Assignment 5

Group name: Foodies with hoodies

Contents

```r
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,stargazer, fixest,
 ↪  modelsummary, knitr)
```

1 Question 1:

1.1 (i)

Suppose the differences in outcomes between the treatment and the control group is:

$$Y_{g1} - Y_{g0} = (\alpha_1 - \alpha_0) + \delta D_g + (U_{g1} - U_{g0}), \tag{1}$$

in which: δ is the estimated treatment effect.

The parallel trends assumption state that without the intervention of the treatment $(\delta = 0)$, the difference of between the control and treatment group ($\alpha_1 - \alpha_0$) remain constant over time. Since in this example, they only look at the pre-treatment period, the parallel trends assumption could be violated due to the fact that after the treatment period, the differences in outcomes of control and treatment groups are not constant over time anymore unrelated to the treatment.

When parallel trend is violated, it means that $\alpha_1 - \alpha_0$ changes over time and this means it is no longer possible to estimate Equation 1 using OLS. If we continue estimating it using OLS, we will have a biased and inconsistent estimator.

One example for violation of this assumption could be that the there exists autocorrelation in the treatment group after getting treated. Specifically, the outcome of the next time lag is influenced by the outcome of the previous lag. Prior to the treatment, both the control and the treatment group have the same time trend because both experience no treatment. However, after receiving the treatment but not due to this, the treated group has a steeper slope in their outcomes due to autocorrelation and is no longer parallel to the control group. Suppose for example that one wishes to investigate the treatment effects on savings among the poorest income bracket, where treatment is giving them a fixed amount of monthly stipend. However, the pre-treatment period was during a crisis, where the poorest income bracket savings were stagnating. After the treatment, there was a booming economic period, where savings were autocorrelated. Then one group turns out to experience a divergent trend in savings not due to the stipend but simply because the economic situation changed the dependent variable's properties.

1.2 (ii)

The difference-in-difference estimator is an OLS estimator Equation 1, which can be written in the form below:

$$Y_{g1} - Y_{g0} = \beta_0 + \delta D_g + U_g. \tag{2}$$

Now suppose that $Corr(D_g, \beta_0) \neq 0$. The main problem with applying the OLS estimator in this case is that the estimator will be biased and inconsistent due to endogeneity issues. This is because the estimator treats the time trend β_0 as a constant, however, in this case, the parallel trends assumption is not satisfied and thus β_0 is not constant. This makes the error term change over time and is correlated with the treatment variable D_g. This implies that the treatment was probably not assigned time-group conditional randomly and this causes the endogeneity of the treatment to the error term, which in turn causes the previously mentioned biasedness and inconsistency.

2 Question 2

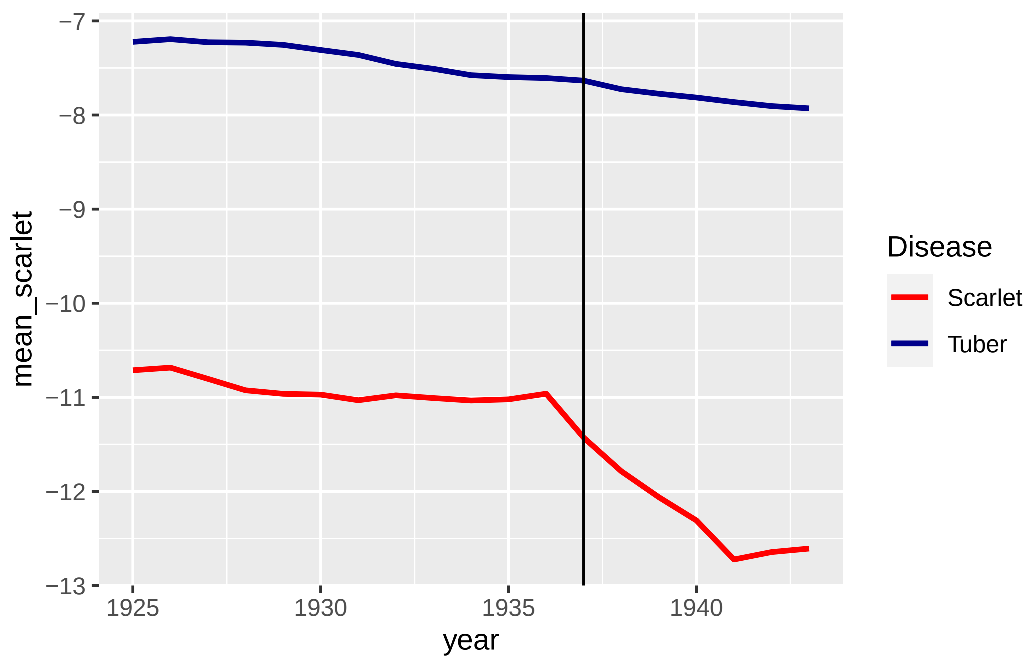
```r
dfData = read.csv("assignment5.csv")
attach(dfData)
```

