Alastair Hall **ECON61001: Econometric Methods** Semester 1

Study Group Questions # 7

This study group assignment is designed to give you an introduction to the use of *Difference in Difference (DiD)* estimators in policy analysis. In these analytical questions, you demonstrate that the DiD estimator can be obtained via OLS estimation of a particular regression model.

Please enter the details of the group:

Group name	
Student ID 1	
Student ID 2	
Student ID 3	
Student ID 4	
Student ID 5	

Consider the case in which we have two independent samples from a population of individuals taken at times $t = 1$ and $t = 2$. Each sample consists of an outcome for a variable y and a dummy variable T that captures whether or not the individual exhibits some time invariant characteristic. For $t = 1$, the sample of outcomes for (y, T) are denoted $\{y_{1,i}, T_{1,i}\}_{i=1}^{N_1}$; for $t = 2$, the sample of outcomes for (y, T) are denoted $\{y_{2,i}, T_{2,i}\}_{i=1}^{N_2}$. In addition, define $N_{\ell,T} = \sum_{i=1}^{N_\ell} T_{\ell,i}$ and $N_{\ell,C} = N_\ell - N_{\ell,T}$ for $\ell = 1, 2$.

Suppose now that the two samples are pooled to create a single sample on (y, T, D) where D is a dummy variable that indicates whether or not the observation pertains to time $t = 2$. We denote this pooled sample as $\{y_i, T_i, D_i\}_{i=1}^{N}$ where

$$(y_i, T_i) \;=\; \begin{cases} (y_{1,j}, T_{1,j}) & \text{for } \; i = j, \; j = 1, 2, \ldots, N_1, \\ (y_{2,j}, T_{2,j}) & \text{for } \; i = N_1 + j, \; j = 1, 2, \ldots, N_2, \end{cases}$$

and

$$D_i \;=\; \begin{cases} 0, & \text{if } i \leq N_1, \\ 1, & \text{if } N_1 < i \leq N. \end{cases}$$

Consider the pooled regression model

$$y_i \;=\; (1 - D_i)(1 - T_i)\eta_1 + (1 - D_i)T_i\eta_2 + D_i(1 - T_i)\eta_3 + D_iT_i\eta_4 + \text{“error”},$$

and let $\hat{\eta}$ denote the OLS estimator of the regression parameter vector $\eta_0 = (\eta_1, \eta_2, \eta_3, \eta_4)$.

1. Show that $\hat{\eta} = (\bar{y}_{1,C}, \bar{y}_{1,T}, \bar{y}_{2,C}, \bar{y}_{2,T})'$ where, for $\ell = 1, 2$, $\bar{y}_{\ell,C} = N_{\ell,C}^{-1} \sum_{i=1}^{N_\ell} (1 - T_{\ell,i}) y_{\ell,i}$ and $\bar{y}_{\ell,T} = N_{\ell,T}^{-1} \sum_{i=1}^{N_\ell} T_{\ell,i} y_{\ell,i}$.

the pooled regression model : $y_i = (1-D_i)(1-T_i)\eta_1 + (1-D_i)T_i\eta_2 + D_i(1-T_i)\eta_3 + D_iT_i\eta_4 + \text{'error'}$

Consider the model as : $y = X \cdot \eta$, then , X can be written as :

$$X = \begin{bmatrix} (1-D_i)(1-T_i) & (1-D_i)T_i & D_i(1-T_i) & D_iT_i \end{bmatrix} \quad (X \text{ is a } N\times 4 \text{ matrix}).$$

$$\Rightarrow \quad X'X = \begin{bmatrix} (1-D_i)(1-T_i) \\ (1-D_i)T_i \\ D_i(1-T_i) \\ D_iT_i \end{bmatrix} \begin{bmatrix} (1-D_i)(1-T_i) & (1-D_i)T_i & D_i(1-T_i) & D_iT_i \end{bmatrix}$$

