Problem Set 6. Difference in Difference + Synthetic Control

Relevant material will be covered by **Nov 9**. Problem set is due **Nov 16**.

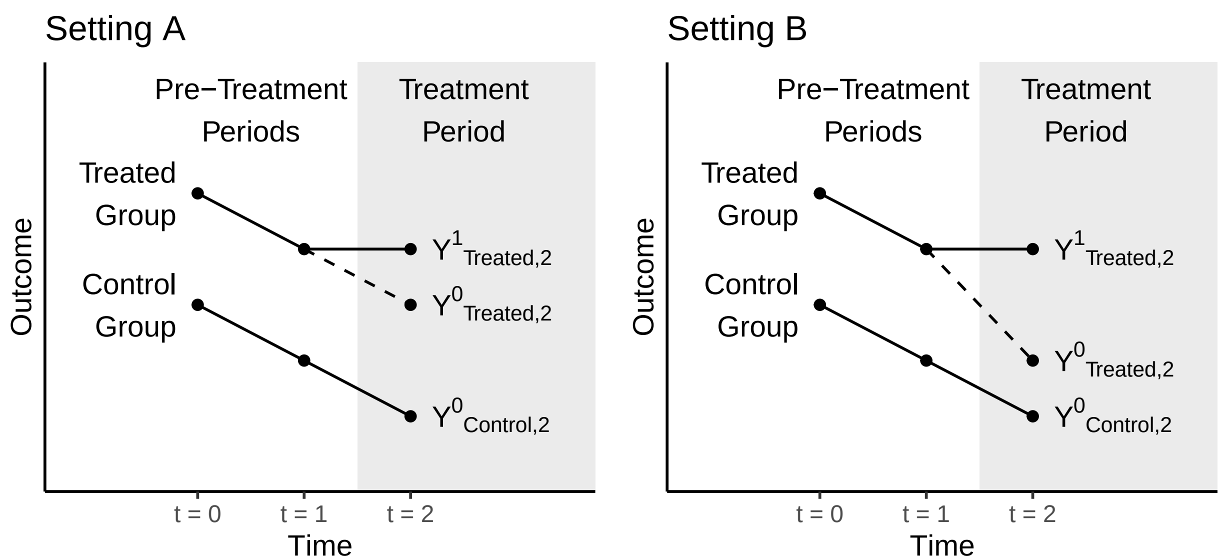
To complete the problem set, feel free to **Download the .Rmd**. Omit your name so we can have anonymous peer feedback. Submit the PDF on Canvas.

The learning goals of completing this problem set are to engage with conceptual assumptions for difference in difference and synthetic control.

```
library(tidyverse)
```

1. (25 points) Difference in Difference

In the figures below, the treated group becomes treated between time 1 and time 2. The control group never becomes treated. Figures are hypothetical scenarios that depict true potential outcomes even if those outcomes would not be observed in an actual study.



1.1 (5 points)

In which setting does the parallel trends assumption hold: A, B, neither, or both?

Answer.

- In **Setting A**, the parallel trends assumption holds. This is because, during the pre-treatment periods (time 0 to time 1), the treated and control groups have parallel trends in their outcomes.
- In **Setting B**, the parallel trends assumption does not hold. This is evident from the fact that, during the pre-treatment periods, the treated and control groups have different trends in their outcomes.

1.2 (5 points)

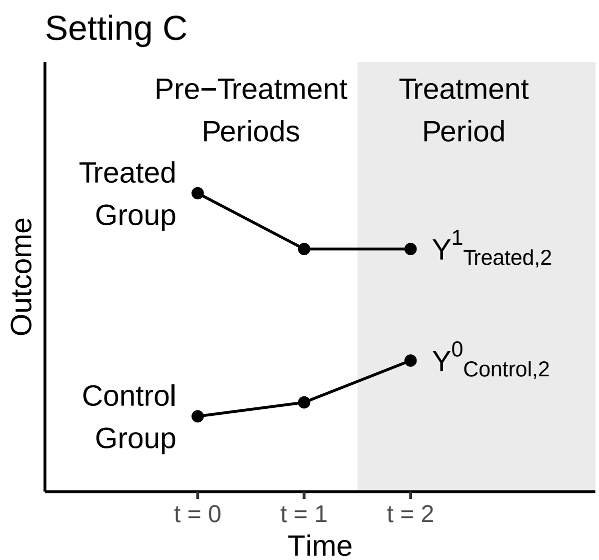
In actual data analysis, can we ever know for certain whether we are in Setting A or Setting B? If the answer is no, then tell us which outcome cannot be observed.

