Problem Set 1

Answer Key

Fall 2020

I. Econometrics

1. (a) The OLS estimator is given by the following minimization problem:

$$\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \beta I_{it}\right)^2,$$

where $I_{it} = 1$ is a constant for all i and t. Taking the derivative with respect to β; the first-order condition implies that

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{N} \sum_{t=1}^{T} I_{it}^2\right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} I_{it} Y_{it}\right) \\ &= \frac{1}{NT} \sum_{t=1}^{N} \sum_{t=1}^{T} Y_{it} = \bar{Y}\end{aligned}$$

Substituting Y_{it} with $\beta + u_{it}$, we have

$$\widehat{Var\left(\hat{\beta}\right)} = \frac{Var\left(\beta + u_{it}\right)}{NT} = \frac{\sigma_\theta^2 + \sigma_\epsilon^2}{NT},$$

where the second equality holds because we assume that u_{it} is i.i.d.

(b) The actual variance of $\hat{\beta}$ is

$$\begin{aligned}
Var\left(\hat{\beta}\right) &= Var\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\beta+\theta_i+\epsilon_{it}\right)\right) \\
&= \mathbf{E}\left\{\left[\frac{1}{NT}\sum_{i}\sum_{t}(\theta_i+\epsilon_{it})\right]^2\right\} \\
&= \frac{1}{N^2T^2}\mathbf{E}\left\{\left[\sum_{i}\sum_{t}(\theta_i+\epsilon_{it})\right]\left[\sum_{j}\sum_{s}\left(\theta_j+\epsilon_{js}\right)\right]\right\} \\
&= \frac{1}{N^2T^2}\sum_{i}\sum_{t}\mathbf{E}\left\{(\theta_i+\epsilon_{it})\left[\sum_{j}\sum_{s}\left(\theta_j+\epsilon_{js}\right)\right]\right\} \\
&= \frac{1}{N^2T^2}\sum_{i}\sum_{t}\mathbf{E}\left[\sum_{j}\sum_{s}(\theta_i+\epsilon_{it})\left(\theta_j+\epsilon_{js}\right)\right] \\
&= \frac{1}{N^2T^2}\sum_{i}\sum_{t}\mathbf{E}\left[\sum_{s=1}^{T}\left(\theta_i^2+2\theta_i\epsilon_{is}+\epsilon_{it}\epsilon_{is}\right)+\sum_{j\neq i}\sum_{s}(\theta_i+\epsilon_{it})\left(\theta_j+\epsilon_{js}\right)\right] \\
&= \frac{1}{N^2T^2}\sum_{i}\sum_{t}\left[T\sigma_\theta^2+\sigma_\epsilon^2+\mathbf{E}\left(\sum_{j\neq i}\sum_{s}(\theta_i+\epsilon_{it})\left(\theta_j+\epsilon_{js}\right)\right)\right] \\
&= \frac{T\sigma_\theta^2+\sigma_\epsilon^2}{NT}+\frac{1}{N^2T^2}\sum_{i}\sum_{t}\sum_{j\neq i}\sum_{s}\mathbf{E}\left[(\theta_i+\epsilon_{it})\left(\theta_j+\epsilon_{js}\right)\right] \\
&= \frac{\sigma_\theta^2}{N}+\frac{\sigma_\epsilon^2}{NT}
\end{aligned}$$

The last equality holds because $\mathbf{E}_{j\neq i}\left[(\theta_i+\epsilon_{it})\left(\theta_j+\epsilon_{js}\right)\right]=0$. It is obvious that $Var\left(\hat{\beta}\right) > \widehat{Var\left(\hat{\beta}\right)}$ whenever $T>1$. This means that if we ignore the panel structure of the data, we will over estimate the variance.

(c) We know that

$$\begin{aligned}
\hat{\beta} &= \frac{1}{NT}\sum_{i}\sum_{t}\left(\beta+u_{it}\right) \\
&= \beta+\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}
\end{aligned}$$

To consider the clustered standard errors, we define

$$v_i \equiv \sum_{t-1}^{T}u_{it}$$

Note that v_i is iid because $Cov\left(v_i, v_j\right)=0$ when $i\neq j$, and $T^2\sigma_\theta^2+T\sigma_\epsilon^2$ all i when $i=$

j.[1] Then we can write the estimator of β as

$$\hat{\beta} = \beta + \frac{1}{NT} \sum_{i=1}^{N} v_i$$

This implies the variance of $\hat{\beta}$ when we cluster the standard error is

$$\begin{aligned} Var\left(\hat{\beta}\right)_{cluster} &= \frac{\sum_i Var\left(v_i\right)}{N^2 T^2} \\ &= \frac{NT^2\sigma_\theta^2 + NT\sigma_\epsilon^2}{N^2 T^2} \\ &= \frac{\sigma_\theta^2}{N} + \frac{\sigma_\epsilon^2}{NT}, \end{aligned}$$

which is the same as the actual variance.