$\Rightarrow$ $X'X$ will be a 4×4 matrix. For convenience, write $\sum_{i=1}^{N}$ as Σ.

$$X'X = \begin{bmatrix} \Sigma(1-D_i)^2(1-T_i)^2 & \Sigma(1-D_i)^2(1-T_i)T_i & \Sigma(1-D_i)D_i(1-T_i)^2 & \Sigma(1-D_i)D_i(1-T_i)T_i \\ \Sigma(1-D_i)^2T_i(1-T_i) & \Sigma(1-D_i)^2T_i^2 & \Sigma(1-D_i)D_i(1-T_i)T_i & \Sigma(1-D_i)D_iT_i^2 \\ \Sigma D_i(1-D_i)(1-T_i)^2 & \Sigma D_i(1-D_i)T_i(1-T_i) & \Sigma D_i^2(1-T_i)^2 & \Sigma D_i^2T_i(1-T_i) \\ \Sigma D_i(1-D_i)T_i(1-T_i) & \Sigma D_i(1-D_i)T_i^2 & \Sigma D_i^2(1-T_i)T_i & \Sigma D_i^2T_i^2 \end{bmatrix}$$

Since we have D_i , T_i , $(1-D_i)$ and $(1-T_i)$ are dummy variables taking values of 0 and 1, we have: $\begin{cases} D_i^2 = D_i \\ T_i^2 = T_i \\ (1-D_i)^2 = 1-D_i \\ (1-T_i)^2 = 1-T_i, \end{cases}$ and all other combinations will equal 0 .

From the definition of T_i (as $T_{1,I}$ and $T_{2,I}$) and D_i (as 0 and 1) we have:

$$\Rightarrow \Sigma(1-D_i)^2(1-T_i)^2 = \Sigma(1-D_i)(1-T_i) = \sum_{i=1}^{N_1}(1-D_i)(1-T_{1,i}) + \sum_{i=N_1+1}^{N_2}(1-D_i)(1-T_{2,i})$$

$$= \sum_{T=1}^{N_1}(1-0)(1-T_{1,i}) + \sum_{i=N_1+1}^{N_2}(1-1)(1-T_{2,i})$$

$$= N_{1,C} .$$

Similarly, we have :

$$\Sigma(1-D_i)^2T_i^2 = \Sigma(1-D_i)T_i = \sum_{i=1}^{N_1}(1-D_i)T_{1,i} + \sum_{i=N_1+1}^{N_2}(1-D_i)T_{2,i}$$

$$= \sum_{i=1}^{N_1}(1-0)T_{1,i} + \sum_{i=N_1+1}^{N_2}(1-1)T_{2,i} = N_{1,T}$$

Following the same procedure, we have:

$$\Sigma D_i^2(1-T_i)^2 = N_{2,C}$$

$$\Sigma D_i^2T_i^2 = N_{2,T} .$$

$$\Rightarrow (X'X)^{-1} = \begin{bmatrix} N_{1,C}^{-1} & 0 & 0 & 0 \\ 0 & N_{1,T}^{-1} & 0 & 0 \\ 0 & 0 & N_{2,C}^{-1} & 0 \\ 0 & 0 & 0 & N_{2,T}^{-1} \end{bmatrix}$$

Now consider $X'y$:

$$X'y = \begin{bmatrix} (1-D_i)(1-T_i) \\ (1-D_i)T_i \\ D_i(1-T_i) \\ D_i T_i \end{bmatrix} y = \begin{bmatrix} \Sigma(1-D_i)(1-T_i)y_i \\ \Sigma(1-D_i)T_i y_i \\ \Sigma D_i(1-T_i)y_i \\ \Sigma D_i T_i y_i \end{bmatrix}$$ with Σ is short for $\sum_{i=1}^{N}$