2.1 (i)

```r
# Get subgroups and mean per year
df_scarlet = dfData[dfData$treated==1,]
mean_scarlet = df_scarlet %>% group_by(year) %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop')
df_tuber = dfData[dfData$treated==0,]
mean_tuber = df_tuber %>% group_by(year) %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop')

# reformat into dataframe
df_grouped =
 ↪  data.frame(mean_tuber$year,mean_scarlet$mean_rate,mean_tuber$mean_rate)
names(df_grouped) = c("year","mean_scarlet","mean_tuber")
```

```r
# Plot
ggplot() +
  geom_line(data=df_grouped,aes(y=mean_scarlet,x=
   ↪  year,colour="Scarlet"),size=1 )+
  geom_line(data=df_grouped,aes(y=mean_tuber,x=
   ↪  year,colour="Tuber"),size=1) +
  scale_color_manual(name = "Disease", values = c("Tuber" = "darkblue",
   ↪  "Scarlet" = "red")) +
  geom_vline(xintercept = 1937) #add a vertical line indicating
   ↪  treatment year
```



From our graph, we can observe that while both diseases had a relatively slight downwards trend (so at first glance, we may assume parallel trends), the treatment appearance seems to affect the scarlet fever with a strong drop-off in means. This might imply that the drop in the scarlet fever against the tuberculosis is due to the treatment (the appearance of the drug is the treatment for fever but not a treatment for tuberculosis). However, we have to make a lot of assumptions for this to indeed hold true.

2.2 (ii)

```r
# get mean effects
mean_treated_1936 = as.numeric(df_scarlet[df_scarlet$year==1936,] %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop'))
mean_treated_1937 = as.numeric(df_scarlet[df_scarlet$year==1937,] %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop'))

mean_control_1936 = as.numeric(df_tuber[df_tuber$year==1936,] %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop'))
mean_control_1937 = as.numeric(df_tuber[df_tuber$year==1937,] %>%
  summarise(mean_rate=mean(lnm_rate),
            .groups = 'drop'))

Before_Treatment_1936 <- c(mean_treated_1936, mean_control_1936,
 ↪  (mean_treated_1936-mean_control_1936))
After_Treatment_1937 <- c(mean_treated_1937, mean_control_1937,
 ↪  (mean_treated_1937-mean_control_1937))
Difference <- c((mean_treated_1936 - mean_treated_1937),
 ↪  (mean_control_1936 - mean_control_1937), ((mean_treated_1936 -
 ↪  mean_treated_1937) - (mean_control_1936 - mean_control_1937)))

dfTable <- data.frame(Before_Treatment_1936, After_Treatment_1937,
 ↪  Difference)
rownames(dfTable) <- c("Treatment", "Control", "Difference")

kable(dfTable, caption="Treatment and time differences of treatment and
 ↪  control groups", digits=3, label = "tab_did")
```

Table 1: Treatment and time differences of treatment and control groups

	Before_Treatment_1936	After_Treatment_1937	Difference
Treatment	-10.962	-11.429	0.467
Control	-7.607	-7.635	0.028
Difference	-3.355	-3.794	0.439

From Table **??**, we can see that the difference-in-differences estimator is 0.439. This DiD estimator is obtained by the differences in means, for treatment group pre-treatment, treatment group post-treatment, control group pre-treatment and control group post-treatment.

