FLS 6415: Replication 6 - Difference-in-Differences

April 2020

To be submitted (code + answers) by midnight, Wednesday 30th April.

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 eexercise).

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

1. What is treatment and control in this study? What is the treatment assignment mechanism?

Treatment is recentralization. Control is maintaining the local councils. The treatment assignment mechanism is an unusual mix of a semi-experimental design by the Ministry of Home Affairs, with some stratification based on their selection criteria and awareness of the need for diversity, but without explicit randomization. While direct self-selection was not allowed, informal political lobbying may well have been possible, along with any conscious or unconscious biases of the Ministry.

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 %>% filter(time == 1) %>% lm(index_infra ~ treatment, data = .) %>% texreg(caption = "Q2",
    caption.above = T, stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3)
```

There appears to be no significant cross-sectional relationship between recentralization and the infrastructure index.

One omitted variable which could bias this estimate would be if larger, wealthier provinces were more likely to be recentralized as they would also be more likely to have better infrastructure.

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.

Table 1: Q2

	Model 1
(Intercept)	3.327***
	(0.026)
treatment	0.061
	(0.071)
R^2	0.000
Adj. R^2	-0.000
Num. obs.	2048
RMSE	1.086

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 2: Q3

	Model 1
(Intercept)	2.898***
	(0.067)
time	0.490***
	(0.093)
R^2	0.050
Adj. R^2	0.048
Num. obs.	528
RMSE	1.070

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 3: Q4

	Model 1
(Intercept)	3.086***
	(0.025)
time	0.241***
	(0.035)
treatment	−0.188***
	(0.070)
time:treatment	0.249**
	(0.098)
R^2	0.018
Adj. R^2	0.018
Num. obs.	4129
RMSE	1.049

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

```
d %>% filter(treatment == 1) %>% lm(index_infra ~ time, data = .) %>% texreg(caption = "Q3",
    caption.above = T, stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3)
```

The quality of infrastructure improved in the second period when treatment took place.

One alternative explanation for this is overall trend bias not associated with treatment. For example, that infrastructure was generally improving over time throughout the country as it got richer and Chinese investment and expertise improved infrastructure.

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 = .) %>% texreg(caption = "Q4",
    caption.above = T, stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3)
```

Recentralization is associated with a 0.25 improvement on the infrastructure index, which is statistically significant at the $p = 0.01$ level. This value is the additional amount that treated provinces improved between 2008 and 2010 *more* than the control provinces.

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 of the paper.

```
d %>% lm(index_infra ~ time + treatment + time * treatment + lnarea + lnpopden +
    city + factor(Region), data = .) %>% texreg(caption = "Q5", caption.above = T,
```

Table 4: Q5

	Model 1
(Intercept)	1.039***
	(0.256)
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)
R^2	0.099
Adj. R^2	0.097
Num. obs.	4126
RMSE	1.006

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

```
    stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3)
```

Controlling for area, population density, city status, and comparing provinces within the same region, recentralization is associated with a 0.225 unit improvement in the quality of infrastructure, statistically significant at thr $p = 0.05$ level. The results are comparable to those in Table 3 but with slightly different standard errors.

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) %>% texreg(caption = "Q6",
    caption.above = T, stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3)
```

Clustering standard errors at the district level increases the uncertainty around our coefficients and reduces their statistical significance, including for the crucial interaction term of the differences-in-differences methodology.

7. Using your regression model from Question 6 applied to all of the outcome variables, try to replicate all of the columns of Panel 1 of Table 3 of Malesky et al. (Some of them might not be the same).

Table 5: Q6

	Model 1
(Intercept)	1.039***
	(0.372)
time	0.224***
	(0.053)
treatment	−0.269**
	(0.115)
lnarea	0.170***
	(0.060)
lnpopden	0.313***
	(0.052)
city	0.126
	(0.103)
factor(Region)2	0.116
	(0.111)
factor(Region)3	0.041
	(0.168)
factor(Region)4	0.216
	(0.176)
factor(Region)5	0.248*
	(0.113)
factor(Region)7	0.631***
	(0.120)
factor(Region)8	−0.006
	(0.133)
time:treatment	0.225
	(0.129)
R^2	0.099
Adj. R^2	0.097
Num. obs.	4126
RMSE	1.006

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Q7

	Infra	Agric	Health	Educ	Comms	Business
(Intercept)	1.039***	−0.611	1.019***	−0.017	1.705***	−1.253**
	(0.372)	(4.319)	(0.157)	(0.316)	(0.354)	(0.562)
time	0.224***	−0.154	−0.014	0.075**	−0.046**	−0.011
	(0.053)	(0.501)	(0.017)	(0.028)	(0.022)	(0.032)
treatment	−0.269**	−0.946	−0.013	0.057	−0.197**	−0.034
	(0.115)	(0.993)	(0.022)	(0.090)	(0.084)	(0.160)
lnarea	0.170***	1.473**	−0.078***	0.231***	0.032	0.368***
	(0.060)	(0.644)	(0.023)	(0.045)	(0.046)	(0.073)
lnpopden	0.313***	1.237**	−0.130***	0.200***	0.089*	0.454***
	(0.052)	(0.569)	(0.020)	(0.041)	(0.047)	(0.072)
city	0.126	1.049	0.030	0.236*	−0.022	0.189
	(0.103)	(3.150)	(0.018)	(0.092)	(0.041)	(0.190)
time:treatment	0.225	2.006	0.123***	0.091	0.152*	0.007
	(0.129)	(1.567)	(0.033)	(0.091)	(0.076)	(0.100)
R^2	0.099	0.076	0.139	0.039	0.131	0.116
Adj. R^2	0.097	0.073	0.137	0.037	0.128	0.113
Num. obs.	4126	4126	4126	4126	4126	4126
RMSE	1.006	10.762	0.386	0.853	0.664	1.021

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