Examining the elements of $X'y$, we have:

$$\Sigma(1-D_i)(1-T_i)y = \sum_{i=1}^{N_1}(1-D_i)(1-T_{1,i})y_{1,i} + \sum_{i=N_1+1}^{N_2}(1-D_i)(1-T_{2,i})y_{2,i}$$

$$= \sum_{i=1}^{N_1}(1-T_{1,i})y_{1,i}$$

Similarly, we have:

$$\Sigma(1-D_i)T_i y_i = \sum_{i=1}^{N_1} T_{1,i} y_{1,i}$$

$$\Sigma D_i(1-T_i)y_i = \sum_{i=N_1+1}^{N_2}(1-T_{2,i})y_{2,i}$$

$$\Sigma D_i T_i y_i = \sum_{i=N_1+1}^{N_2} T_{2,i} y_{2,i}$$

Thus,

$$(X'X)^{-1}X'y = \begin{bmatrix} N_{1,C}^{-1}\sum_{i=1}^{N_1}(1-T_{1,i})y_{1,i} \\ N_{1,T}^{-1}\sum_{i=1}^{N_1} T_{1,i} y_{1,i} \\ N_{2,C}^{-1}\sum_{i=N_1+1}^{N_2}(1-T_{2,i})y_{2,i} \\ N_{2,T}^{-1}\sum_{i=N_1+1}^{N_2} T_{2,i} y_{2,i} \end{bmatrix} = \begin{bmatrix} \bar{y}_{1,C} \\ \bar{y}_{1,T} \\ \bar{y}_{2,C} \\ \bar{y}_{2,T} \end{bmatrix} = \hat{\eta}$$

$$\Rightarrow \hat{\eta} = (\bar{y}_{1,C} \quad \bar{y}_{1,T} \quad \bar{y}_{2,C} \quad \bar{y}_{2,T})' \quad \text{(Q.E.D)}.$$

Now consider the regression model

$$y_i = \beta_1 + T_i\beta_2 + D_i\beta_3 + D_iT_i\beta_4 + \text{“error”}.$$

Define $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ to be the OLS estimator of $(\beta_1, \beta_2, \beta_3, \beta_4)$.

2. Show that $\hat{\beta}_4$ is the difference-in-difference estimator. Provide mathematical formulae for $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ as functions of the data and provide a verbal description of what each estimator captures. *Hint: Use the result in Question 2 from the Tutorial session.*

Consider the regression model $y_i = \beta_1 + T_i\beta_2 + D_i\beta_3 + D_iT_i\beta_4 + \text{'error'}$

Since T_i and D_i are dummy, we can write the value of y_i conditioning on T_i and D_i:

If $T_i = 0,\ D_i = 0 \Rightarrow \quad y_i = \beta_1 + \text{'error'}$

$T_i = 1,\ D_i = 0 \Rightarrow \quad y_i = \beta_1 + \beta_2 + \text{'error'}$

$T_i = 0,\ D_i = 1 \Rightarrow \quad y_i = \beta_1 + \beta_3 + \text{'error'}$

$T_i = 1,\ D_i = 1 \Rightarrow \quad y_i = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \text{'error'}$

Take the expectation of both sides

$$\rightarrow \begin{cases} E[y_i \mid T_i = 0,\ D_i = 0] = \beta_1 \\ E[y_i \mid T_i = 1,\ D_i = 0] = \beta_1 + \beta_2 \\ E[y_i \mid T_i = 0,\ D_i = 1] = \beta_1 + \beta_3 \\ E[y_i \mid T_i = 1,\ D_i = 1] = \beta_1 + \beta_2 + \beta_3 + \beta_4. \end{cases}$$