2. (a) In the two-period case, the FE estimator is the same as the FD estimator, which is given by the following differenced equation

$$Y_{i2} - Y_{i1} = \alpha \left(\tau_{g(i)2} - \tau_{g(i)1}\right) + \delta + \varepsilon_{i2} - \varepsilon_{i1}$$

Then the regression result for α is

$$\hat{\alpha} = \frac{\sum_{i=1}^N \left(\tau_{g(i)2} - \tau_{g(i)1} - (\bar{\tau}_2 - \bar{\tau}_1)\right)(Y_{i2} - Y_{i1})}{\sum_{i=1}^N \left[\tau_{g(i)2} - \tau_{g(i)1} - (\bar{\tau}_2 - \bar{\tau}_1)\right]^2}$$

Note that $\tau_{g(i)2} - \tau_{g(i)1} = 1$ if $g(i) = 1$ and zero otherwise. Define $N_t = \sum_{g(i)=1} 1$ and $N_c = \sum_{g(i)=0} 1$, we have

$$\bar{\tau}_2 - \bar{\tau}_1 = \frac{N_t}{N_t + N_c}$$

This means that

$$\begin{aligned} \sum_{i=1}^N \left[\tau_{g(i)2} - \tau_{g(i)1} - (\bar{\tau}_2 - \bar{\tau}_1)\right]^2 &= \sum_{g(i)=1} \left(\frac{N_c}{N_t + N_c}\right)^2 + \sum_{g(i)=0} \left(\frac{N_t}{N_t + N_c}\right)^2 \\ &= N_t \left(\frac{N_c}{N_t + N_c}\right)^2 + N_c \left(\frac{N_t}{N_t + N_c}\right)^2 \end{aligned}$$

[1] You may verify by yourself that $Var(v_i) = T^2\sigma_\theta^2 + T\sigma_\epsilon^2$.

and

$$\sum_{i=1}^{N}\left(\tau_{g(i)2}-\tau_{g(i)1}-\left(\bar{\tau}_2-\bar{\tau}_1\right)\right)\left(Y_{i2}-Y_{i1}\right)=\frac{N_c}{N_t+N_c}\sum_{g(i)=1}\left(Y_{12}-Y_{11}\right)$$
$$-\frac{N_t}{N_t+N_c}\sum_{g(i)=0}\left(Y_{02}-Y_{01}\right)$$
$$=\frac{N_cN_t}{N_t+N_c}\left(\bar{Y}_{12}-\bar{Y}_{11}\right)-\frac{N_cN_t}{N_t+N_c}\left(\bar{Y}_{02}-\bar{Y}_{01}\right)$$

Combining the two results above, we know that

$$\hat{\alpha}=\frac{\frac{N_cN_t}{N_t+N_c}\left(\bar{Y}_{12}-\bar{Y}_{11}\right)-\frac{N_cN_t}{N_t+N_c}\left(\bar{Y}_{02}-\bar{Y}_{01}\right)}{N_t\left(\frac{N_c}{N_t+N_c}\right)^2+N_c\left(\frac{N_t}{N_t+N_c}\right)^2}$$
$$=\left(\bar{Y}_{12}-\bar{Y}_{11}\right)-\left(\bar{Y}_{02}-\bar{Y}_{01}\right)$$

This is just the well-known Diff-In-Diff estimator.
When the number of periods is greater than 2, we have a similar result. To see this, notice that

$$Y_{it}-\bar{Y}_i=\alpha\left(\tau_{g(i)t}-\bar{\tau}_{g(i)}\right)+\delta\left(t-\bar{t}\right)+\varepsilon_{it}-\bar{\varepsilon}_i,\ t=1,2,\cdots,T$$

Define that $\ddot{Y}_{it}\equiv Y_{it}-\bar{Y}_i$,$\ddot{\tau}_{it}\equiv\tau_{g(i)t}-\bar{\tau}_{g(i)}$, and $\ddot{t}=t-\bar{t}$. Then the FE estimator for α is obtained by solving the following minimization problem

$$\min_{\alpha,\delta}L(\alpha,\delta)\equiv\sum_i\sum_t\left(\ddot{Y}_{it}-\alpha\ddot{\tau}_{g(i)t}-\delta\ddot{t}\right)^2$$

The first-order conditions are

$$\frac{\partial L(\alpha,\delta)}{\partial\alpha}=-2\sum_i\sum_t\ddot{\tau}_{it}\left(\ddot{Y}_{it}-\alpha\ddot{\tau}_{g(i)t}-\delta\ddot{t}\right)=0$$
$$\frac{\partial L(\alpha,\delta)}{\partial\alpha}=-2\sum_i\sum_t\ddot{t}\left(\ddot{Y}_{it}-\alpha\ddot{\tau}_{g(i)t}-\delta\ddot{t}\right)=0$$