```
vars <- c("index_infra", "index_agric", "index_health", "index_education", "index_comms",
    "index_bus_dev")

regs <- vars %>% map(~lm_robust(as.formula(paste0(.x, " ~ time + treatment + time*treatment + lnarea + l
    data = d, cluster = District))

regs %>% texreg(caption = "Q7", caption.above = T, omit.coef = "Region", stars = c(0.01,
    0.05, 0.1), include.ci = F, digits = 3, custom.model.names = c("Infra",
    "Agric", "Health", "Educ", "Comms", "Business"))
```

The agricultural services index seems the least well replicated.

8. Assess the balance in land area (totalland) of the treated and control units in time $t = 0$ using a simple t-test. (Focus on the substantive difference more than 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
```

While the difference is not significantly different from zero, provinces in the control group on average have a

somewhat larger land area than those in the treatment group. This would be a particular risk of omitted variable bias if we thought that smaller regions were likely to have better outcomes, for example on the infrastructure index, as this could serve as an alternative explanation for our finding.

However, the difference-in-differences methodology is not vulnerable to this risk as it removes all non-time-varying confounders by comparing the treated units before and after treatment, using themselves as controls. So this is not a problem at all.

9. The 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 even though we have used a differences-in-differences approach.

The concern is that provinces in the treatment group were already improving more rapidly on the outcome indicators (eg. the infrastructure index) before the treatment was introduced, so subsequent measured differences are due to unit-specific trends and not the treatment itself. For example, the bureaucrats in charge of the selection may have chosen the most 'up-and-coming' places to test the recentralization as these are the places they are more interested in or spend more time interacting with and have more connections with.

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 from Question 6 (with control variables and clustered standard errors), 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")

vars <- c("index_infra", "index_agric", "index_health", "index_education", "index_comms",
    "index_bus_dev")

regs_pre <- vars %>% map(~lm_robust(as.formula(paste0(.x, " ~ time + treatment + time*treatment + lnarea
    data = d2, cluster = District))