2.3 (iii)

```
# Create indicator variable
dfData$indicator <- ifelse(dfData$year >=1937, 1, 0)

# Get subdata for the year 1936 and 1937
dfSub = dfData[dfData$year==1936 | dfData$year==1937,]

# DiD model
DiD1 = feols(lnm_rate ~ indicator*treated|year + treated, data = dfSub,
↪   se="standard")
```

2.4 (iv)

```
DiD2= feols(lnm_rate ~ indicator*treated| year + treated, data = dfData,
↪   se="standard")
msummary(list(DiD1,DiD2), stars = c('*' = .1, '**' = .05, '***' = .01))
```

(For (iii) and (iv), I took out the group-specific and time-specific effect. However, if you want to interpret those you can take out the "|year + treated" part in the model specification)

2.5 (v)

For this question, we can take out the two-way fixed effect of group and year and create an interaction variable of year and treated group, making the year 1936 the reference year and thus normalize its correficients to 0.

```
es <- feols(lnm_rate ~ i(year, treated, ref = 1936)|year+treated, data =
↪   dfData)
summary(es)
```

```
OLS estimation, Dep. Var.: lnm_rate
Observations: 1,721
```

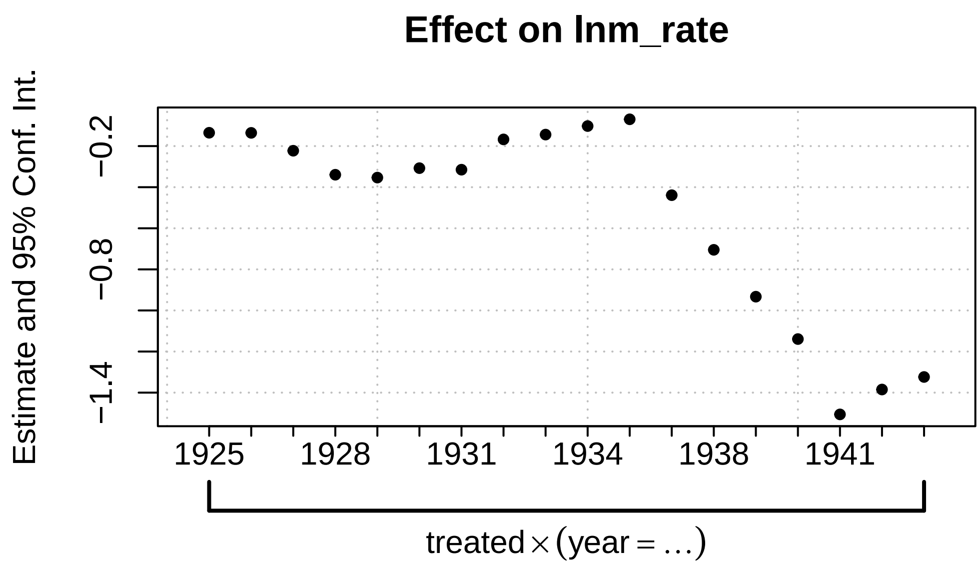
	(1)	(2)
indicator × treated	-0.439^{**}	-0.867^{***}
	(0.219)	(0.060)
Num.Obs.	192	1721
R2	0.851	0.916
R2 Adj.	0.849	0.915
R2 Within	0.021	0.108
R2 Within Adj.	0.016	0.107
AIC	442.6	3200.1
BIC	455.6	3314.5
RMSE	0.75	0.61
Std.Errors	IID	IID
FE: year	X	X
FE: treated	X	X

