# did

*haiyan yu*

*1/18/2020*

```
#https://towardsdatascience.com/causal-inference-using-difference-in-differences-causal-impact-and-synth
  library(forecast)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
## Registered S3 methods overwritten by 'forecast':
##   method             from
##   fitted.fracdiff    fracdiff
##   residuals.fracdiff fracdiff
```

```
  library(ggplot2)
  library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
  library(tidyr)
  library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
  library(ggplot2)
  library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
  library(tableone)
  library(lattice)
  library(pwr)
  library(rcompanion)
```

```
##
## Attaching package: 'rcompanion'

## The following object is masked from 'package:forecast':
##
##     accuracy
```

```
  library(scales)
  library(plm)
```

```
##
## Attaching package: 'plm'

## The following object is masked from 'package:data.table':
##
##     between

## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
  library(readxl)
  library(MatchIt)
  library(lfe)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

##
## Attaching package: 'lfe'

## The following object is masked from 'package:plm':
##
##     sargan
```

```
  library(Synth)
```

```
## ##
## ## Synth Package: Implements Synthetic Control Methods.

## ## See http://www.mit.edu/~jhainm/software.htm for additional information.
```

```
  library(gsynth)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
  library(panelView)
  library(CausalImpact)
```

```
## Loading required package: bsts

## Loading required package: BoomSpikeSlab

## Loading required package: Boom

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'Boom'

## The following object is masked from 'package:stats':
##
##     rWishart

##
## Attaching package: 'BoomSpikeSlab'

## The following object is masked from 'package:stats':
##
##     knots

## Loading required package: zoo

##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: xts

##
## Attaching package: 'xts'

## The following objects are masked from 'package:data.table':
##
##     first, last

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'bsts'

## The following object is masked from 'package:BoomSpikeSlab':
##
##     SuggestBurn
```

```
library(knitr)
```

Correlation is not causation.

Then what is causation? How can it be measured?

Causation is measuring the real impact on Y because of X. E.g., What is the effect of ad campaigns on the sales of a product?

It is critical to exactly understand the causal effects of these interventions on the subject. One of the main threats to causal inference is the confounding effect from other variables. In case of ad campaigns, it could be a reduction in price of the product, change in the overall economy or various other factors that could be inducing the change in sales at the same time. So how do we correctly attribute the change in sales because of the ad campaign?.

There are two ways to estimate the true causal impact of the intervention on the subject.

- Randomized experiment: It's the most reliable method to infer the actual causal impact of the treatment, where the user induces a change in the process at random and measure the corresponding change in the outcome variable. However, in most cases, it would be impossible to conduct experiments and control the whole system to be truly at random.
- Causal Inference in Econometrics: This method involves application of statistical procedures to arrive at the causal estimate while controlling for confounders. Some approaches under this method are what we'll be looking at in this analysis. The following are the approaches:
  - Difference in Differences (DD)
  - Causal Impact
  - Synthetic Control

Using the Basque dataset, we'll estimate the economic impact of terrorist conflict in the Basque country, an autonomous community in Spain, with the help of data from 17 other regions as well. The data can be found here. Let's look at some facts about the data and the experimental design.

## Data Description:

Some facts about the data:

- Dataset contains information from year 1955 - 1997
- Information about 18 Spanish regions is available
  - One of which is average for the whole country of Spain (we'll remove that)
- The treatment year is considered to be year 1975
- The treatment region is "Basque Country (Pais Vasco)"
- The economic impact measurement variable is GDP per capita (in thousands)

```
data(basque)
```

