Regression II
Vanderbilt University
Prof. Sean P. Corcoran

September 29, 2022

Lecture 3 In-Class Exercises

Part 1 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 Social Security survivor's benefits 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. 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 best way to define the "treatment period" here, and which students are "treated"?

2. Estimate a <u>first difference</u> model for the effect of survivor's benefits by limiting the analysis to the "ever treated," comparing outcomes in the treated and non-treated 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.

3. Now estimate a <u>difference-in-differences</u> model of the effect of survivor's benefits 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.

4. 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 fir assessing the parallel trends assumption?

Part 2 This exercise will replicate the simple difference-in-differences results from chapter 5 of *Mastering 'Metrics*. This example comes from 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. 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 two-way fixed effects, 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 (use clustered standard errors) but with other mortality rates as the outcome: suicide, internal causes, and all deaths. You will need to re-load the dataset if you dropped the other mortality measures in part 2. What do you find? One might expect to find an impact of alcohol consumption on MVAs, but less so suicide and internal causes. What would you conclude if you also found significant effects on these other two causes of death?

6. Estimate the same model—for each of the four mortality outcomes—but this time using the age 21-24 group. Use the *legal1820* variable as "treatment" which represents the percent of adults in the state age 18-20 who can legally drink. (*legal* is always 1 for the age 20-24 group). What do you find? One might not expect to find an impact of changes in MLDA laws on those aged 21-24. What would you conclude if you found significant effects on this group?

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

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

9. Repeat part 4 but as a *triple difference* model using the age 21-24 group as a second comparison group.

10. Repeat part 4 but include state-specific time trends. How does this affect the source of variation that is used to identify the effect of the MLDA?

11. Use the following syntax to create a cruder version of the treatment variable (for purposes of doing a simple event study graph).

```
// identify when states switched from 0 to 1 (or 1 to 0)
sort state year
gen legalchange=(legal2~=legal2[_n-1] & state==state[_n-1])
gen temp1=year if legalchange==1
// identify first year of change and later (last) change (if any)--for this
// example I only want states that switched treatment status once
egen changeyear1=min(temp1),by(state)
egen changeyear2=max(temp1),by(state)
preserve
collapse (max) legal2 legalchange (first) changeyear1 changeyear2, by(state_name)
table legal2 legalchange
list, noobs
restore
// drop Illinois and Michigan for this example--they change treatment status twice
drop if state==17 | state==26
drop temp1 changeyear2
// create relative time to treatment (0 in first year of treatment)
gen timetoevent=year-changeyear1
table year timetoevent
```

    Note that Illinois and Michigan are dropped as they change "legal" status more than once. Now that the *timetoevent* variable is created, use the user-created `eventdd` command to estimate and plot an event study.

Part 3 This exercise will perform additional analysis of the MLDA dataset used in Part 2. The focus here will be on the effects of differential treatment timing.

1. Another nifty user-written Stata command called `labmask` allows you to easily attach value labels to a variable based on what is contained in another variable. This will come in handy when labelling graphs later. Try:

```
net install labutil.pkg, from("http://fmwww.bc.edu/RePEc/bocode/l/")
labmask state, values(state_name)
table state
label dir
label list state
```

2. Use the `tabstat` command by `state` to identify states that varied over time in *legal* and *legal2* (created in Part 2). Which states were always treated? Never treated? Were treated approximately half of the time?

3. You can also use the panel command `xtline` to show time variation in *legal* or *legal2* by cross-sectional unit. These time series graphs can be overlaid or shown as a panel of graphs. With 51 states an overlaid graph can look messy. You can limit the graph to display a handful of states, like FIPS codes 1, 2, 4, 9, and 10—Alabama, Alaska, Arizona, Connecticut, and Delaware. The value labels you applied in #3 are useful here. Note you can also use `xtline` to look at trends in the outcome variable (*mrate*) by state.

4. As in Part 2, estimate a generalized difference-in-differences regression model (without controls) where the dependent variable is *mrate*. (As a reminder you have limited the dataset to motor vehicle accident fatalities and age 18-20). This model should include state and year effects, and use the simple 0/1 *legal2* treatment indicator. Use only observations before 1984, when the National Minimum Drinking Age Act was passed. Make note of the estimated treatment effect.

5. Install the user-written command called `bacondecomp`, which performs a "Bacon decomposition" of the generalized DD estimate. (Use `ssc install bacondecomp`). The following command will perform the decomposition and produce a nice scatterplot of 2x2 DD estimates against their weights. Be sure to use the dummy variable treatment *legal2*. How much heterogeneity in 2x2 effects do you see? Are there any outliers that receive disporportionate weight?

   ```
   bacondecomp mrate legal2, ddetail
   ```

   Note the user-written command `ddtiming` does a similar decomposition:

   ```
   net describe ddtiming, from(https://tgoldring.com/code)
   net install ddtiming
   ddtiming mrate legal2, i(state) t(year)
   ```

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.