Problem Set 6: Diff-in-diff

MGMT 737

1. **Diff-in-diff:** We first consider a simple diff-in-diff with one treatment, using our dataset from Mian and Sufi's 2009 paper from Homework 3. We study the impact of supply elasticity on house prices. [You may use standard regression packages for this] Recall that `hpi_all` is the house price measure, and `pop` is the population in the county.

   (a) Using the years 1998 to 2006, estimate the impact of supply elasticity on log house prices, using a two-way fixed effects approach with county and year FE, estimating the impact of the supply elasticity for the 2002 to 2006 period (e.g. treat 2002 as the first treated year), and weight by the county population. Report this effect, and its standard error (make a case for what you think the appropriate standard error adjustment is).

   (b) Now using the same sample, estimate the same regression with county + year FE, but estimate the year-by-year effect of supply elasticity on log house prices, relative to 2001. Report the effect in 2005, and interpret the coefficient.

   (c) How would you test the pre-trend condition? (You can ignore the uniform testing issue from class) Report the p-value for this test.

   (d) Re-estimate the same specification as part (b), but use the level of house prices (in thousands of dollars). Do the pre-trends look different?

   (e) Finally, use the dummy variable `elas_bin` (which is split on the median value of elasticity) and reestimate part (b) on log house prices.

2. **Event Study** Next, we consider an event study approach. We will use data from Sun and Abraham (2021)'s application, which replicates results from Dobkin et al. (2020)'s results using the HRS data (which is publicly available). Variables: `hhidpn` household identifier – this is the identifier for an individual in the panel; `wave` time identifier (wave of survey) – this is the time index of the survey; `wave_hosp` time of event – time when the individual is hospitalized; and `oop_spend` Out-of-pocket spending.

   (a) We will be following Sun and Abraham's notation for describing this setup. Denote the initial time period of treatment for a unit as E_i. What variable corresponds to E_i in our dataset? Construct a variable $D_{it} = 1(E_i <= t)$ which is equal to one when an individual is treated. What share of individuals are treated in period 7,8,9,10?

   (b) Estimate the traditional static two-way fixed effects estimation for this setup:

   $$Y_{it} = \alpha_i + \lambda_t + D_{it}\beta + \epsilon_{it} \tag{1}$$

   where Y_{it} is `oop_spend`, α_i is a unit fixed effect and λ_t is a time fixed effect. Report the estimate for β and its standard error (adjust for appropriate inference in the panel setting).

   (c) Now, consider the estimation group by group. Denote our control group as the last group ever treated. For each other treated wave, estimate the treatment effect relative to this group, excluding the last period of data. Report the coefficients and standard errors for each of these waves. How do these estimates compare to your last result? For Wave 8 cohort, what is the relative comparison period for the diff-in-diff? In other words, the diff across units is Cohort Wave 8 vs. Cohort Wave 11. What is the diff across time comparing?

   (d) Now thing back to the traditional static equation – what is the relative comparison period for this diff-in-diff?

   (e) We now consider the dyanamic versions of Equation 1. Denote $D^l_{it} = 1(t - E_i = l)$

   $$Y_{it} = \alpha_i + \lambda_t + \sum_{l \in -3,-2} D^l_{it}\beta_l + \sum_{l \in 0,1,2,3} D^l_{it}\beta_l + \epsilon_{it}. \tag{2}$$

   Report the β coefficients and their standard errors.

(f) Now, repeat this exercise, but consider the estimation group-by-group again. Focusing just on the Cohort Wave 8 vs. Cohort Wave 11 comparison, how would you run the above specification? What coefficients are you able to estimate? Report these estimates. Now repeat and estimate β_0 for each of the groups. How do these estimates compare to your estimates from Equation 2?

(g) Now focus on the estimate for β_{-2} from Equation 2. This is traditionally the pre-trend test. Sun and Abraham show that under the standard diff-in-diff assumptions, the β_{-2} coefficient in Equation 2 specification, this coefficient is the weighted combination of multiple treatments in other periods. Denote $CATT_{e,l}$ as the average treatment effect l periods from the initial treatment for the cohort of units first treated at time e. Then, Sun and Abraham show that

$$\beta_{-2} = \sum_{e=8}^{11} \omega_{e,-2}^{-2} CATT_{e,-2} + \sum_{l=-3,0,1,2,3} \sum_{e=8}^{11} \omega_{e,l}^{-2} CATT_{e,l} + \sum_{l' \in \{-4,-1\}} \sum_{e=8}^{11} \omega_{e,l'}^{-2} CATT_{e,l'}, \tag{3}$$

where the ω are weights that we can calculate. We can estimate these by replacing Y_{it} in Equation 2 with $D_{i,t}^{l} 1(E_i = e)$ as the outcome variable, and reporting the coefficient on D_{it}^{-2}. Do so for each l and e. Your results should match Figure 2 in Sun and Abraham. How does this affect your interpretation of the pre-trend test?

(h) Finally, we estimate Sun and Abraham's alternative estimator, which avoids the contamination bias. This approach *pools* our cohort-by-cohort comparison from before. First, we estimate

$$Y_{it} = \alpha_i + \lambda_t + \sum_{e=8,9,10} \sum_{l=-3, l \neq -1}^{l=3} 1(E_i = e) \times D_{it}^{l} \delta_{e,l} + \epsilon_{it}, \tag{4}$$

where we exclude the last time period and treat the Cohort Wave 11 as our control group. Take the $\delta_{e,l}$ estimates, and report $\delta_{e,0}$ for all 3 groups. The final estimate μ_0 weights each of these δ by the cohort sample weight $\pi_e = Pr(E_i = e | l = 0)$. Report this estimate of μ_0.