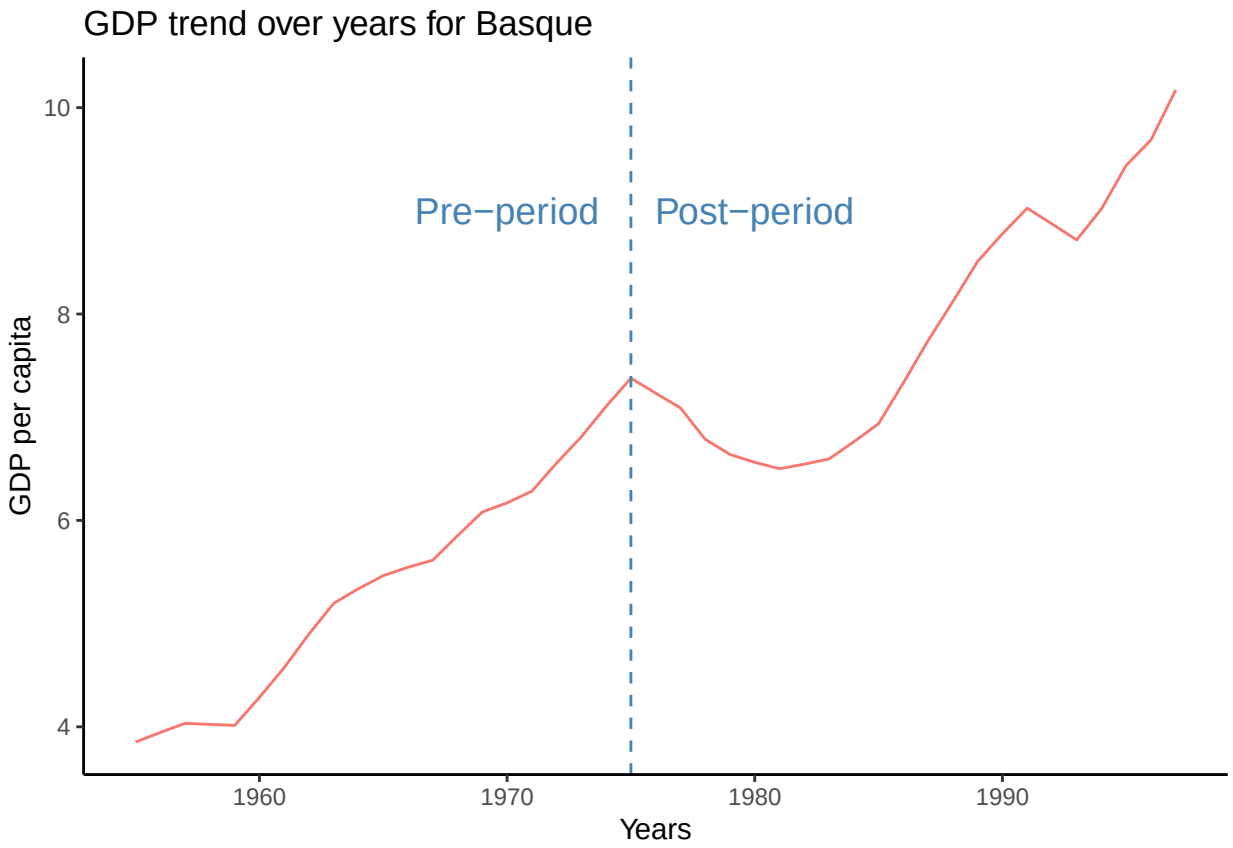
```
unused <- c("sec.agriculture", "sec.energy" , "sec.industry" , "sec.construction" ,
            "sec.services.venta" , "sec.services.nonventa", "school.illit", "school.prim",
            "school.med", "school.high", "school.post.high", "popdens")
basq_clean <- basque[,!(names(basque) %in% unused)]
basq_clean <- basq_clean %>%
  mutate(post = ifelse(year > 1975, 1, 0),
         treat = ifelse(regionname == "Basque Country (Pais Vasco)", 1, 0),
         regionname = as.factor(regionname)) %>%
  filter(regionno != 1)
```

The analysis has been done in R and the source codes can be found in my GitHub. We'll get started with the approaches mentioned.

## First Difference Estimate

Before going into Difference in Differences method, let's look at First Differences and what it does. Our goal here is to quantify the impact of GDP before and after the terrorist conflict in Basque country. Naively, we can actually achieve this by constructing a first difference regression and observing the estimate. Let's look at the general trend of GDP per capita for the Basque country.

```
# Calculating first differences;  message=FALSE, warning=FALSE
basq_fdid <- basq_clean %>%
  filter(treat == 1)
ggplot(basq_fdid, aes(x=year, y=gdpcap)) +
  geom_line(color = "#F8766D") + theme_classic() +
  geom_vline(xintercept=1975, color = "steelblue", linetype = "dashed") +
  labs(title="GDP trend over years for Basque",
       y="GDP per capita",x="Years", color = "Region") +
  annotate("text", x = 1970, y = 9, label = "Pre-period", size  =5, color = "steelblue") +
  annotate("text", x = 1980, y = 9, label = "Post-period", size  =5, color = "steelblue")
```




