Problem Set 4

Tao Wu

October 24, 2023
Professor Suleyman Taspinar
ECON 387: Advanced Econometrics

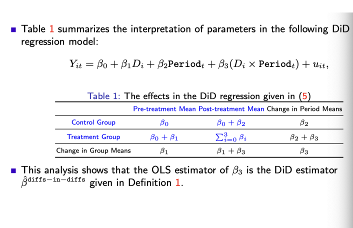
1 Question 1

(20 pts) Define the following terms in your own words.

(a) A quasi-experiment is like a real experiment except that it lacks the elements of true randomization of assignment of participants to either treatment group or control group.

(b) A natural experiment like a quasi-experiment often involves pre-existing groups that are not conceived by the experimenter. This kind of experiment happens in the real world in a more naturalistic setting that makes the results more generalizable.

(c) Because of the lack of full randomization in the quasi-experiment, the treatment group and control group are not completely uncorrelated. In other words, some differences between the 2 groups might remain if we simply use the differences estimator even with the help of control variables.



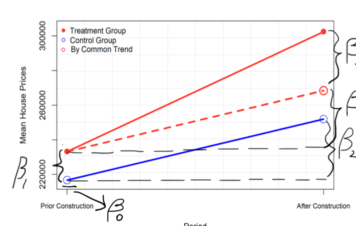
- Table 1 summarizes the interpretation of parameters in the following DiD regression model:

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 \texttt{Period}_t + \beta_3 (D_i \times \texttt{Period}_t) + u_{it},$$

Table 1: The effects in the DiD regression given in (5)

	Pre-treatment Mean	Post-treatment Mean	Change in Period Means
Control Group	β_0	$\beta_0 + \beta_2$	β_2
Treatment Group	$\beta_0 + \beta_1$	$\sum_{i=0}^{3} \beta_i$	$\beta_2 + \beta_3$
Change in Group Means	β_1	$\beta_1 + \beta_3$	β_3

- This analysis shows that the OLS estimator of β_3 is the DiD estimator $\hat{\beta}^{\texttt{diffs-in-diffs}}$ given in Definition 1.



effects of DiD regression in table and graph

(d) When the treatment group and the control group in an experiment are being compared, the pre-existing differences between them remain more or less the same before any treatments are introduced. In other words, the treatment group and the control group are actually the same group before any treatments are introduced. So, without treatments, their trends will rise and fall the same way, again, in the absence of intervention or treatments. This is the common trend assumption.

(e) A regression discontinuity design is a method to find the causal effect of a treatment or intervention which is determined by a continuous variable. The assignment of the treatment group and control group is based on the cutoff point in a continuous variable.

(f) For studying changes in populations over time such as trends and variations, repeated cross-sectional data is the desired form of long-term data. Unlike panel data that track the same individuals or entities over extended periods, repeated cross-sectional data collect observations on different individuals or entities over time.

2 Question 2

By transferring some children into the small classes, the school principals introduce selection bias into the study. This is obviously a "Failure to randomize" problem. Think about it for a moment. The performance of students in small classes is now suddenly correlated with all kinds of unobserved factors such as family background, income, and the personality of the parents because those parents who exert pressure on the principals are perhaps more influential and know how to use politics to give their kids an advantage. Influence might come from wealth or power. Therefore, the internal validity of the study is certainly compromised.

Due to the principals' intervention, the independent variable is now "dirty" meaning its correlation with the unobserved factors in the error term. We have learned in the previous lecture that we can use a technique called two-stage least regression by creating an instrumental variable which should be the solution to the violation of the zero mean condition assumption in this problem. Since we have data on the original random assignment of each student before the principal's intervention, we can create an IV based on the data because the initial randomization of the assignment ensures its uncorrelation with all unobserved characteristics of the outcome variable and the error term. This IV also must correlate with the current "dirty" independent variable class size. The rest is just to estimate a two-stage regression model with this new IV. Regress the "contaminated" variable on the IV, and then regress the outcome variable i.e. students' performance on the first predicted outcome variable. The internal validity of the study can be restored this way!