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

```
Fixed-effects: year: 19,  treated: 2
Standard-errors: Clustered (year)
                    Estimate Std. Error       t value  Pr(>|t|)
year::1925:treated -0.135176   5.59e-13 -2.419209e+11 < 2.2e-16 ***
year::1926:treated -0.135567   5.59e-13 -2.427222e+11 < 2.2e-16 ***
year::1927:treated -0.222575   5.60e-13 -3.973709e+11 < 2.2e-16 ***
year::1928:treated -0.339383   5.62e-13 -6.034650e+11 < 2.2e-16 ***
year::1929:treated -0.353351   5.59e-13 -6.319517e+11 < 2.2e-16 ***
year::1930:treated -0.307318   5.60e-13 -5.484003e+11 < 2.2e-16 ***
year::1931:treated -0.314916   5.60e-13 -5.626575e+11 < 2.2e-16 ***
year::1932:treated -0.167430   5.60e-13 -2.987859e+11 < 2.2e-16 ***
year::1933:treated -0.144043   5.62e-13 -2.565220e+11 < 2.2e-16 ***
year::1934:treated -0.102580   5.60e-13 -1.830538e+11 < 2.2e-16 ***
year::1935:treated -0.069522   5.62e-13 -1.236576e+11 < 2.2e-16 ***
year::1937:treated -0.439008   5.58e-13 -7.867799e+11 < 2.2e-16 ***
year::1938:treated -0.704925   5.54e-13 -1.271874e+12 < 2.2e-16 ***
year::1939:treated -0.932996   5.51e-13 -1.692227e+12 < 2.2e-16 ***
year::1940:treated -1.139170   5.49e-13 -2.076428e+12 < 2.2e-16 ***
year::1941:treated -1.505930   5.79e-13 -2.603004e+12 < 2.2e-16 ***
year::1942:treated -1.384778   5.44e-13 -2.547761e+12 < 2.2e-16 ***
year::1943:treated -1.323763   5.87e-13 -2.256581e+12 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.593665     Adj. R2: 0.917231
```

```
Within R2: 0.142828
```

```
coefplot(es)
```



2.6 (vi)

We need to cluster standard errors at he fixed effect levels: namely year and treated

2.7 (vii)

For this, we can use the Wald test to check to check on the pre-trends from the event study model.

```
# https://lrberge.github.io/fixest/reference/wald.html
wald(es,
↪   keep=c("year::1925:treated","year::1926:treated","year::1927:treated","year::1928:trea
```

```
Wald test, H0: joint nullity of year::1925:treated, year::1926:treated, year::1927:treated, y
 stat = 8.522e+22, p-value < 2.2e-16, on 11 and 1,683 DoF, VCOV: Clustered (year).
```

We reject the null, which mean the parallel condition is not satisfied we can use the placebo test, in which we pick fake treatment periods before the actual treatment period and see if there is a significant effect.

```
# Create fake indicator variables
dfData$D_fake1 <- ifelse(dfData$year >=1928, 1, 0)
dfData$D_fake2 <- ifelse(dfData$year >=1930, 1, 0)
dfData$D_fake3 <- ifelse(dfData$year >=1932, 1, 0)
dfData$D_fake4 <- ifelse(dfData$year >=1934, 1, 0)

# Test fake models
DiD1_fake = feols(lnm_rate ~ D_fake1*treated|year + treated, data =
 ↪  dfData, cluster = "treated^year")
DiD2_fake = feols(lnm_rate ~ D_fake2*treated|year + treated, data =
 ↪  dfData, cluster = "treated^year")
DiD3_fake = feols(lnm_rate ~ D_fake3*treated|year + treated, data =
 ↪  dfData, cluster = "treated^year")
DiD4_fake = feols(lnm_rate ~ D_fake4*treated|year + treated, data =
 ↪  dfData, cluster = "treated^year")
msummary(list(DiD1_fake,DiD2_fake,DiD3_fake,DiD4_fake), stars = c('*' =
 ↪  .1, '**' = .05, '***' = .01))
```

Using the placebo test, we also see that the parallel trend pre-treatment period is not satisfied

	(1)	(2)	(3)	(4)
D_fake1 × treated	-0.407***			
	(0.092)			
D_fake2 × treated		-0.358***		
		(0.105)		
D_fake3 × treated			-0.386***	
			(0.116)	
D_fake4 × treated				-0.513***
				(0.127)
Num.Obs.	1721	1721	1721	1721
R2	0.907	0.907	0.907	0.909
R2 Adj.	0.905	0.906	0.906	0.908
R2 Within	0.011	0.014	0.020	0.040
R2 Within Adj.	0.011	0.013	0.020	0.039
AIC	3376.8	3372.3	3361.1	3326.8
BIC	3491.3	3486.7	3475.6	3441.2
RMSE	0.64	0.64	0.63	0.63
Std.Errors	by: treated^year	by: treated^year	by: treated^year	by: treated^year
FE: year	X	X	X	X
FE: treated	X	X	X	X

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$