Exercise: Difference-in-Differences

This week, we will simulate a difference-in-differences methodology and see how well we recover the assumed treatment effect.

1. Let's create fake data for Brazil, working at the level of the 5570 municipalities. One characteristic (of these municipalities - which will be a potential omitted variable - is the presence of oil reserves. Note that this is a (more or less) *time-invariant* characteristic; if municipality i has oil reserves in year t it also has oil reserves in year $t+1$ or $t+10$. Let's generate a binomial `oil` variable which gives each of our municipalities a 50% chance of possessing oil.

```
N <- 5570
set.seed(54321)
d <- tibble(oil=rbinom(N,1,0.5))
```

2. This is a panel dataset so let's make sure that in our dataset we have (i) an indicator for each `municipality` (a number from 1 to 5570), and (ii) an indicator for each time period. We will work with 3 time periods (years, $t=0$, $t=1$ and $t=2$), so create a column for `year` with entries 0, 1 and 2 so that each municipality has three rows in your dataset, one for each year. (*Hint:* Try using `uncount` in R to duplicate your dataset rows).

```
d <- d %>% mutate(municipality=1:N) %>%
  uncount(3) %>%
  group_by(municipality) %>%
  mutate(year=0:2) %>%
  ungroup()
```

3. Now let's simulate potential outcomes - let's say the outcome is 'voter turnout' - for each municipality. We will assume that in general voter turnout is declining over time (-2% per year), and that the presence of oil reserves also reduces turnout (-3%). The treatment effect we assume will give a 10% boost to voter turnout.

Create variables so that:

$$y_{0,i,year} = N(60,5) - 2 * year - 3 * oil$$

$$y_{1,i,year} = y_{0,i,year} + 10$$

```
d <- d %>% mutate(y_0=rnorm(N,60,5) - 2*year - 3*oil,
                  y_1=y_0+10)
```

4. Treatment D is participation by the municipality in a federal government program and occurs between time periods $t=1$ and $t=2$. We will assume that only municipalities with oil receive treatment. Make an indicator variable where each municipality with oil is coded as being in the treated group and the rest as control. (*Note* that we are not coding municipalities as treated only in $t=2$, we are coding for whether they are a treated 'unit' which applies even in $t=0$, $t=1$).

```
d <- d %>% mutate(D=oil)
```

5. Now calculate the observed outcome based on the potential outcomes, unit treatment status *AND time period.*

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1 & year<2~y_0,
                                  D==1 & year==2~y_1))
```

6. First, what would happen if we ignored differences-in-differences and ran the 'naive' **cross-sectional** observational regression of observed outcomes on treatment? We can only do this in $t = 2$ when the treatment has been activated so we have both control and treated units present. Using only the data from $t = 2$, run the basic regression of observed outcomes on treatment. How does the result compare to our assumed treatment effect? Why do you get this different answer?

```
d %>% filter(year==2) %>%
  lm(y_obs ~D, data=.) %>%
  stargazer(single.row=T, header=F, title="Q6")
```

Table 1: Q6

	Dependent variable:
	y_obs
D	6.842*** (0.134)
Constant	56.061*** (0.095)
Observations	5,570
R^2	0.320
Adjusted R^2	0.320
Residual Std. Error	4.985 (df = 5568)
F Statistic	2,623.208*** (df = 1; 5568)
Note:	*p<0.1; **p<0.05; ***p<0.01

7. Next, let's run the 'naive' comparison of the treated units *before* and *after* treatment is activated. Run the regression of observed outcomes on year, using only the data from $t = 1$ and $t = 2$ for the treated units. How does the result compare to our simulation assumptions? Why do you get this different answer?