3 Question 3

(a) When looking at the problem, I realize that we are using observational data to estimate the average causal effect of the minimum wage increase on restaurant employment. However, the setting reeks of quasi-experiment "scent." For starters, a new law is passed to increase minimum wages, which somehow only affects city A but not city B. This seems like randomization to me. I know it is not pure randomization. Public policy has many many considerations and forces that shape it before passing. But still, it seems "as if" randomized. Based on that, I am going to use the Differences-in-difference estimator to find the average causal effect of the minimum wage increase on restaurant employment by comparing the change in both the treatment and control group pre- and posttreatment.

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 Period_t + \beta_3 (D_i \times Period_t) + u_{it}$$

β_0 represents the average outcome of the same restaurants this year (control group) before the treatment.

β_1 represents the difference between the average outcome of the same restaurants next year (treatment group) and the same restaurants this year (control group) in the pre-treatment period.

β_2 represents how much the average outcome of the same restaurants (control group) has changed in the post-treatment period i.e. next year.

β_3 represents how much the average outcome of the treatment group has changed in the period after the treatment, compared to what would happened to the same group had the intervention not occurred.

In plain English, the analysis goes like this. Before the increase in minimum wages, restaurants' employment levels in two time periods would have had the same trend or trajectory had a new law not been passed. The posttreatment effect of the treatment group is an overestimate and we need to account for the pre-existing difference in the same restaurants' employment levels between the two time periods. Once we get rid of that pre-existing difference, what we are left with is what we want - the average causal effect of the minimum wage increase on restaurant employment.

(b) If I design my analysis and I get to sample different, independently selected restaurants this year and next year, I will use the data to estimate the average causal effect of the minimum wage increase on restaurant employment level by using regression discontinuity design. The data I collect this way is the type called repeated cross-sectional data which will enable me to run a quasi-experiment whereby the treatment is "as if" randomly assigned conditional on W, in this case, time. For the sake of clarity, let's say the time that the new law goes into effect is May 20, 2023, a cutoff point before which all the restaurants analyzed are the in control group. After that day, all the restaurants analyzed were the in treatment group. The law takes effect immediately and if the market is assumed to be efficient, it will adapt very quickly, which is the reason we will see a jump or "discontinuity" in the data right before and after the cutoff point. We can then run a regression on the data and get a regression discontinuity estimator β_1.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

Since all the data are collected from different, independently selected restaurants, X_i is uncorrelated with the error term. Regression discontinuity estimator β_1 is more likely to be generalizable to a different sample within the same population.

(c) Using the same restaurants in (a) and using different restaurants in (b) can both yield a precise estimate of the average causal effect. However, I think using the same sample of restaurants across two time periods, if possible, should be preferred because the same entities are likely to be more comparable. Also, we can use a fixed effects model to control for the time-invariant factors that might influence the precision of the estimate. Sampling different restaurants in the RDD approach presents a little problem which is that there should be sufficient numbers of independently selected restaurants near the cutoff point W_0 or the discontinuity for precise estimation. Again, when the sample size is large enough, RDD approach can also produce precise estimates.

4 Question 4

Consider the following model:

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 D_t + \beta_3 (G_i \times D_t) + u_t$$

Show that the OLS estimator of β_3 is the DiD estimator $\hat{\beta}^{diffs-in-diff}$ given in Definition 1 shown in the screenshot below.

- Using this notation, we can define the DiD estimator in the following way.

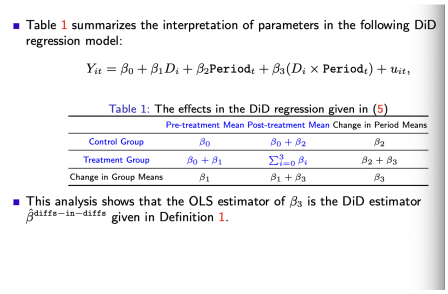
Definition 1

The DiD estimator is the average change in Y for those in the treatment group minus the average change in Y for those in the control group:

$$\begin{aligned} \hat{\beta}^{\texttt{diffs-in-diffs}} &= \left(\bar{Y}^{\texttt{treatment,after}} - \bar{Y}^{\texttt{treatment,before}}\right) \\ &- \left(\bar{Y}^{\texttt{control,after}} - \bar{Y}^{\texttt{control,before}}\right) \\ &= \Delta\bar{Y}^{\texttt{treatment}} - \Delta\bar{Y}^{\texttt{control}} \end{aligned} \tag{1}$$

where $\Delta\bar{Y}^{\texttt{treatment}}$ is the average change in Y in the treatment group and $\Delta\bar{Y}^{\texttt{control}}$ is the average change in Y in the control group.

