Assignment 5

Bas Machielsen & Walter Verwer

February 3, 2021

Problem 1

We assume the following regression model: $Y_{t,g} = \alpha_{t,g} + \delta \cdot D_{t,g} + \eta_g + u_{t,g}$ for $g \in \{T, C\}$ and $t \in \{0, 1\}$. Then, the difference-in-difference estimator equals $\hat{\delta} = \delta \cdot D_1 + \alpha_{1,T} - \alpha_{0,T} + U_{1,T} - U_{0,T} - [\alpha_{1,C} - \alpha_{0,C} + U_{1,C} - U_{0,C}]$. The ATT = expected value of the DiD estimator is then:

$$\mathbb{E}[\hat{\delta}|D = 1] = \delta + [\alpha_{1,T} - \alpha_{0,T}] - [\alpha_{1,C} - \alpha_{0,C}]$$

Hence, the expected value of $\hat{delta}$ depends on the assumption that the sum of the terms containing the α's equal zero, in other words, if there is a common time trend between treatment and control groups.

If we assume that the program is known beforehand (by the students), and grades are a function of effort and ability $\in$ {High, Low}, and high-ability students are all in the treatment group, then a fraction of the treatment group will also consist of high-effort and low-ability students, whereas the control group will consist of low-ability students only. After being provided with the incentive of housing, they will readjust their effort in the 2nd year, and hence, obtain lower grades. This causes a violation of the common trend, because the low-ability individuals who are in the treatment group will revert back to their effort level that is unincentivized by housing.

If we assume the program is not known beforehand, the students have no differing incentives, irrespective of their ability and effort. Hence, the common time trend assumption is justified and the estimator is unbiased.

Problem 2

(i) Regress the number of out-of-wedlock births on the sex ratio, using only the observations from the pre-war period. Discuss your result. How can a difference-in-differences approach using the military mortality rate during WWI improve on this estimation strategy?

```
model1 <- data %>%
    filter(post == 0) %>%
    lm(formula = "illeg ~ sr")

cov1 <- vcovHC(model1, type = "HC1")
robust_se <- sqrt(diag(cov1))

stargazer(model1,
          dep.var.labels = "Illegal Births",
          omit.stat = "ser",
          header = F,
          se = list(robust_se))
```

Table 1:

	Dependent variable:
	Illegal Births
sr	−0.089
	(5.050)
Constant	6.772
	(5.786)
Observations	87
R^2	0.00001
Adjusted R^2	−0.012
F Statistic	0.0005 (df = 1; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimates focus on the cross-section and relies on between-department variation to find the correlation between the sex-rate and illegal births. This gives a biased estimate, because the (unaccounted for) department effects might be correlated with the sex ratio. The difference in difference estimator can improve on this by considering the within-departement estimate, thereby eliminating time- and department-specific effects from the estimate for the influence of the sex ratio.

(ii) Generate a dummy variable that indicates whether the military mortality in a region is above the median military mortality or not. Make a table with the mean percentage of out-of-wedlock births for the high and low mortality regions, both before and after the war. Use the numbers from the table to calculate the difference-in-differences estimator.

```
data <- data %>%
    mutate(htm_mortality = if_else(mortality > median(mortality, na.rm = T),
                                   1,
                                   0)
           )

table <-  data %>%
    filter(!is.na(htm_mortality)) %>%
    group_by(post, htm_mortality) %>%
    summarize(mean_illeg = mean(illeg, na.rm = T))

kable(table)
```

post	htm_mortality	mean_illeg
0	0	7.960866
0	1	5.086213
1	0	8.448658
1	1	6.154916

```
did <- (6.154916 - 5.086213) - (8.448658 - 7.960866)
```

The difference in difference estimator is equal to 0.580911, indicating an increase of illegal births by 0.580911 percentage points, consistent with the hypothesis.

(iii) Estimate the following model, which estimates the difference-in-differences estimator in a regression equation. What is the interpretation of the coefficients β_1 and β_2? What do you conclude about the effect of male scarcity on the number of out-of-wedlock births?

Table 3:

	Dependent variable:
	Illegal Births
post	9.376*
post:mortality	−0.515
Constant	6.672
	(5.786)
Observations	174
R^2	0.095
Adjusted R^2	0.085
F Statistic	8.990*** (df = 2; 171)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient β_1 represents the average illegal births rate post-war, and the constant term represents the average illegal births pre-war. The coefficient β_2 represents the effect on illegal births of a 1-percentage point increase in mortality. The coefficient is negative, but not significant. The point estimate is, however, in accordance with the hypothesis, in that an increased mortality leads to an increase in out-of-wedlock births.

The coefficient, however, might be biased because the model does not take into account systematic differences in the out-of-wedlock birth rate between various departments.

(iv) Run the same regression but now include dummies for all the departments. Discuss your results. Do you prefer this estimation over the estimation of question iii)? Why?

```
library(plm)
model_q4 <- plm(data = data,
                formula = "illeg ~ post + post:mortality",
                model='within',
                index = 'depc')

stargazer(model_q4,
          header = F)
```

We prefer this estimation over the estimation of question iii, because in question iii we do not take the department specifc fixed effects into account. These are likely to bias our estimate of the treatment effect if excluded. This is because it is likely that there are department specific levels of out of wedlock births, thus not including department dummies, means that in the regression of question iii there could be a correlation between the error term (includes the department specific effects) and the mortality rate. This correlation would then violate the zero conditional mean assumption.

(v) What is the key assumption when you apply difference-in-differences? What would be a way to investigate the plausibility of this assumption? Why is that not possible with this dataset?

Table 4:

	Dependent variable:
	illeg
post	−1.740**
	(0.673)
post:mortality	0.148***
	(0.040)
Observations	174
R^2	0.444
Adjusted R^2	−0.132
F Statistic	33.939*** (df = 2; 85)
Note:	*p<0.1; **p<0.05; ***p<0.01

The key assumption is that the treatment and control groups show "parallel trends", or more broadly, that the treatment group is a plausible counterfactual of the treated group. In this case, the way to investigate that would be two verify whether the trend of out-of-wedlock births in the departments shows a parallel development over time preceding the treatment (WWI). In this dataset, this is impossible, because $t = 2$.