Solving for the system of equations, we obtain that

$$\hat{\alpha}_{FE}=\frac{\left(\sum_i\sum_t\ddot{\tau}_{it}\ddot{Y}_{it}\right)\left(\sum_i\sum_t\ddot{t}^2\right)-\left(\sum_i\sum_t\ddot{t}\ddot{Y}_{it}\right)\left(\sum_i\sum_t\ddot{t}\ddot{\tau}_{it}\right)}{\left(\sum_i\sum_t\ddot{t}^2\right)\left(\sum_i\sum_t\ddot{\tau}_{it}^2\right)-\left(\sum_i\sum_t\ddot{t}\ddot{\tau}_{it}\right)^2}$$

After some cumbersome algebra, we may show that[2]

$$\hat{\alpha}_{FE} = \frac{N(T+1)\left(\bar{Y}_{12} - \bar{Y}_{11}\right) - 6\sum_i \sum_t t\ddot{Y}_{it}}{N(T+1) - 3N_t T_0 T_1 (T-1)},$$

where $\bar{Y}_{12} = \frac{1}{N_t T_1} \sum_{g(i)=1} \sum_{t \geq t_0} Y_{it}$, and $\bar{Y}_{11} = \frac{1}{N_t T_0} \sum_{g(i)=1} \sum_{t < t_0} Y_{it}$.
(b) The FD estimator is the pooled OLS estimator for the following equation:

$$\Delta Y_{it} = \alpha \Delta \tau_{g(i)t} + \delta + \Delta \varepsilon_{it},\ t = 2, 3, \cdots, T,$$

where $\Delta Y_{it} = Y_{it} - Y_{it-1}, \Delta \tau_{g(i)t} = \tau_{g(i)t} - \tau_{g(i)t-1}$, and $\Delta \varepsilon_{it} \equiv \varepsilon_{it} - \varepsilon_{it-1}$. Then the OLS estimator of α is

$$\hat{\alpha}_{FD} = \frac{\sum_i \sum_{t=2}^T \left(\Delta \tau_{g(i)t} - \overline{\Delta \tau}\right) \Delta Y_{it}}{\sum_i \sum_{t=2}^T \left(\Delta \tau_{g(i)t} - \overline{\Delta \tau}\right)^2}$$

Note that

$$\overline{\Delta \tau} = \frac{\sum_i \left(\tau_{g(i)T} - \tau_{g(i)1}\right)}{N(T-1)} = \frac{N_t}{N(T-1)}$$

$$\Delta \tau_{g(i)t} = \begin{cases} 1 & if\ t \geq t_0\ and\ g(i) = 1 \\ 0 & otherwise \end{cases}$$

Plugging them back to the expression of $\hat{\alpha}_{FD}$, we obtain that

$$\begin{aligned} \hat{\alpha}_{FD} &= \frac{\sum_{g(i)=1} \left[\sum_{t \geq t_0} \left(1 - \frac{N_t}{N(T-1)}\right) + \sum_{t < t_0} \left(-\frac{N_t}{N(T-1)}\right)\right] \Delta Y_{it} + \sum_{g(i)=0} \sum_t \left(-\frac{N_t}{N(T-1)}\right) \Delta Y_{it}}{\sum_{g(i)=1} \left[\sum_{t \geq t_0} \left(1 - \frac{N_t}{N(T-1)}\right)^2 + \sum_{t < t_0} \left(-\frac{N_t}{N(T-1)}\right)^2\right] + \sum_{g(i)=0} \sum_t \left(-\frac{N_t}{N(T-1)}\right)^2} \\ &= \frac{N(T-1)\left(\bar{Y}_{1T} - \bar{Y}_{1t_0-1}\right) - N_t\left(\bar{Y}_{1T} - \bar{Y}_{11}\right) - N_c\left(\bar{Y}_{0T} - \bar{Y}_{01}\right)}{T_1(T-1)N - (2T_1 - 1)N_t} \end{aligned}$$

When $T = 2$, the FE and FD estimators are identical and requires the same exogeneity assumption for the error term. When $T > 2$, unlike the FE estimator which employs all of the information on Y_{it}, calculating the FD estimate only requires data on Y_{iT}, Y_{it_0-1},and Y_{i1}.[3] Moreover, the FE estimator is more efficient under assumption that ε_{it} are serially uncorrelated, while the FD estimator is more efficient when ε_{it} follows a random walk.

[2]We may verify that when $T = 2$, this estimator degenerates into $\left(\bar{Y}_{12} - \bar{Y}_{11}\right) - \left(\bar{Y}_{02} - \bar{Y}_{01}\right)$.
[3]Verify by yourself that $\hat{\alpha}_{FD} = \left(\bar{Y}_{12} - \bar{Y}_{11}\right) - \left(\bar{Y}_{02} - \bar{Y}_{01}\right)$ when $T = 2$ and $T_1 = 1$.