Answer.

- In actual data analysis, we cannot know for certain whether we are in Setting A or Setting B. The reason is that the counterfactual outcomes, represented by "{Yˆ0}['Treated,2']", cannot be observed. These are the outcomes that would have occurred for the treated group in the absence of treatment during the treatment period (time 1 to time 2).
- We observe the actual outcomes for the treated group during the treatment period, but we do not observe what would have happened if they had not received treatment. Therefore, we cannot directly compare the observed outcomes to the counterfactual outcomes to determine whether the parallel trends assumption holds.

1.3 (5 points)

A researcher comes to you with the data below, which depict only observed outcomes. That researcher wants to run a difference in difference analysis. Here, we have not depicted the counterfactual outcome because the researcher would not know it.



Why is the parallel trends assumption doubtful in this setting?

Answer. In the plotprovided for Setting C above, the parallel trends assumption appears doubtful because the treatment group (yT) and control group (yC) do not seem to have parallel trends before the treatment is introduced. Prior to the treatment period, there is a visible difference in the outcomes between the treatment and control groups. In a well-behaved DiD setting, you would expect the trends in the treatment and control groups to be similar in the absence of treatment.

1.4 (5 points)

A researcher is interested in the causal effect of a minimum wage increase on employment. They plan to study only the U.S., and they are interested in a time when the minimum wage rose simultaneously at every place in the U.S. Why won't a difference in difference design work for the researcher's question?

Answer. A difference-in-differences (DiD) design won't work for the researcher's question because it requires a treatment group experiencing the minimum wage increase and a control group not experiencing the increase. In a scenario where the minimum wage rises simultaneously across every place in the U.S., there is no genuine control group, undermining the key assumption of parallel trends.

1.5 (5 points)

Propose another design that the researcher could use to answer the question in (1.4), which may involve data outside the U.S. Answer this question in no more than 3 sentences. Many answers are possible.

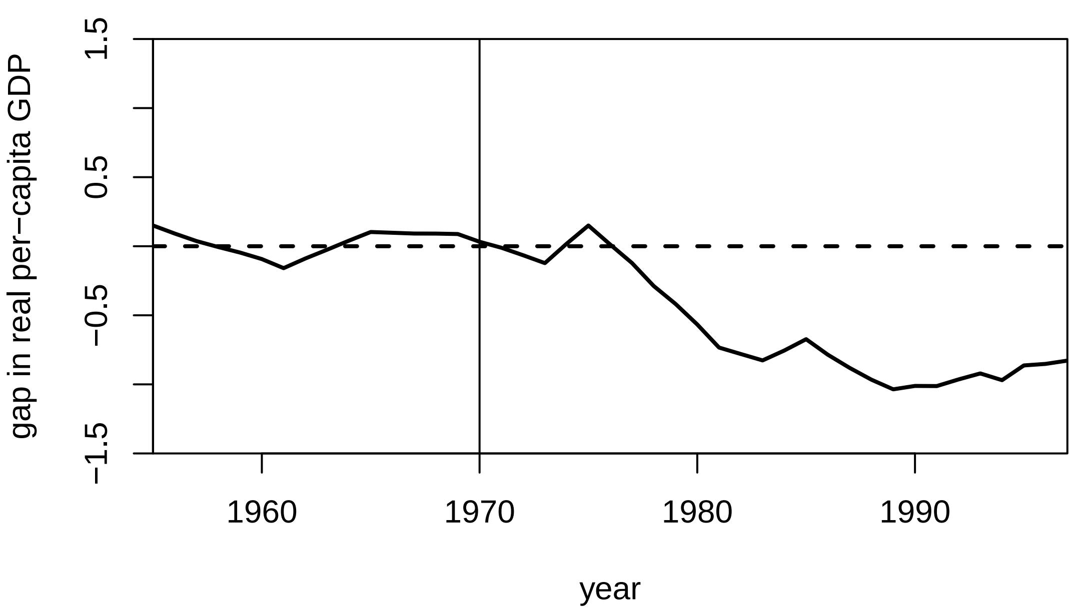
Answer. The researcher could consider using a synthetic control group approach, which involves creating a weighted combination of control units that best approximates the characteristics of the treated unit (U.S. in this case) before the minimum wage increase. This method allows the researcher to construct a plausible counterfactual scenario by selecting appropriate comparison countries with similar pre-treatment trends in employment.

