Self-study 1

Econometrics

1. **FD and FE for 2-periods data**: Consider the following simple panel data model

$$y_{i,t} = x_{i,t}\beta + \alpha_i^* + \varepsilon_{i,t}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T,$$

where β is one-dimensional, and where it is assumed that

$$\alpha_i^* = \bar{x}_i\lambda + \alpha_i, \quad \text{with} \quad \alpha_i \sim NID(0, \sigma_\alpha^2), \quad \varepsilon_{i,t} \sim NID(0, \sigma_\varepsilon^2),$$

mutually independent and independent for all $x_{i,t}$, where $\bar{x}_i = (1/T)\sum_{t=1}^{T} x_{i,t}$.

The parameter β can be estimated by the fixed effects (or within) estimator given by

$$\hat{\beta}_{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{i,t} - \bar{x}_i)(y_{i,t} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{i,t} - \bar{x}_i)^2}.$$

   (a) As an alternative, the correlation between the error term $\alpha_i^* + \varepsilon_{i,t}$ and $x_{i,t}$ can be handled by an instrumental variables approach.

   Give an expression for the IV estimator $\hat{\beta}_{IV}$ using $x_{i,t} - \bar{x}_i$ as an instrument for $x_{i,t}$. Show that $\hat{\beta}_{IV}$ and $\hat{\beta}_{FE}$ are identical.

   (b) Another way to eliminate the individual effects α_i^* from the model is obtained by taking first differences. This results in

$$y_{i,t} - y_{i,t-1} = (x_{it} - x_{i,t-1})\beta + (\varepsilon_{i,t} - \varepsilon_{i,t}), \quad i = 1, \cdots, N, \quad t = 2, \cdots, T.$$

   Denote the OLS estimator based on the equation just above by $\hat{\beta}_{FD}$. Show that $\hat{\beta}_{FD}$ is identical to $\hat{\beta}_{IV}$ and $\hat{\beta}_{FE}$ if $T = 2$. This identity does no longer hold for $T > 2$.

2. **Differences-in-differences**

   In this exercise, you are going to study the differences-in-differences (DiD) estimator, a useful form of panel data models further.

   Reading material: From Moodle, read Chapter 5 of *Mostly Harmless Econometrics* to get a better understanding of DiD estimation. As complementary literature, you can read *Impact evaluation using DiD* (also available in Moodle).

   Background: One of the most famous uses of DiD is by Card and Krueger (1994)* (attached) on the effect of increasing the minimum wage on unemployment. You are going to replicate some of their results.

   *Card was awarded the Nobel prize in Economics this week for this and related articles. Krueger died on 2019 and they do not award the Nobel prize posthumously.

   Exercise: On April 1, 1992, the minimum wage in New Jersey was raised from \$4.25 to \$5.05. In the neighbouring state of Pennsylvania, however, the minimum wage remained constant at \$4.25. Card and Krueger (1994) analyze the impact of the minimum wage increase on employment in the fast–food industry, since this is a sector which employs many low-wage workers.

The authors collected data on the number of employees in 331 fast–food restaurants in New Jersey and 79 in Pennsylvania. The survey was conducted in February 1992 (before the minimum wage was raised) and in November 1992 (after the minimum wage was raised).

The dataset *m_wage.csv* (attached) includes the information necessary to replicate the Card and Krueger analysis. The dataset is stored in a "wide" format, i.e. there is a single row for each unit (restaurant), and different columns for the outcomes and covariates in different years. The dataset includes the following variables (as well as some others which we will not use):

- nj – a dummy variable equal to 1 if the restaurant is located in New Jersey
- emptot – the total number of full-time employed people in the pre-treatment period
- emptot2 – the total number of full-time employed people in the post-treatment period
- wage_st – a variable measuring the average starting wage in the restaurant in the pre-treatment period
- wage_st2 – a variable measuring the average starting wage in the restaurant in the post-treatment period
- pmeal – a variable measuring the average price of a meal in the pre-treatment period
- pmeal2 – a variable measuring the average price of a meal in the post-treatment period
- co_owned – a dummy variable equal to 1 if the restaurant was co-owned
- bk – a dummy variable equal to 1 if the restaurant was a Burger King
- kfc – a dummy variable equal to 1 if the restaurant was a KFC
- wendys – a dummy variable equal to 1 if the restaurant was Wendys

(a) Calculate the DiD estimate for the average wage in NJ and PA. Noting that the wage is not the outcome of interest in this case, what does this analysis suggest about the effectiveness of the minimum-wage policy?

*Note that there are some observations with missing data in this exercise (these are coded as NA in the data). You can calculate the mean of a vector with missing values by setting the *na.rm* argument to be equal to TRUE in the mean function.

Hint: To compute DiD, you need to compute the difference in average wages in the pre-treatment period for NJ and PA. Do the same for the post-treatment period. Then compute the difference between pre- and post-treatment periods.

(b) Calculate the DiD estimator for the outcome of interest (the number of full-time employees).

(c) Calculate the DiD estimator for the price of an average meal. Do restaurants that were subject to a wage increase raise their prices for fast–food?

(d) Convert the dataset from a "wide" format to a "long" format (i.e. where you have two observations for each restaurant, and an indicator for the time period in which the restaurant was observed).

*Note: The easiest way to achieve the data conversion is to notice that you can simply "stack" one *data.frame* (with information from the pre-treatment period) on top of another *data.frame* (with information from the post-treatment period). So, first create two *data.frame*s with the relevant variables. Second, bind these two *data.frame*s together using the *rbind()* function (the *data.frame*s must have the same column names before they are joined). Note that you will have to create the relevant treatment period indicator before binding the *data.frame*s together.

(e) Estimate DiD using linear regression. You should run two models: one which only includes the relevant variables to estimate the DiD, and one which additionally includes restaurant-level covariates which do not vary over time (kfc, wendys,co_owned). Do your estimates of the treatment effect differ?