

FLS 6415: Replication 6 - Difference-in-Differences

May 2019

To be submitted (code + answers) by midnight, Wednesday 22nd May.

First read the paper by Malesky et al (2014) on the course website.

The replication data is in the files *Vietnam0810.csv* (for the main analysis) and *Vietnam0608.csv* (at the end of the exercise).

```
d <- read_csv("Vietnam0810.csv")
```

1. What is treatment and control in this study? What is the treatment assignment mechanism?
2. Run the ‘naive’ cross-sectional OLS regression of the infrastructure index (one of the 6 presented in Table 3 of Malesky et al) on treatment. How do you interpret the results? Provide at least one specific reason why the treatment effect in your regression may be a biased estimate.

```
d %>% lm(index_infra ~ treatment, data = .) %>% stargazer(header = F, keep.stat = c("n"),  
  title = "Q2")
```

Table 1: Q2

<i>Dependent variable:</i>	
	index_infra
treatment	−0.053 (0.049)
Constant	3.205*** (0.018)
Observations	4,129

Note: *p<0.1; **p<0.05; ***p<0.01

3. Run the ‘naive’ before-after OLS regression of the infrastructure index on the time variable (1 for 2010, 0 for 2008) for the treated units only. How do you interpret the results? Provide at least one specific reason why the treatment effect in your regression may be a biased estimate.

```
d %>% filter(treatment == 1) %>% lm(index_infra ~ time, data = .) %>% stargazer(header = F,  
  keep.stat = c("n"), title = "Q3")
```

4. Now perform the main Difference-in-differences analysis for the Infrastructure Index outcome. Don’t cluster your standard errors or include any control variables yet. Interpret the results.

```
d %>% lm(index_infra ~ time + treatment + time * treatment, data = .) %>% stargazer(header = F,  
  keep.stat = c("n"), title = "Q4")
```

5. Repeat Q4 but now add the control variables (lnarea,lnpopden,city, and Region fixed effects) used in Table 3 of Malesky et al. Compare your answers to those in Table 3.

Table 2: Q3

	<i>Dependent variable:</i>
	index_infra
time	0.490*** (0.093)
Constant	2.898*** (0.067)
Observations	528
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Table 3: Q4

	<i>Dependent variable:</i>
	index_infra
time	0.241*** (0.035)
treatment	-0.188*** (0.070)
time:treatment	0.249** (0.098)
Constant	3.086*** (0.025)
Observations	4,129
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

```
d %>% lm(index_infra ~ time + treatment + time * treatment + lnarea + lnpopden +
  city + factor(Region), data = .) %>% stargazer(header = F, keep.stat = c("n"),
  title = "Q5")
```

Table 4: Q5

	<i>Dependent variable:</i>
	index_infra
time	0.224*** (0.034)
treatment	-0.269*** (0.068)
lnarea	0.170*** (0.039)
lnpopden	0.313*** (0.033)
city	0.126* (0.066)
factor(Region)2	0.116* (0.062)
factor(Region)3	0.041 (0.089)
factor(Region)4	0.216*** (0.060)
factor(Region)5	0.248*** (0.068)
factor(Region)7	0.631*** (0.067)
factor(Region)8	-0.006 (0.056)
time:treatment	0.225** (0.094)
Constant	1.039*** (0.256)
Observations	4,126
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

6. Repeat Q5 but now with clustered standard errors at the District level. How does this

alter your results?

```
d %>% lm_robust(index_infra ~ time + treatment + time * treatment + lnarea +
  lnpopden + city + factor(Region), data = ., cluster = District) %>% stargazer(header = F,
  keep.stat = c("n"), title = "Q6")
```

% Error: Unrecognized object type.

7. Replicate all of the columns of Table 3 of Malesky et al.

8. Assess the balance in land area (totalland) of the treated and control units in time $t = 0$. Focus on the substantive difference, not the p-value. Is there any evidence of imbalance? Would this create a risk of bias for our difference-in-differences analysis?

```
d %>% filter(time == 0) %>% t.test(totalland ~ treatment, data = .)

##
## Welch Two Sample t-test
##
## data: totalland by treatment
## t = 0.89132, df = 375.52, p-value = 0.3733
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -185.1575 492.2012
## sample estimates:
## mean in group 0 mean in group 1
## 2190.021 2036.499
```

9. Difference-in-differences methodology cannot protect us against *time-varying* confounders. Provide an example of an omitted (confounding) variable that might create bias in our results.

10. One way of testing for the presence of time-varying confounders is to check that there are *parallel pre-treatment trends* in the outcomes for treated and control units. Using the second dataset, Vietnam0608.csv, and your main difference-in-differences regression, assess if treated units had a different trend to control units before treatment, i.e. between 2006 and 2008, for each of the 6 outcome indices. This should replicate Panel 2 of Table 3 in Malesky et al.

```
d2 <- read_csv("Vietnam0608.csv")

d2 %>% lm_robust(index_infra ~ time + treatment + time * treatment + lnarea +
  lnpopden + city + factor(Region), data = ., cluster = District) %>% texreg(caption = "Q10")
```

11. Create a Difference-in-differences chart showing the average Infrastructure Index outcome by treatment group between 2008 and 2010. Compare this to the same chart between 2006 and 2008. What do these charts suggest about the validity of our difference-in-differences methodology?

```
d %>% mutate(time = time + 1) %>% bind_rows(d2) %>% group_by(time, treatment) %>%
  summarize(mean_index_infra = mean(index_infra, na.rm = T)) %>% ggplot() +
  geom_line(aes(x = time, y = mean_index_infra, group = factor(treatment),
    colour = factor(treatment))) + theme_classic()
```

	Model 1
(Intercept)	1.87*
	[0.95; 2.78]
time	−0.00
	[−0.06; 0.06]
treatment	−0.16
	[−0.46; 0.15]
lnarea	0.11
	[−0.02; 0.24]
lnpopden	0.20*
	[0.07; 0.33]
city	0.10
	[−0.46; 0.65]
factor(Region)2	0.00
	[−0.24; 0.24]
factor(Region)3	−0.02
	[−0.56; 0.51]
factor(Region)4	0.01
	[−0.30; 0.31]
factor(Region)5	0.09
	[−0.15; 0.32]
factor(Region)7	0.44*
	[0.23; 0.66]
factor(Region)8	−0.15
	[−0.41; 0.10]
time:treatment	−0.11
	[−0.27; 0.04]
R ²	0.05
Adj. R ²	0.05
Num. obs.	4220
RMSE	0.98

* 0 outside the confidence interval

Table 5: Q10