2. (25 points) Synthetic Control

In discussion, we considered the paper by Abadie and Gardeazabal (2003) which estimates the effect of terrorist conflict in the Basque Country on GDP per capita. Using synthetic control, they construct a synthetic version of Basque Country. We show the selected weights and plot the gap between the observed and synthetic Basque Country below.

```
## Warning: package 'Synth' was built under R version 4.2.3
```

```
##    w.weights                   unit.names unit.numbers
## 2      0.000                    Andalucia            2
## 3      0.000                       Aragon            3
## 4      0.000      Principado De Asturias            4
## 5      0.000            Baleares (Islas)            5
## 6      0.000                    Canarias            6
## 7      0.000                   Cantabria            7
## 8      0.000             Castilla Y Leon            8
## 9      0.000          Castilla-La Mancha            9
## 10     0.851                    Cataluna           10
## 11     0.000        Comunidad Valenciana           11
## 12     0.000                 Extremadura           12
## 13     0.000                     Galicia           13
## 14     0.149        Madrid (Comunidad De)           14
## 15     0.000          Murcia (Region de)           15
## 16     0.000 Navarra (Comunidad Foral De)           16
## 18     0.000                  Rioja (La)           18
```



2.1 (10 points) Motivation

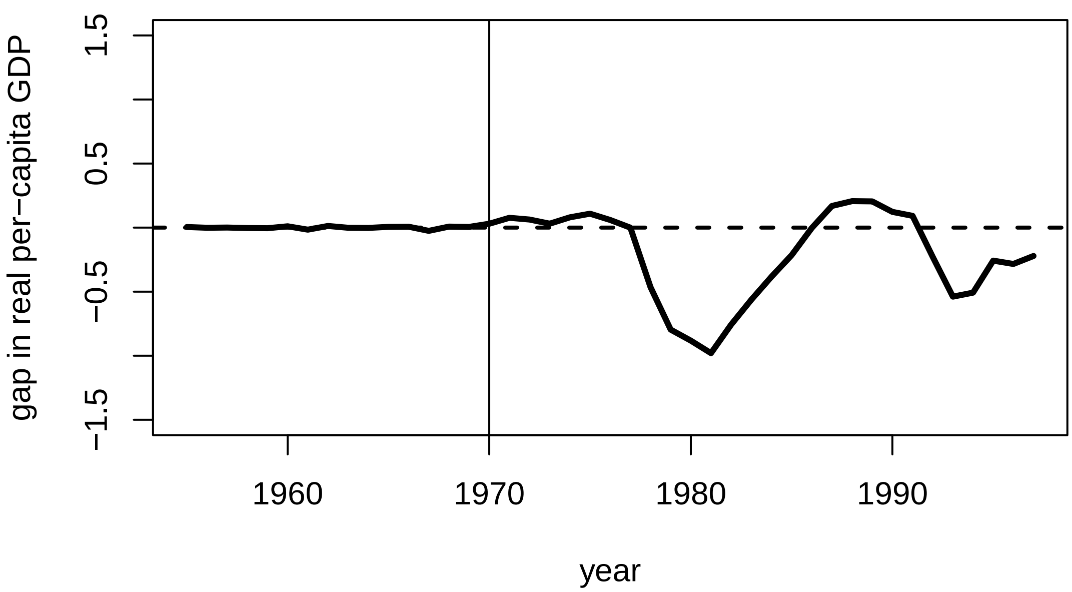
Instead of selecting the weights using synthetic control, we could have instead estimated the potential outcome for Basque Country using a regression approach. Specifically, we consider the data prior 1970 and simply regress the GDP per capita in the Basque region onto the GDP per capita in other regions to find coefficients $\hat{\beta}$ so that

$$\widehat{Y^0}_{t,Basque} = \sum_j \hat{\beta}_j Y^0_{t,j}.$$

We then use the estimated $\hat{\beta}$ to predict $\widehat{Y^0}_{t,Basque}$ after treatment occurs. The weights and gap plot are shown below. We haven't included all the regions in the code below, but you don't need to worry about that.

Looking at the estimated weights sand the gap plots, why might you prefer the synthetic control estimate over the regression based estimate? Why might you prefer the regression estimate over the synthetic control estimate?

```
##   weights                          name id
## 1  -0.632                        Aragon  3
## 2   1.256       Principado De Asturias  4
## 3  -0.586             Baleares (Islas)  5
## 4   0.438                     Cantabria  7
## 5   0.594                     Cataluna 10
## 6  -0.788         Comunidad Valenciana 11
## 7   0.155         Madrid (Comunidad De) 14
## 8  -0.245 Navarra (Comunidad Foral De) 16
## 9   1.136                   Rioja (La) 18
```



Answer.