```
d %>% filter(year>0 & D==1) %>%
  lm(y_obs ~ year, data=.) %>%
  stargazer(single.row=T, header=F, title="Q7")
```

Table 2: Q7

	Dependent variable:
	y_obs
year	7.901*** (0.135)
Constant	47.101*** (0.213)
Observations	5,580
R^2	0.382
Adjusted R^2	0.382
Residual Std. Error	5.027 (df = 5578)
F Statistic	3,446.277*** (df = 1; 5578)
Note:	*p<0.1; **p<0.05; ***p<0.01

8. Now, using the data for $t = 1$ and $t = 2$, let's run a basic difference-in-differences regression of the

observed outcomes on treatment, year, and the interaction of treatment and year. How do you interpret the results?

```
d %>% filter(year>0) %>%
  lm(y_obs ~ D + year + D*year, data=.) %>%
  stargazer(single.row=T, header=F, title="Q8")
```

Table 3: Q8

	Dependent variable:
	y_obs
D	−12.762*** (0.299)
year	−1.901*** (0.134)
D:year	9.802*** (0.189)
Constant	59.863*** (0.211)
Observations	11,140
R^2	0.270
Adjusted R^2	0.270
Residual Std. Error	4.985 (df = 11136)
F Statistic	1,376.058*** (df = 3; 11136)
Note:	*p<0.1; **p<0.05; ***p<0.01

9. Our standard errors are wrong here. Cluster your errors by the cross-sectional unit (municipality). (In this case the difference is very small).

```
d %>% filter(year>0) %>%
  lm_robust(y_obs ~ D + year + D*year, data=., cluster=municipality) %>%
  texreg(caption="Q9", include.ci = F)
```

10. We can also do a simpler differences-in-differences-in-means estimate without a regression. Create a 2*2 table for the four groups as shown in the table below. Fill in the table with the average observed outcomes. Then calculate the differences in average outcomes for each cell in the rows and/or the columns, and finally the difference in the differences. Interpret your result.

	Model 1
(Intercept)	59.86***
	(0.21)
D	−12.76***
	(0.30)
year	−1.90***
	(0.13)
D:year	9.80***
	(0.19)
R^2	0.27
Adj. R^2	0.27
Num. obs.	11140
RMSE	4.99

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Q9

Treatment:	D=0	D=1	Difference	Diff-in-Diff
t=1				
t=2				

```
d_in_d_means <- d %>% filter(year>0) %>%
  group_by(D, year) %>%
  summarize(mean_y=mean(y_obs,na.rm=T))