$$\Rightarrow \beta_1 = E[y_i \mid T_i = 0,\ D_i = 0]$$

$$\beta_2 = E[y_i \mid T_i = 1,\ D_i = 0] - E[y_i \mid T_i = 0,\ D_i = 0]$$

$$\beta_3 = E[y_i \mid T_i = 0,\ D_i = 1] - E[y_i \mid T_i = 0,\ D_i = 0]$$

and $\beta_4 = E[y_i \mid T_i = 1,\ D_i = 1] - \beta_3 - \beta_2 - \beta_1$

$$= E[y_i \mid T_i = 1,\ D_i = 1] - \{E[y_i \mid T_i = 0,\ D_i = 1] - E[y_i \mid T_i = 0,\ D_i = 0]\} - \{E[y_i \mid T_i = 1,\ D_i = 0] - E[y_i \mid T_i = 0,\ D_i = 0]\} - E[y_i \mid T_i = 0,\ D_i = 0].$$

$$= \{E[y_i \mid T_i = 1,\ D_i = 1] - E[y_i \mid T_i = 1,\ D_i = 0]\} - \{E[y_i \mid T_i = 0,\ D_i = 1] - E[y_i \mid T_i = 0,\ D_i = 0]\}$$

which is a difference between two differences by T and D.

$\rightarrow \beta_4$ is a difference-in-difference coefficient that captures the causal effect of some treatment under their common trend controlled by the differences between their time invariant characteristics and time trend.

(to be continued).

Let's reconsider the model from Question 1 which has $\hat{\eta} = (\bar{y}_{1,C}; \bar{y}_{1,T}; \bar{y}_{2,C}; \bar{y}_{2,T})$

$$y_i = (1-D_i)(1-T_i)\eta_1 + (1-D_i)T_i\eta_2 + D_i(1-T_i)\eta_3 + D_iT_i\eta_4 + \text{'error'} \quad (1)$$

the matrix of covariates of (1) is X; $y = X\eta + \text{'error'}$

$$X = [(1-D_i)(1-T_i) \quad (1-D_i)T_i \quad D_i(1-T_i) \quad D_iT_i]$$
$$= [1 - D_i - T_i + D_iT_i \quad T_i - D_iT_i \quad D_i - D_iT_i \quad D_iT_i]$$

From Tutorial question 2, we know that if there is an admissible square matrix S: $\tilde{X} = XS$, then the estimation from the model $y = \tilde{X}\beta + \text{'error'}$ is subject to $\hat{\eta} = S\hat{\beta}$.

In this case, consider our model $y_i = \beta_1 + T_i\beta_2 + D_i\beta_3 + D_iT_i\beta_4 + \text{'error'}$ (2)

$$\tilde{X} = [1 \quad T_i \quad D_i \quad D_iT_i]$$

Then, we can choose $S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$ because:

$$[1 \quad T_i \quad D_i \quad D_iT_i] = [1 - D_i - T_i + D_iT_i \quad T_i - D_iT_i \quad D_i - D_iT_i \quad D_iT_i] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

or $\tilde{X} = XS$. Then, the estimator of β for (2) is:

$$\hat{\beta} = S^{-1}\cdot\hat{\eta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \bar{y}_{1,C} \\ \bar{y}_{1,T} \\ \bar{y}_{2,C} \\ \bar{y}_{2,T} \end{bmatrix} = \begin{bmatrix} \bar{y}_{1,C} \\ \bar{y}_{1,T} - \bar{y}_{1,C} \\ \bar{y}_{2,C} - \bar{y}_{1,C} \\ (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}) \end{bmatrix}$$

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \bar{y}_{1,C} \\ \hat{\beta}_2 = \bar{y}_{1,T} - \bar{y}_{1,C} \\ \hat{\beta}_3 = \bar{y}_{2,C} - \bar{y}_{1,C} \\ \hat{\beta}_4 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C}) \end{cases}$$

So, $\hat{\beta}_4$ is the diff-in-diff estimator as it shows the difference of two differences of mean outcomes between the treated and control group between two time periods.

$\hat{\beta}_1$ is the outcome sample mean of the control group in period 1.

$\hat{\beta}_2$ is the difference between the outcome sample means of treated and control groups in period 1.

$\hat{\beta}_3$ is the difference between the outcome sample means of the control groups between 2 time periods