Problem Set 4: Diff-in-diff

MGMT 737

1. **Diff-in-diff:** Consider the following simulated data in `dind_data.csv`. There are 10 time periods (`timeid`), 1000 units (`ids`), and the treatment turns on in period 5 (`post`). The treated group (`treated_group`) receives the treatment in period 5 and the control group does not. The treatment is fully absorbing. For this exercise, you may use canned linear regression packages, or your own constructed estimator. Please be specific on what standard errors you are reporting.

   (a) First, focus on `y_instant` and estimate the treatment effect for the following three regressions:

   - The treatment effect, controlling for group status and the post period $y_{it} = \alpha_{post} + \alpha_{treated} + \beta \texttt{post} \times \texttt{treated}$
   - The treatment effect, controlling for group status and time fixed effects $y_{it} = \alpha_t + \alpha_{treated} + \beta \texttt{post} \times \texttt{treated}$
   - The treatment effect, controlling for unit and time fixed effects $y_{it} = \alpha_t + \alpha_i + \beta \texttt{post} \times \texttt{treated}$

   Compare the point estimates and standard errors. How do they differ? Why? Plot the means of the outcome over time, split out by treatment group, to help explain your answer.

   (b) Now, we want to estimate the dynamic effects for this outcome. Estimate the effect of the treatment relative to time period 4, for all time periods, controlling for unit and time fixed effects. Plot these coefficients, and report the estimate and standard error in period 6.

   N.B. Verify to yourself that you must omit the estimated effect in a given time period. What would your estimator estimate if you failed to omit a given period?

   (c) How does your point estimate from part a with the same specification compare to taking the simple average of the dynamic coefficients from the regression in part b?

   (d) Now estimate the effect of the treatment relative to time period 3. How does the estimate change? Use your figure to explain why.

   (e) Now, we want to estimate the effect for the `y_dynamic` outcome. Replicate part a and b using the this outcome. Explain why this outcome looks different (consider plotting the means as before).

   (f) Now replicate part b and d using `y_dynamic2`. Do you think `y_dynamic2` satisfies the necessary condiitions for diff-in-diff?

   (g) Consider the standard errors that you chose in this exercise. Repeat part a, but try with both robust standard errors, and clustered by id. Report the difference in standard errors for the estimate of the treatment effect, controlling for unit and time fixed effects.

   To get intuition on this result, do the following exercise:

   - Calculate the residuals from the specification
   - Estimate the autocorrelation within unit by period

   What does the autocorrelation structure imply about how the shocks line up with the treatment timing?

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

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

where the ω are weights that we can calculate. We can estimate these by replacing Y_{it} in Equation 2 with $D^l_{i,t} 1(E_i = e)$ as the outcome variable, and reporting the coefficient on D^{-2}_{it}. 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

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

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.