Econometrics I (Second Half): Problem Set 2

Due: December 8th at 11:00AM

Question 1: Difference-in-Differences Estimator and its Matching Counterpart

Our goal is to quantify the effect of a job training program (the national Job Training Partnership Act – JTPA – in the United States) on individuals' earnings. This question is motivated by the analysis conducted in the 1997 paper by Heckman, Ichimura and Todd titled *Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme*. For this question, we will use the data in the csv file named `jtpa_ps2.csv` which has information on individuals' earnings around the assignment into the JTPA program. While the data contains information on the age and education of the participants, we will abstract from the use of covariates to focus on the construction of the estimators of choice. The variables needed are id (person identifier), d (indicator of treatment assignment, $= 1$ for treated individuals), qtr (quarter relative to treatment assignment, with -6 denoting 6 quarters before treatment assignment), $earn$ (monthly earnings), and p (propensity score provided to us – note that this is specific to each individual and fixed over time). It is worth noting that there is no data for the year in which the treatment assignment was done ($qtr = 0$).

(a) Compute the difference-in-differences estimator manually by *computing and reporting* – in each of the cases below – the following: $\bar{y}^1_{t_1}$ (mean earnings for individuals assigned to treatment during the quarters after the treatment assignment was made), $\bar{y}^1_{t_0}$ (mean earnings for individuals assigned to treatment during the quarters before the treatment assignment was made), $\bar{y}^0_{t_1}$ (mean earnings for individuals *not* assigned to treatment during the quarters after the treatment assignment was made), $\bar{y}^0_{t_0}$ (mean earnings for individuals *not* assigned to treatment during the quarters before the treatment assignment was made)

C1: Restrict the analysis to include only observations for $t \in [-4, 4]$, report $\widehat{\alpha}^{DID}_4 = (\bar{y}^1_{t_1} - \bar{y}^1_{t_0}) - (\bar{y}^0_{t_1} - \bar{y}^0_{t_0})$.

C2: Restrict the analysis to include only observations for $t \in [-5, 5]$, report $\widehat{\alpha}^{DID}_5 = (\bar{y}^1_{t_1} - \bar{y}^1_{t_0}) - (\bar{y}^0_{t_1} - \bar{y}^0_{t_0})$.

C3: Restrict the analysis to include only observations for $t \in [-6, 6]$, report $\widehat{\alpha}^{DID}_6 = (\bar{y}^1_{t_1} - \bar{y}^1_{t_0}) - (\bar{y}^0_{t_1} - \bar{y}^0_{t_0})$.

Do you notice a pattern? Explain.

(b) For each of the cases (C1-C3) covered in part (a), run the following linear regression

$$Y_{it} = \beta_0 + \beta_1 d_i + \beta_2 Post_t + \alpha(Post_t \times d_i) + \epsilon_{it} \tag{1}$$

where $Post_t = \mathbb{1}[qtr > 0]$. Using your results from part (a), verify that your OLS estimates for α coincide with the estimates for α^{DID} you calculated manually and interpret the following in terms of $\bar{y}^1_{t_1}$, $\bar{y}^1_{t_0}$, $\bar{y}^0_{t_1}$, $\bar{y}^0_{t_0}$:

- $\widehat{\beta}_0$
- $\widehat{\beta}_1$
- $\widehat{\beta}_2$

(c) So far, we have obtained the DID estimates of the ATT of JTPA on individuals' earnings. Since we have data on earnings (outcomes) for more than 1 quarter before the treatment assignment, we want to check for violations of the parallel trends assumption behind the DID. Plot the mean earnings per quarter for treated and non-treated individuals (in the same graph) and describe the trends you observe. **NOTE:** This is not a trick question. You are on the right track if you notice something suspicious going on before the treatment assignment was made.

(d) We have enough reasons to suspect that significant pre-treatment differences between un-treated and treated individuals – potentially driving individuals' program participation decision – might be contaminating our estimates of the ATT of JTPA. We then want to implement the Matching DID (MDID) estimator which explicitly accounts for the participation decision by exploiting information on each individuals' probability to participate in the program. In our case, this is captured in the variable p. Using p, compute and report $\widehat{\alpha}_4^{MDID}$, $\widehat{\alpha}_5^{MDID}$, and $\widehat{\alpha}_6^{MDID}$ (for each of the cases covered in part (a)) implementing the following matching algorithms:

- 2 nearest neighbor matching (with replacement)
- Epanechnikov kernel matching using Silverman's rule of thumb for bandwidth selection

Interpret your results.

(e) Recall from our lecture that one of the main assumptions of the MDID relies on the implementation of the estimator over a region of common support. This requires implementing the estimator using only observations for which their propensity scores fall within the region in which the distribution of p for treated and un-treated overlap. Plot the distribution of p for both treated and un-treated individuals (separately). What is the range of p for treated individuals? What is it for their un-treated counterparts? Using a minmax rule, define what the region of common support would be in this case.

(f) Using only observations in the region of common support defined in part (e), re-do part (d). Compare your results with those obtained in part (d) and interpret your results.

Question 2: Event Study Design

We are going to quantify the dynamic impact of COVID-19 on married women's quarterly labor market rates. For this exercise, use the data contained in the csv file `enoe_q219-q122_married_female`. The file includes the variables $newid$ which captures a women's identifier in the survey, $time$ which captures the quarter of survey, eda capturing age, $dent2 - dent32$ denoting state-specific dummy variables, $dchild2\_12$ denoting an indicator for the presence of children in the household younger than 12, edu which contains information on educational attainment, $inact$ being an indicator of inactivity, $unemp$ an indicator of unemployment, $formal\_new$ an indicator of formal employment and $informal\_new$ an indicator of informal employment.

$$Y_{ist} = \sum_{j \in [-3,-2] \cup [0,7]} \alpha_j D_{it}^j + \boldsymbol{\beta} \mathbf{X}_{it} + \eta_s + \epsilon_{ist} \tag{2}$$

where X_{it} includes a constant, $age, edu, edusq, dchild2\_12$. η_s denotes state-specific fixed effects.

(a) Letting $time = 4$ denote the quarter capturing the onset of COVID-19, define an event-time variable centered around this quarter so that $event = time - 4$. To check that this has been defined correctly, compute and report the mean of the variable $event$ by $time$. Before implementing the event study design in 2, generate a set of dummies specific to each event time (for example, $Dm3 = 1$ if $event = -3$ and 0 otherwise, $D1 = 1$ if $event = 1$ and 0 otherwise, and so on).

(b) Using the specification described in 2, run an event study of

- Unemployment and inactivity on pandemic quarters
- Formal employment and informal employment on pandemic quarters

As a normalization, leave out the dummy variable corresponding to the quarter just before the onset of COVID-19. Make a plot of the coefficients associated with the event dummies with confidence intervals. Interpret your results

(c) Notice that the specification described in 2 does not include quarter-specific fixed effects. Explain why adding quarter-specific fixed effects is not necessary in this application.