Regression II
Vanderbilt University
Prof. Sean P. Corcoran

September 24, 2024

Lecture 4 In-Class Exercises: Part 1

Q1 This exercise will replicate the simple difference-in-differences result from chapter 8 of Murnane & Willett. This example comes from Dynarski (2003), who looked at the effects of the Social Security Student Benefit Program (SSBP) on college enrollment. The dataset used here is from the NLSY and consists of high school seniors in 1979-1983. The file is called *dynarski.dta* and is available via Github:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/dynarski.dta, clear
```

1. Is this dataset a panel or a pooled cross-section? What is the unit of observation in the data?

2. Do a cross-tabulation of *yearsr* (the year in which the student is a senior) and *offer* (= 1 if the student was a senior in a year in which Social Security survivor's benefits were available). What is the "group" here (the level at which treatment is assigned)? Which group is "treated" and when are they treated?

3. Estimate a <u>first difference</u> (ITS) model for the effect of SSBP by limiting the analysis to the "ever treated" group and contrasting outcomes in the treated and untreated periods. The outcomes of interest are *coll* (whether the student was enrolled full time in college by age 23) and *hgc23* (the highest grade completed by age 23). You can do this with an OLS regression and be sure to use the sampling weights `weight=[wt88]`. Your results can be compared to Table 8.1 in Murnane & Willett.

4. Now estimate a <u>difference-in-differences</u> model of the effect of SSBP by including the "never treated" group in the regression. Again be sure to use the sampling weights. Your results can be compared to Table 8.2 in Murnane & Willett.

5. For practice, plot the mean outcome *coll* by year for the two groups and note whether the parallel trends assumption appears to hold. Which time period makes the most sense for assessing the parallel trends assumption?

Q2 This exercise will replicate the simple difference-in-differences results from chapter 5 of *Mastering 'Metrics* based on Carpenter & Dobkin (2011). The research question is whether a lower state minimum legal drinking age (MLDA) is associated with a higher mortality rate from motor vehicle accidents among 18-20 years olds. The dataset used here is called *deaths.dta* and is available via Github:

```
use https://github.com/spcorcor18/LPO-8852/raw/main/data/deaths.dta, clear
```

1. Open the *deaths.dta* file to get familiar with it. How is it structured? How many years of data are there? How many states? How many state-year observations? Why are there multiple observations per state and year?

2. Keep only the age 18-20 group and observations pertaining to mortality due to motor vehicle accidents (MVA). Ensure there remain only one observation per state and year.

3. The variable *legal* represents the proportion of persons in a state, year, and age group who can legally drink. For example, it equals 0 if the MLDA in the state-year is 21, and 1 if the MLDA is 18. It can take on fractional values if (say) the legal drinking age is 19 or 20, or if the law changed in the middle of the year. (For example, if the MLDA changed from 18 to 21 in the exact middle of the year, the value of *legal* would be 0.5). Plot a time series that shows the mean percent of adults age 18-20 in a state who are of legal drinking age, by year. Note the changes after 1971 (the 26th Amendment) and 1984 (the National Minimum Drinking Age Act).

4. Estimate a generalized difference-in-differences using <u>two-way fixed effects</u>, where the treatment is defined by *legal* and the outcome is mortality by MVAs. Use only observations before 1984 when the National Minimum Drinking Age Act was passed. Do this once with the traditional standard error calculation, and then again allowing for clustering by state. Your results can be compared to Table 5.2 in *Mastering 'Metrics*.

5. Repeat part 4 but including the *beertax* as an additional covariate. How do the results change if at all?

6. Repeat part 4 but weighting by state population age 18-20. What does this accomplish, and how do the results change if at all?

Table 8.1 "First difference" estimate of the causal impact of an offer of \$6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23 in the United States

(a) *Direct Estimate*

H.S. Senior Cohort	Number of Students	Was Student's Father Deceased	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* (standard error)	Between-Group Difference in Avg Value of *COLL*	$H_0: \mu_{OFFER} = \mu_{NO\ OFFER}$	
						t-statistic	*p*-value
1979–81	137	Yes	Yes (*Treatment Group*)	0.560 (0.053)	**0.208***	2.14	0.017†
1982–83	54	Yes	No (*Control Group*)	0.352 (0.081)			

~p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

(b) *Linear-Probability Model (OLS) Estimate*

Predictor	Estimate	Standard Error	$H_0: \beta = 0$;	
			t-statistic	*p*-value
Intercept	0.352***	0.081	4.32	0.000
OFFER	0.208*	0.094	2.23	0.013†
R^2	0.036			

~p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

Table 8.2 Direct "difference-in-differences" estimate of the impact of an offer of \$6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23, in the United States

H.S. Senior Cohort	Number of Students	Was Student's Father Deceased?	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* (*standard error*)	Between-Group Difference in Avg Value of *COLL*	"Difference in Differences"	
						Estimate (*standard error*)	*p*-value
1979–81	137	Yes	Yes (*Treatment Group*)	0.560 (0.053)	0.208 (*First Diff*)	0.182* (0.099)	0.033†
1982–83	54	Yes	No (*Control Group*)	0.352 (0.081)			
1979–81	2,745	No	No	0.502 (0.012)	0.026 (*Second Diff*)		
1982–83	1,050	No	No	0.476 (0.019)			

~p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.