regs_pre %>% texreg(caption = "Q10", caption.above = T, omit.coef = "Region",
    stars = c(0.01, 0.05, 0.1), include.ci = F, digits = 3, custom.model.names = c("Infra",
        "Agric", "Health", "Educ", "Comms", "Business"))
```

Applying the same regression specification to the earlier time period reveals no significant effect, and in fact a substantively negative coefficient on the interaction term. This provides considerable confidence that our earlier results are not due to unit-specific trends.

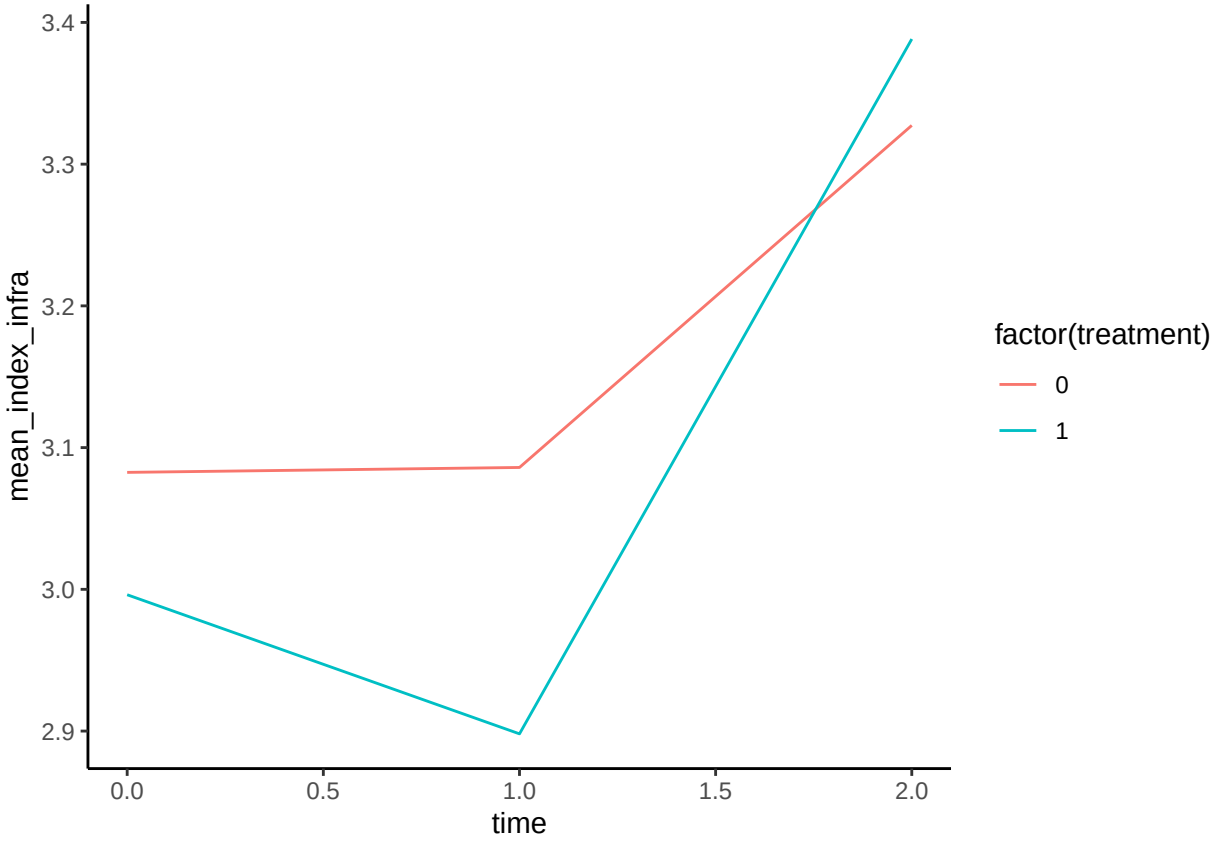
11. Create a Difference-in-differences chart showing the average Infrastructure Index outcome by treatment group between 2008 and 2010. Compare this to the same chart between 2006 and 2008. What do these charts suggest about the validity of our difference-in-differences methodology?

```
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()
```

Table 7: Q10

	Infra	Agric	Health	Educ	Comms	Business
(Intercept)	1.866***	1.380	1.058***	0.022	2.050***	−1.565**
	(0.454)	(4.658)	(0.152)	(0.311)	(0.371)	(0.647)
time	−0.000	1.149**	−0.007	0.022	0.025	0.032
	(0.031)	(0.438)	(0.016)	(0.028)	(0.022)	(0.033)
treatment	−0.158	0.106	0.021	−0.002	−0.146	−0.033
	(0.135)	(1.160)	(0.031)	(0.101)	(0.085)	(0.162)
lnarea	0.108	1.012	−0.089***	0.234***	−0.024	0.408***
	(0.065)	(0.666)	(0.021)	(0.048)	(0.051)	(0.083)
lnpopden	0.198***	0.960	−0.132***	0.202***	0.041	0.501***
	(0.064)	(0.618)	(0.020)	(0.041)	(0.049)	(0.086)
city	0.095	2.755	0.035	0.187	−0.006	0.148
	(0.205)	(1.511)	(0.032)	(0.146)	(0.055)	(0.124)
time:treatment	−0.114	−1.566	−0.033	0.033	−0.051	−0.024
	(0.069)	(0.926)	(0.033)	(0.050)	(0.051)	(0.074)
R^2	0.049	0.067	0.139	0.035	0.129	0.123
Adj. R^2	0.046	0.065	0.137	0.032	0.127	0.120
Num. obs.	4220	4220	4220	4220	4220	4220
RMSE	0.978	9.559	0.393	0.846	0.683	1.043

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$



The charts also highlight that the treated group's infrastructure was declining before the recentralization, so it is even more surprising that it started improving faster than the control group after the recentralization.