Self-Study 1: Differences-in-differences (DiD) estimation

Econometrics

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

Reading material: From the repository, read Chapter 5 of *Mostly Harmless Econometrics* as literature for DiD estimation. Additionally, you can read *Impact evaluation using DiD*.

Background: One of the most famous uses of DiD is by Card and Krueger (1994)[1] (in the repo) on the effect of increasing the minimum wage on unemployment. You are going to replicate some of their results.

Experiment: 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) analyzed the impact of the minimum wage increase on employment in the fast-food industry, 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).

Data: The file *m_wage.csv* (in the repo) includes the information necessary to replicate Card and Krueger's 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 others which we will not use):

Variable name	Description
nj	dummy equal to 1 if the restaurant is located in NJ
$emptot$	total number of full-time employed pre-treatment
$emptot2$	total number of full-time employed post-treatment
$wage\_st$	average starting wage in the restaurant, pre-treatment
$wage\_st2$	average starting wage in the restaurant, post-treatment
$pmeal$	average price of a meal in the pre-treatment period
$pmeal2$	average price of a meal in the post-treatment period
co_owned	dummy variable equal to 1 if restaurant was co-owned
bk	dummy variable equal to 1 if restaurant was a Burger King
kfc	dummy variable equal to 1 if restaurant was a KFC
$wendys$	dummy variable equal to 1 if restaurant was Wendys

[1] Card was awarded the Nobel prize in Economics in 2021 for this and related articles. Krueger died on 2019 and they do not award the Nobel prize posthumously.

1. Load the dataset into Julia using the ***CSV*** and ***DataFrames*** packages. You can follow the example here.

2. You are going to calculate the DiD estimate for the average wage in NJ and PA by using the following formula:

$$(\overline{wage}_{NJ,post} - \overline{wage}_{NJ,pre}) - (\overline{wage}_{PA,post} - \overline{wage}_{PA,pre}).$$

   Hence, do the following:

   i. Calculate the average wage in NJ in the pre- and post-treatment periods.
   ii. Obtain the difference between the average wage in NJ in the pre- and post-treatment periods.
   iii. Calculate the average wage in PA in the pre- and post-treatment periods.
   iv. Obtain the difference between the average wage in PA in the pre- and post-treatment periods.
   v. Calculate the DiD estimator for the average wage to obtain the treatment effect of the minimum wage increase in NJ.
   vi. 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 1: You can index a Julia *DataFrame* by using the *:colname* syntax like in the example here. In particular, you can access the values of the variable *nj* by using *df[!,:nj]*, where *df* is the name of the *DataFrame*.

Note 2: There are some observations with missing data in this exercise. You can calculate the mean of a vector with missing values by using the auxiliar *skipmissing()* function inside the *mean()* computation, like in the example here. Recall that you must load the ***Statistics*** package to use the *mean()* function.

3. Following similar steps as in 2., calculate the DiD estimator for the outcome of interest (the number of full-time employees). What does this analysis suggest about the effect of the minimum-wage policy on employment?

4. 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?

5. You are going to estimate DiD using linear regression, so first you must convert the dataset from a "wide" format to a "long" format. That is, construct a data frame where you have two observations for each restaurant, and an indicator for the time period (pre- or post-treatment) in which the restaurant was observed. You can do this by doing the following:

   i. Generate a dummy variable *treatment* which indicates the time period (pre- or post-treatment) in which the restaurant was observed. It should be equal to 1 if the observation is from the post-treatment period, and 0 otherwise. Hence, it should be of length 820, double the size of the original dataset.

ii. Create a new data frame considering only the observations before treatment for the *nj, emptot, wage, pmeal, co_owned, bk, kfc, wendys* variables.

iii. Create a new data frame considering only the observations after treatment (indexed with a 2 at the end of the name) for the same variables as in ii. The dummy variables that do not change pre- and post-treatment should be duplicated.

iv. Change the column names for the data frame in iii. to be the same as in ii.. You can see an example here

v. Create a new data frame by stacking the two data frames just constructed using the *vcat()* function, like here. The resulting data frame should have 820 rows and 8 columns (*nj, emptot, wage, pmeal, co_owned, bk, kfc, wendys*).

vi. Attach the *treatment* variable (add it as an extra column) to the data frame formed in ii. using the *insertcols()* function, like here. The resulting data frame should have 820 rows and 9 columns (*nj, emptot, wage, pmeal, co_owned, bk, kfc, wendys, treatment*).

vii. Remove any row with missing values using the *completecases()* function, like here. The resulting data frame should have 721 rows and 9 columns (*nj, emptot, wage, pmeal, co_owned, bk, kfc, wendys, treatment*).

6. You are going to estimate DiD using linear regression. Do the following:

   i. Generate the matrices of regressors and vector of the dependent variable from the data above.

   ii. Estimate DiD using linear regression in the equation

   $$emptot_{it} = \beta_0 + \beta_1 nj_i + \beta_2 treatment_t + \beta_3 nj_i \times treatment_t + \epsilon_{it}.$$

   What are the DiD estimates? Is the estimated effect similar to the one computed in 3.?

   iii. Obtain the standard errors of the DiD estimates and compute the t-statistic. Is the effect statistically significant?

Note: Notice that we have not corrected for correlation of the error term across time for the same restaurant. A robust standard error would be appropriate in this case, but we will not consider it in this self-study.

7. Following similar steps as in 6., estimate DiD using linear regression in the the extended specification controlling for whether the restaurant was co-owned, a Burger King, a KFC, or a Wendys. You may have to construct a new matrix of regressors. The specification is

   $$\begin{aligned} emptot_{it} = \quad & \beta_0 + \beta_1 nj_i + \beta_2 treatment_t + \beta_3 nj_i \times treatment_t + \beta_4 co\_owned_i \\ & + + \beta_5 bk_i + \beta_6 kfc_i + \beta_7 wendys_i + \epsilon_{it}. \end{aligned}$$

   Do your estimates of the treatment effect differ? Are they statistically significant?