATE of .477 with standard error of .321. T-test fails to reject the null hypothesis and we conclude there is no LATE. This is a consistent to the findings in 5.1.

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.

The result shows that ITT estimator is -.589 with standard error of .340. Although this is not significant in 5% level, but we can reject t-test if we use 10% significant level. This does not make sense because we expect this to be insignificant because there is no policy implementation in region 0. This problem occurred because the misspecification of the model violates the additive separability assumption, deteriorating the accuracy of estimation.

Part 2

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.

The IIT estimator is .328 and is significant in 1% level. This is larger than the true estimate of .203. The effect tend to be to larger in case of expected treatment because individuals can adjust their effort to meet the threshold to get subsidy. This finding is consistent with Blundell and Costa Dias paper.

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?

The result shows that the subsidy positively affects family background. This is a different result from 4.2 where the introduction of subsidy was not significantly different from zero. Since subsidy is only provided for those scores higher than the threshold and go to university, it makes economic sense that eligibility status has positive correlation to family background.

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.

To compare the difference between the regions, we take t, x and t*x as an explanatory variable. Estimated ITT is .028 with standard error of .020. Corresponding t-statistics and p-value is 1.43 and 0.15, failing to reject the null hypothesis of zero ITT. This is probably explained by the small

proportion of the eligibles (12.1%). The effect could be crowded out when we take the average of the region.

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.

The result shows the ITT effect is insignificant probably because of the crowding out of averaging. Also noted that x has a negative significant coefficient indicating that region 1 is less z score on average.

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

2SLS based LATE is identified using the same approach as 3.3. Here, since we compare regions, we replace from 3.3 `sub` to `x` and remove `x==1` restriction. Estimated LATE is .839