ECON 8000/9000 Empirical Energy Economics
PS6: LP-DID Example
Christy Zhou

Instruction and Notes:

- Deadline: Sunday 3/22 @ 5pm
- Submission: Please compile all results into a single PDF.
- Group: Everyone can work together, but only **up to two students** can submit the assignment together. If so, make sure you include both names at the top when you submit.
- Here I write instructions using Stata. You can use other software if you like.
- This study has a replication package online by Daniel Girardi (GitHub). I recommend you follow the instructions below, which are slightly simpler than the online package.

Q1. Implement Dube et al. (2025) "A Local Projections Approach to Diff-in-Diff", J of Applied Econometrics

We will estimate Equation (1) below for event times h from -5 to 8 (except for $h = -1$). In this exercise, we will assume the impact of the event begins at time 0. We will normalize the outcome variable, taking an FD relative to Y_{t-1}.

$$\Delta Y_{i,t-1,t+h} = \beta_h \Delta D_{it} + \gamma Z_{it} + \phi_t + \varepsilon_{it} \quad \cdots\cdots \quad (1)$$

While there is an LPDID package, estimating it manually will offer the maximum flexibility (e.g., if you wish to implement LPDID with instrumental variable(s), or use a standard error that is not allowed in the package). In this exercise, we will estimate Equation (1) manually with different event time h.

We will estimate Equation (1) in three versions: In version 1, we will estimate it naively without specifying a clean control condition. In version 2, we will estimate it under the most strict clean-control condition, as suggested by Dube et al. (2025).

$$\begin{cases} \text{C1:Newly treated} & \Delta D_{it} = 1 \\ \text{C2: Clean control} & D_{i,t+h=0} \end{cases}$$

In version 3, we will estimate it using a slightly compromised clean-control condition. We will enforce that units are not yet treated within the next 3 periods. This somewhat compromised version is popular in practice, as the most strict version will remove a lot of observables, and researchers may end up with an unreliable estimate with a very large SE as h increases.

$$\begin{cases} \text{C1:Newly treated} & \Delta D_{it} = 1 \\ \text{C2: Clean control} & D_{i,t+3=0} \end{cases}$$

Info:

- Use data: "PSdata_Dube_et_al_JAppMetrics.dta". This is not the data used in the publication, but data generated on their website.
- Variables:
  - Outcome yvar: Y and treatment D_{it}: treat
  - Unit id and time are called: unit and time
  - I also generate an event time relative to treatment time (date_entry_first), denoted as event_time.
  - We don't have controls to include in Z_{it} to simplicity

Tasks:

- 1. Estimate Equation (1) V1, aka naively. V1 will be a guided exercise.
  - Follow my lines 42-47 to generate some $\Delta Y_{i,t-1,t+h}$ for both pre- and post-periods. You can also generate them outside of a loop to see how they are created.
  - Second, follow my lines 54-57 to generate variables that you will use in Tasks 2-3 for clean control.
  - Third, we will create some empty slots to store beta and confidence intervals (lines 60-78). In R you can create a matrix or a dataframe to store them. In Stata, it is easier to store them in the same data (although it will make your data look messy).

```
38    * 1.  Gen vars
39    * 1.1 Gen xvar
40    gen D_treat = D.treat
41
42    * 1.2 Gen yvar:
43    // Pre-period
44    forvalue j = 2/5 {
45        gen Dm`j'y = L`j'.Y - L.Y
46    }
47
48    // Post-period
49    forvalue j = 0/8 {
50        gen D`j'y = F`j'.Y - L.Y
51    }
52
53
54    * 1.3 Forward vars to define not yet treated unit
55    forvalue j = 0/8 {
56        gen F`j'treat = F`j'.treat
57    }
58
59
60    * 2. Record LPDID estimates
61    gen     t = _n - 5 - 1
62    replace t = .      if t > 8
63
64    gen b_naive       = .
65    gen ci_naive_low  = .
66    gen ci_naive_upp  = .
67    replace b_naive   = 0 if t == -1
68
69
70    gen b_cleanctrl1        = .
71    gen ci_cleanctrl1_low   = .
72    gen ci_cleanctrl1_upp   = .
73    replace b_cleanctrl1    = 0 if t == -1
74
75    gen b_cleanctrl2        = .
76    gen ci_cleanctrl2_low   = .
77    gen ci_cleanctrl2_upp   = .
78    replace b_cleanctrl2    = 0 if t == -1
79
80
81
82    * 3.  Estimation
83    * 3.1 Naive LPDID: No Clean Control
84    forvalues j = 0/8{
85        qui reghdfe D`j'y D_treat, absorb(time) vce(cluser unit)
86        qui replace b_naive      = _b[D_treat]                       if t == `j'
87        qui replace ci_naive_low = _b[D_treat] - 1.96*_se[D_treat] if t == `j'
88        qui replace ci_naive_upp = _b[D_treat] + 1.96*_se[D_treat] if t == `j'
89    }
90
91
92    forvalues j = 2/5{
93        qui reghdfe Dm`j'y D_treat, absorb(time) vce(cluser unit)
94        qui replace b_naive      = _b[D_treat]                       if t == - `j'
95        qui replace ci_naive_low = _b[D_treat] - 1.96*_se[D_treat] if t == - `j'
96        qui replace ci_naive_upp = _b[D_treat] + 1.96*_se[D_treat] if t == - `j'
97    }
98
99
100
101   twoway  (connected b_naive t               if inrange(t, -5, 8), lc(black) mc(black)) ///
102           (rcap ci_naive_low ci_naive_upp t if inrange(t, -5, 8), fcolor(black%75) lcolor(black%75)), ///
103           graphregion(color(white)) plotregion(color(white)) legend(off) ///
104           yline(0, lc(black) lp(shortdash) lw(thin)) xline(-1, lc(black) lp(dot) lw(thin)) ///
105           xtitle("Event Time") xlabel(-5(1)8) scale(1.25)
106   graph export "$myanalysis_output/Q1_lpdid_naive.pdf", replace
```

- Last, run estimation in lines 82-97 and plotting in lines 101-106. (Note: if you were to normalize to a different time, say $h = -2$, then your yvar needs to be created differently. Also, it is easier to write your pre-periods regression as foreach j of numlist 1 3/5{} (to skip h=-2) rather than in a forvalue loop.

- 2. Estimate Equation (2) V2, aka, with the most strict clean control conditions.
    - It means that either C1 or C2 must be satisfied. You need to add if conditions to lines 84-97 to produce LP-DID with clean control.
    - For both pre-period and post-period regression, C1 ($\Delta D_{it} = 1$) simply means D_treat == 1
    - For pre-period regression, clean control C2 will include (i) all never-treated units and (ii) units that are not yet treated. This means you can impose treat == 0
    - For post-period regression, clean control C2 will need to be F`j'.treat == 0, which means for regression on h=5, you will need to include a unit that won't be treated within the next 5 periods.
- 3. Estimate Equation (1) V3, aka, with the slightly weaker clean control conditions that all units shall not be treated in the next 5 periods.
    - Hint: This estimation will be extremely similar to V2, and you only need to change one line of your code.
- 4. Compare three sets of estimates that you plot. Explain what generates the difference between them.