Problem Set 1

Panel Data and Program Evaluation

Due on 2020/12/02

I. Econometrics

1. Let's think about the problem of serial correlation using a simple example. Consider the following sample-mean model for the panel data:

$$Y_{it} = \beta + u_{it}$$

where $t = 1, \cdots, T$ for all i (the panel is balanced). The structure of u_{it} is $u_{it} = \theta_i + \epsilon_{it}$. θ_i is uncorrelated with ϵ_{it}; $Var(\theta) = \sigma_\theta^2$; $Var(\epsilon_{it}) = \sigma_\epsilon^2$.

   (a) Assume that u_{it} is i.i.d, write down the OLS estimate for β (denote it as $\hat{\beta}$) and the OLS estimate of its variance, $\widehat{Var}\left(\hat{\beta}\right)$.

   (b) Write down the actual variance of $Var\left(\hat{\beta}\right)$ and compare it with $\widehat{Var}\left(\hat{\beta}\right)$,which one is larger?

(c) What is the estimate of the variance of $\hat{\beta}$ if we cluster the standard error by each individual i?

2. Suppose we use following model to evaluate a certain policy:

$$Y_{it} = \alpha_0 + \alpha\tau_{g(i)t} + \delta t + \theta_i + \varepsilon_{it}$$

where $g(i) = 1$ is i is treated and zero otherwise, and $\tau_{g(i)t} = 1$ if $g(i)$=1 and the policy is enacted before t.

(a) What is the FE estimator for α? You may work out the case of two periods and then consider the general case with multiple periods.

(b) Compare the FE estimator with the FD (first-difference) estimator for α.

II. Programming exercises

This exercise requires you to partly replicate the results in Bertrand, Duflo, and Mullainathan (2004,QJE) using a simulated dataset. When you work on this exercise, Angrist and Pischke (2008, Chapter 8) and Cameron and Miller (2015) are good references. Also, make sure that you have read the jupyter notebooks for bootstrap methods and Monte Carlo simulation. Though you can work in groups, it's better for you to work independently on coding and then cross check it with your teammates. You may choose from STATA and MATLAB to complete the exercise.

1. *Simulate the Data.* Let y_{igt} be the outcome variable for individual i in group

g (such as a city, province, region, etc.) by time t (such as a month, quarter, or year), τ_{gt} be a dummy for whether the policy intervention has affected group s at time t. The model we consider for DID analysis is usually given by

$$y_{igt} = \alpha + \beta\tau_{gt} + \mathbf{X}'_{igt}\gamma + \theta_i + \delta_t + \epsilon_{igt} \tag{1}$$

For simplicity, we simulate datasets using a following DGP (Data Generation Process):

$$y_{igt} = \gamma x_{igt} + \theta_i + \delta_t + \epsilon_{igt} \tag{2}$$

for $i = 1, \cdots N$, $t = 2010, \cdots T$, and $g = 1, \cdots, G$. The data are simulated using rules as below:

(a) The seed is set to be 10101.

(b) $N = 20,000$, $T = 2019$. We consider the case that $G = 60$.

(c) x_{igt} is draw from a uniform distribution on $[0,\ 1]$, and $\gamma = 0.6$.

(d) $\theta_i \sim N\left(\mu_\theta,\ \sigma_\theta^2\right)$. We choose $\mu_\theta = 0.6$ and $\sigma_\theta = 0.5$;

(e) $\delta_t = \delta(t - 2019)$ and $\delta = 0.1$;

(f) $\epsilon_{igt} = v_{gt} + u_{it}$, where $u_{it} \sim N\left(0,\ \sigma_u^2\right)$ and $\sigma_u = 1$. v_{gt} follows an $AR(1)$ process such that

$$v_{gt+1} = \rho v_{gt} + \xi_{gt} \tag{3}$$

where $\rho = 0.8$ and $\xi_{gt} \sim N\left(0,\ \sigma_\xi^2\right)$. We choose σ_ξ to be 0.2. To start with the simulate, we choose v_{gt} from a normal distribution $N\ (0,\ 0.1)$ for $t = 1$. For $t > 1$, we use equation (3) to simulate forward and construct the

data. (By the way, this technique is used quite often in macroeconomics where economic shocks are modelled as Markov process.)

(g) After you obtain data on x_{igt}, θ_i, δ_t, ϵ_{igt}, you can simulate y_{igt} using equation (2).

(h) Without sorting the data, generate individual, year, group indicators to identify each observation.

Through (a) to (f), you may generate a dataset containing $x_{igt}, \theta_i, \delta_t, \epsilon_{igt}$, and three indicators for individual, year, and group. Then you can use equation (2) to obtain y_{igt}. Save these variables in a dataset named as `simdata.dta` (or `simdata.mat` for MATLAB users). Extract the first 20 groups and save it as `simdata_sub.dta` (or `simdata_sub.mat` for MATLAB users)

2. *Monte Carlo Simulation for OLS Estimates.* Set the seed to be 12345 to initialize the random number generator. Using the simulated data `simdata.dta`, you need to generate a pseudo treatment variable τ_{it} by randomly assign 30 groups to be treated starting from a year draw from a uniform distribution between 2013 and 2016. So τ_{it}=1 for for affected state and after the intervention period. For example, if the group id is 20 and the random draw of year is 2014.1, then τ_{it} equals one if $g = 20$ and $t \geq 2015$. If the year draw is an integer, choose the starting year as the exact year. Then we perform the OLS regression using the model

$$y_{igt} = \beta_0 + \beta\tau_{igt} + \gamma x_{igt} + \delta_g + \delta_t + \epsilon_{igt}$$

where δ_g is the group-specific fixed effects.

- First, repeat this exercise 400 times by clustering the standard error at individual level. In each time, you need to record and store the coefficient estimate for β and the t-statistic, as well as compute the auto-correlations $corr\left(\hat{\epsilon}_{igt}, \hat{\epsilon}_{igt-1}\right)$ for j=1,2,3. Based on the results, you need to:

(a) Plot a histogram of the estimated β by selecting the number of bins to be 50, indicating its mean value and value of zero with vertical lines;

(b) Plot histograms of the estimated first-order, second-order, and third–order auto-correlations, by selecting the number of bins to be 50, indicating its corresponding theoretical value (ρ, ρ^2, ρ^3) with vertical lines on the graph.

(c) Calculate the rejection rate, which is defined as the number of exercises with t-statistic greater than 1.96 divided by 400. Is it larger than 5%?(Note that if OLS were to provide consistent estimates and standard errors, we would expect to reject the null hypothesis of no effect ($\beta = 0$) roughly 5% of the time if we choose the threshold of the $t-$statistic to be 1.96)

- Second, Repeat (a) to (c) by clustering the standard error at group-year level and group level.
- Third, use the simulated data `simdata_sub.mat` generate a pseudo treatment variable τ_{it} by randomly assign 10 groups to be treated starting

from a year draw from a uniform distribution between 2013 and 2016. Repeat exercises (a) to (c) by clustering the standard error at the individual level, group-year level, and group level.

3. *Monte Carlo Simulation for FE Estimates.* The policy variable is generated as in 2 using `simdata.dta`. Now we estimate a two-way fixed effects model

$$y_{igt} = \beta_0 + \beta\tau_{igt} + \gamma x_{igt} + \theta_i + \delta_t + \epsilon_{igt}$$

use the mean-differencing estimator. Repeat exercises (a) to (c) by clustering the standard errors at individual, group-year, and group level. Compare the results with we obtained from OLS regression, explain their differences (or similarities).