- **Simplicity and Interpretability:** The regression approach is simpler and more straightforward. It provides coefficients

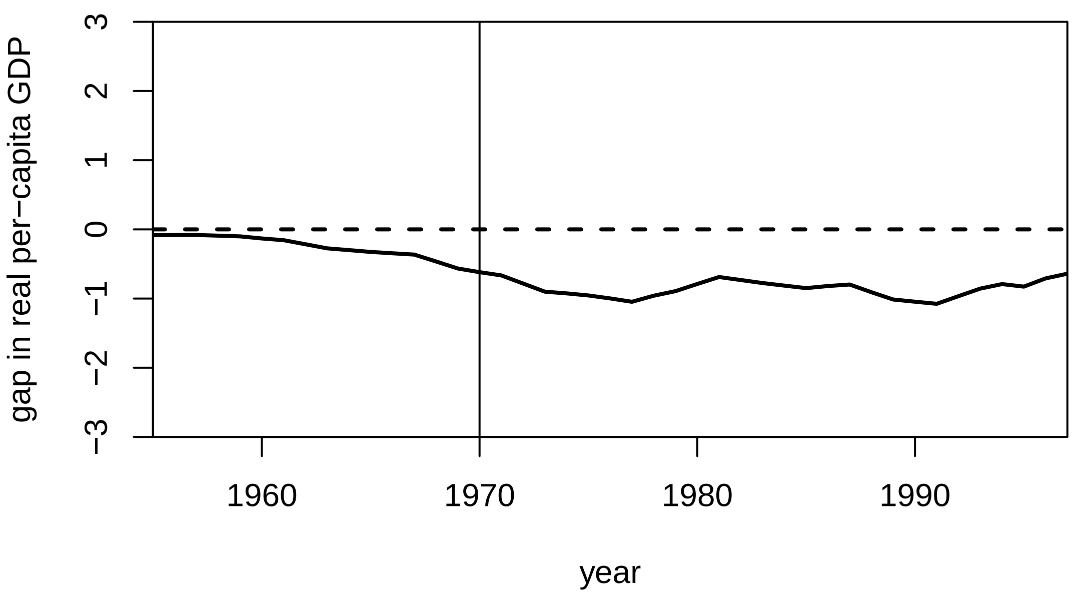
$$\hat{\beta}$$

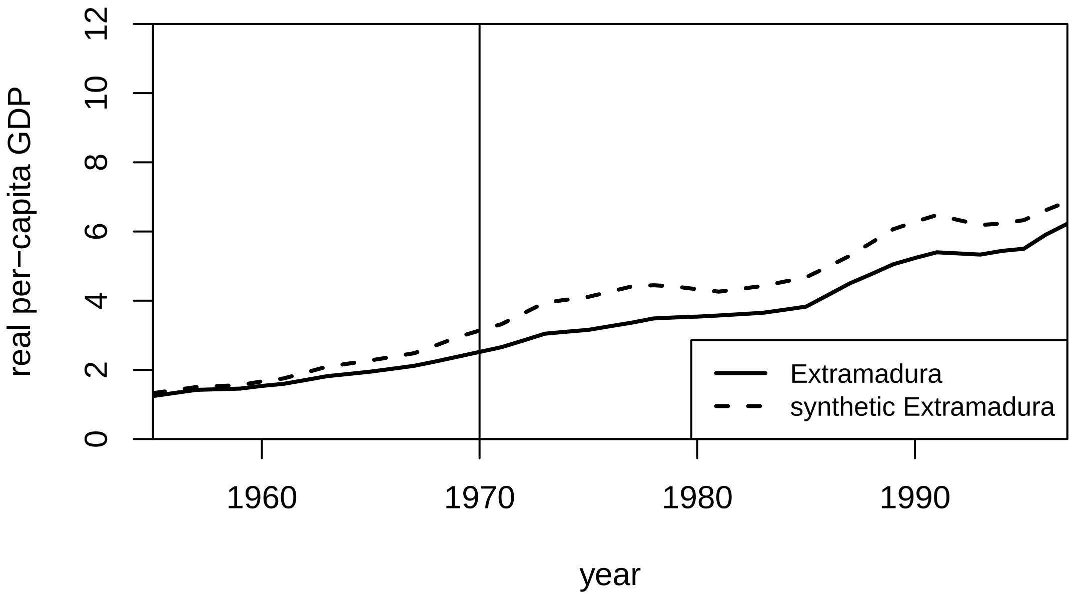
undefined that represent the average relationship between GDP per capita in Basque Country and other regions before the treatment. This simplicity can enhance the interpretability of the estimated effects.