According to the definition, we need to first find the average change in Y in the treatment group and the average change in Y in the control group. As explained in my previous answers to what a Differences-in-differences estimator is in part c. I will demonstrate the difference for β_0, β_1, β_2, β_3 which is the star of the show $\hat{\beta}^{diffs-in-diff}$.



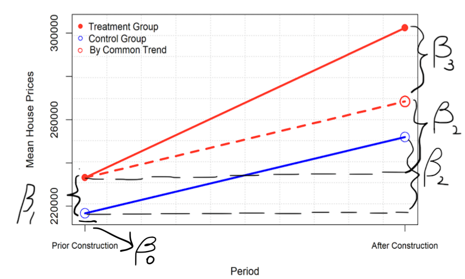
- Table 1 summarizes the interpretation of parameters in the following DiD regression model:

$$Y_{it} = \beta_0 + \beta_1 D_i + \beta_2 \texttt{Period}_t + \beta_3 (D_i \times \texttt{Period}_t) + u_{it},$$

Table 1: The effects in the DiD regression given in (5)

	Pre-treatment Mean	Post-treatment Mean	Change in Period Means
Control Group	β_0	$\beta_0 + \beta_2$	β_2
Treatment Group	$\beta_0 + \beta_1$	$\sum_{i=0}^{3} \beta_i$	$\beta_2 + \beta_3$
Change in Group Means	β_1	$\beta_1 + \beta_3$	β_3

- This analysis shows that the OLS estimator of β_3 is the DiD estimator $\hat{\beta}^{\texttt{diffs-in-diffs}}$ given in Definition 1.



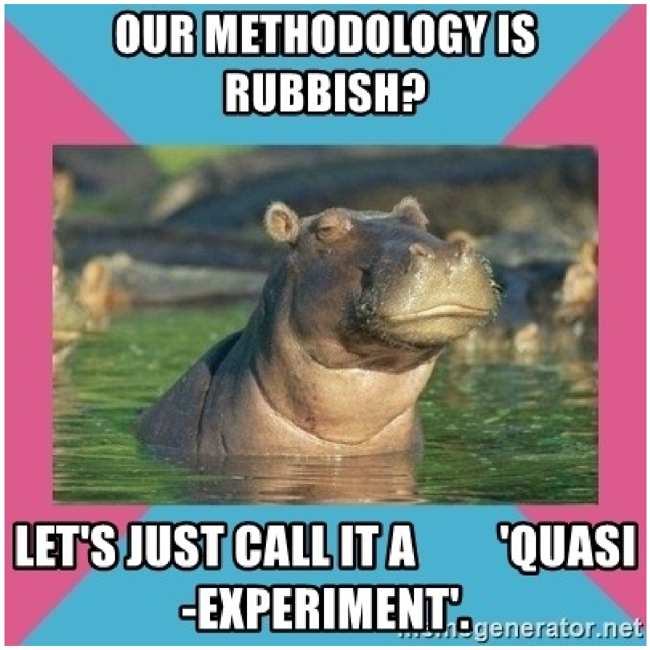
I screenshot the graph in the new housing site example and draw all the βs on it. I hope the picture can show $\hat{\beta}^{diffs-in-diff}$ clearly. If you wonder why we get two β_2, that's because of the common trend assumption. Before treatment is introduced, the treatment group and the control group belong together. They are the same in terms of their growth, trajectory, or built-in trend. So they share the same β_2. Basically,

$$\bar{Y}^{treatment,after} - \bar{Y}^{treatment,before} = \beta_3 + \beta_2$$

$$\bar{Y}^{control,after} - \bar{Y}^{control,before} = \beta_2$$

$$\Delta\bar{Y}^{treatment} - \Delta\bar{Y}^{control} = \beta_3$$

By the way, to the left of the graph is a nifty little table that shows that the order, in which we calculate the differences between the treatment group pre or post-treatment and the control group pre or post-treatment, does not matter. We will be left with the same result.


OUR METHODOLOGY IS RUBBISH?
LET'S JUST CALL IT A 'QUASI -EXPERIMENT'.
memegenerator.net