d_in_d_means %>%
  spread(key="year", value="mean_y") %>%
  ungroup() %>%
  mutate(Diff=`2`-`1`,
         Diff_Diff=Diff-lag(Diff)) %>%
  kable(caption="Q10")
```

Table 6: Q10

D	1	2	Diff	Diff_Diff
0	57.96194	56.06118	-1.900757	NA
1	55.00228	62.90340	7.901112	9.801869

11. One assumption of Difference-in-Differences is that there are parallel trends before treatment occurs. Let's use the $t = 0$ data to test whether the treated and countrol groups display parallel trends in the outcome variable between time $t = 0$ and $t = 1$. One way to do this is to run exactly the same difference-in-differences regression but using time $t = 0$ and $t = 1$, excluding $t = 2$. Interpret your results.

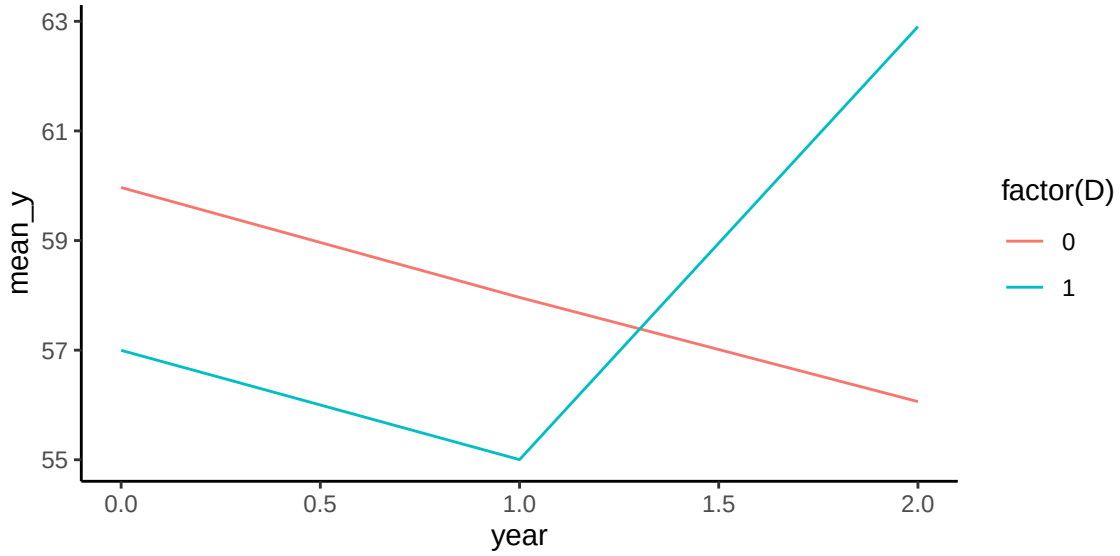
```
d %>% filter(year<2) %>%
  lm(y_obs ~ D + year + D*year, data=.) %>%
  stargazer(single.row=T, header=F, title="Q11")
```

Table 7: Q11

	Dependent variable:
	y_obs
D	-2.971^{***} (0.134)
year	-2.005^{***} (0.134)
D:year	0.011 (0.189)
Constant	59.967^{***} (0.095)
Observations	11,140
R^2	0.114
Adjusted R^2	0.114
Residual Std. Error	4.986 (df = 11136)
F Statistic	477.727^{***} (df = 3; 11136)
Note:	*p<0.1; **p<0.05; ***p<0.01

12. Plot a classic difference-in-differences line graph of the average observed outcome, where the x-axis contains the three time periods, the y-axis the average outcome, and there is one line for the treatment group and one for the control group. Does your graph show there are parallel pre-treatment trends or not?

```
d %>% group_by(D, year) %>%
  summarize(mean_y=mean(y_obs,na.rm=T)) %>%
  ggplot() +
  geom_line(aes(x=year, y=mean_y, group=factor(D), colour=factor(D))) +
  theme_classic()
```



13. Finally, let's try to see what estimate we recover when there are **non-parallel trends**, which are produced by **time-varying confounders**. Recreate your dataset but with the following structure of potential outcomes, which only differs in that the falling turnout trend is only present in oil municipalities. Remember to calculate observed outcomes again, and then run the differences-in-differences regression and interpret the results.

$$y_{0,i,year} = N(60, 5) - 2 * year * oil - 3 * oil$$

$$y_{1,i,year} = y_{0,i,year} + 5$$

```
d2 <- d %>% mutate(y_0=rnorm(N,60,5) - 2*year*oil - 3*oil,
                   y_1=y_0+5)  %>%
  mutate(y_obs=case_when(D==0~y_0,
                         D==1 & year<2~y_0,
                         D==1 & year==2~y_1))

d2 %>% filter(year>0) %>%
  lm(y_obs ~ D + year + D*year, data=.) %>%
  stargazer(single.row=T, header=F, title="Q13")
```

14. Create the difference-in-differences line graph for this new dataset with time-varying confounders. (The same as in Q12). Does your graph show there are parallel pre-treatment trends or not?

```
d2 %>% group_by(D, year) %>%
  summarize(mean_y=mean(y_obs,na.rm=T)) %>%
  ggplot() +
  geom_line(aes(x=year, y=mean_y, group=factor(D), colour=factor(D))) +
  theme_classic()
```

Table 8: Q13

	Dependent variable:
	y_obs
D	−8.551*** (0.299)
year	−0.146 (0.134)
D:year	3.292*** (0.189)
Constant	60.264*** (0.211)
Observations	11,140
R^2	0.154
Adjusted R^2	0.153
Residual Std. Error	4.982 (df = 11136)
F Statistic	674.030*** (df = 3; 11136)
Note:	*p<0.1; **p<0.05; ***p<0.01