- **Assumptions:** The regression approach relies on fewer assumptions compared to synthetic control, which assumes that the control units can be combined linearly to match the treated unit's pre-treatment outcomes. If the assumptions of synthetic control are not met, the estimates may be biased.

2.2 (7.5 points) Assessing fit

Using the same dataset, suppose we wanted to estimate the causal effect for some policy implemented in Extramadura (another region in Spain) in 1970. Running synthetic control gives an estimate of almost -1000 dollars in 1990. Looking at the plots below, why might you be skeptical of the resulting estimate?



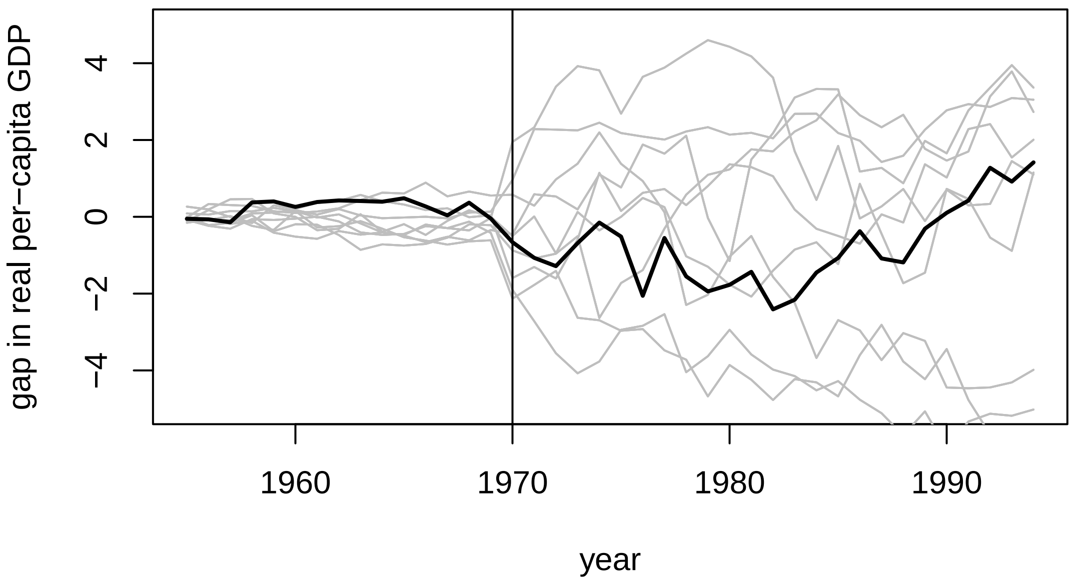


Answer.

- **Overfitting:** The synthetic control method may be prone to overfitting, especially when the number of predictors is large relative to the number of pre-treatment periods. The estimated weights might capture noise in the data rather than a true causal effect.
- **Extrapolation:** The estimated effect is based on extrapolation from the pre-treatment period to the post-treatment period. Extrapolating too far beyond the observed data can result in unreliable predictions.
- **Unaccounted Confounders:** The negative estimate may be indicative of unaccounted confounders or factors not included in the model that affect both the treatment and control units.

2.3 (7.5 points) Hypothesis testing

Suppose we used synthetic control for each of the other regions in Spain as a placebo test. Below, we show a hypothetical plot for the gap between the observed and synthetic values for each region. The difference between observed and synthetic Basque country is shown in the dark black line and the others are shown in gray. Note this is made up data. However, if this were the real plot you saw, would you be confident that there was a non-zero causal effect for Basque Country? Explain why or why not.



Answer.

If the placebo test using synthetic control for other regions in Spain produces gap plots similar to the hypothetical plot shown (gray linesclustering around zero), it would raise doubts about the confidence in the estimated causal effect for Basque Country. The dark black line represents the observed-synthetic gap for Basque Country, and if other regions exhibit similar gap patterns, it suggests that the observed gap in Basque Country might be within the range of what could happen by chance. Therefore, confidence in attributing the observed gap to a non-zero causal effect would be diminished.