Homework 5

Economics 7103

You have access to an imaginary dataset on hourly household energy consumption from June 2021 (*energy_staggered.dta*). You are interested in estimating the effect of adoption of a smart energy device on energy consumption. Adoption of the device was voluntary, so you are worried about endogenous selection on gains (Roy selection). Moreover, selection is staggered. In your data, you have the following variables:

Variable	Description
datetime	Date and time string
temperature	Outdoor temperature in degrees F
precipitation	Inches of precipitation
relativehumidity	Relative humidity (%)
zip	Zip code of the household (10 zip codes observed)
id	Household identifier
size	Square feet of home
occupants	Number of occupants
devicegroup	Binary indicator for eventual adoption of the device
treatment	Binary indicator equal to one for all hours after adoption of the device
energy	kWh of electricity consumed

Table 1: Variable descriptions.

1 Hourly data - Stata

1. Generate a Stata time variable using the `clock` command. Generate a treatment cohort variable and report the number of treated cohorts. *Hints*: Make sure to create a double precision date variable to not lose precision when working with Stata times. The `egen` command will probably be most helpful in creating a cohort variable. Finally, it may be helpful to create a simpler time variable that counts up from one each hour using the `group` function from egen on time-sorted data.

2. Install and use the `fetwowayweights` command to estimate the TWFE weights. Make treatment cohort the group variable and use the `feTR` option. Report the number of negative weights.

3. Estimate the following regression:

$$Y_{i,t} = \alpha_i + \lambda_t + \beta D_{i,t} + \gamma X_{i,t} + \varepsilon_{i,t} \tag{1}$$

   where $Y_{i,t}$ is energy consumption, α_i is a household fixed effect, λ_t i is a set of hourly indicators, and $X_{i,t}$ is a vector of time varying controls including temperature, precipitation, and relative humidity. Report the estimated ATT and clustered standard error. What level did you cluster at and why?

2 Daily data - Stata

Collapse the data down to daily data. You should create a dataset with average daily temperature, humidity, and precipitation, the sum of energy consumed, a treatment variable that is equal to one for the first day of treatment and all days thereafter, and a new treatment cohort variable. *Hints*: You will want a daily date variable. Use `help datetime`. You should end up with 30,000 observations.

1. Estimate the same TWFE-style regression as above, but on the daily data. Report your new estimate of the ATT and the clustered standard error. How much do these differ?

2. Estimate the following event-study regression using `reghdfe`:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\tau=-30}^{-2} \beta_\tau^{pre} 1(R_{i,t} = \tau) + \sum_{\tau=0}^{30} \beta_\tau^{post} 1(R_{i,t} = \tau) + \gamma X_{i,t} + \varepsilon_{i,t} \tag{2}$$

   where $R_{i,t}$ is the relative time to the first-treatment period. Cluster at the correct level. Display the event-study coefficients using `coefplot`. Do not plot coefficients of the control variables. Your plot should be somewhat clear, but do not spend hours trying to get everything perfect. *Hints*: Generate a time to treatment variable and then create lead and lag indicator variables. You should write loops to do this, not 60 lines of code.

3. Replicate your results exactly using the command `eventdd`. Display the plot automatically generated by the command. *Hint*: You should specify to `eventdd` that you would like it to use the `hdfe` option and the correct standard errors.

4. Implement the Callaway and Sant'Anna estimator using the `CSDID` package. Use the default doubly robust estimator (you must specify covariates) and 50 bootstrap replications. Plot your results using the postestimation command `csdid_plot`. *Hint*: This took three or four minutes for my computer to run, and may take longer on other machines.