From the graph, we can see how the trend in GDP per capita plunges right after the terrorist intervention and then increases back all over again. Our goal is to capture the magnitude of plunge that we see. The first difference estimate will tell us the difference in GDP before and after the treatment. Let's construct a first difference equation by having GDP as the dependent variable and pre-post indicator as the independent variable.

```
# Calculating first differences; message=FALSE, warning=FALSE
f_did <- lm(data = basq_fdid, gdpcap ~ post)
stargazer(f_did, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                gdpcap
## ------------------------------------------------
## post                          2.484***
##                                (0.352)
##
## Constant                      5.382***
##                                (0.252)
##
## ------------------------------------------------
## Observations                     43
## R2                              0.549
## Adjusted R2                     0.538
## Residual Std. Error       1.153 (df = 41)
```

```
## F Statistic           49.921*** (df = 1; 41)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The coefficient of post indicator suggests that there is an increase of GDP per capita by ~2.5 units being in the post period after the terrorist conflict and that's not quite what we want.

That's because there's an expected problem as we mentioned earlier. The trend in GDP could have been altered because of a lot of other variables, besides the terrorist conflict, occurring at the same time - otherwise known as confounders. Possible confounders in this case are:

- Passing of a trade law which would affect local businesses and GDP
- Mutiny within local groups
- Perception of corrupt or dysfunctional government

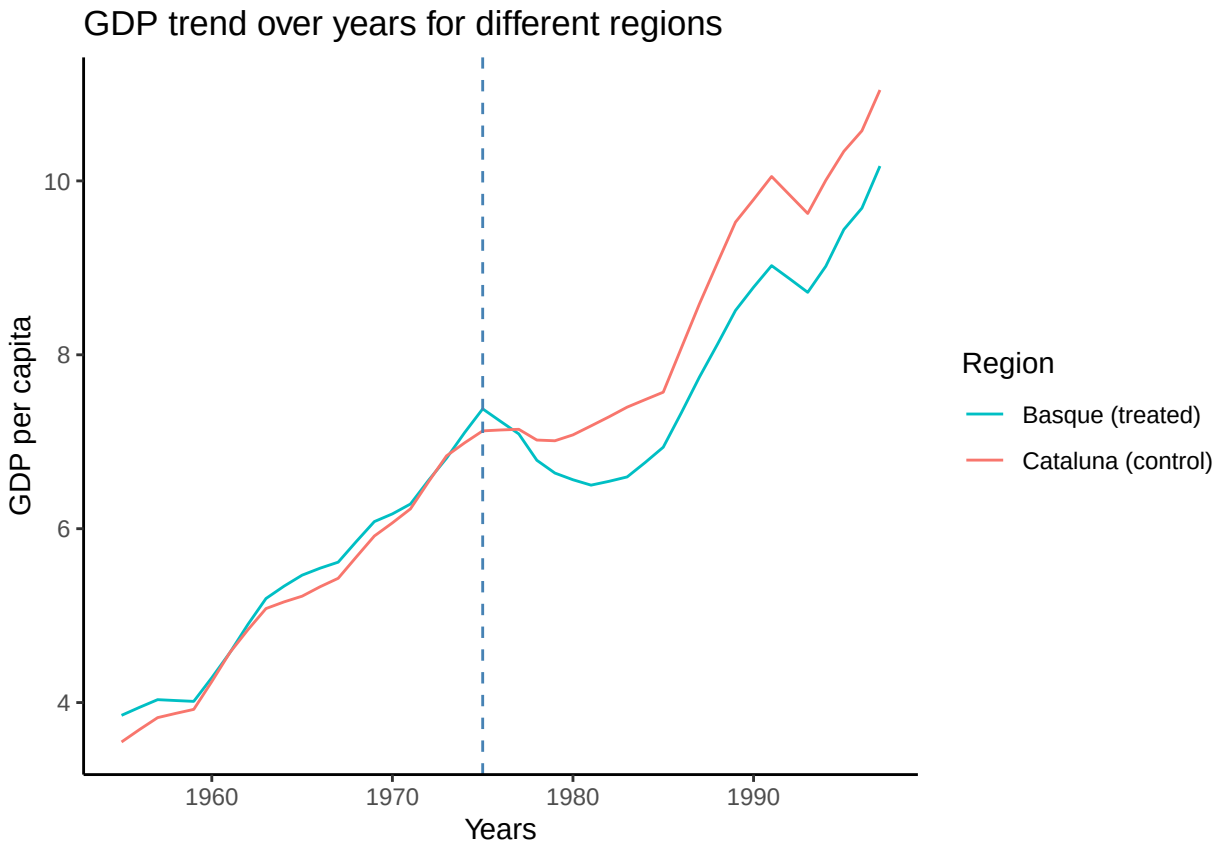
The solution to this is to compare the trend to a control region which was not impacted by terrorist conflict. This comparison allows us to remove the confounding effect after the intervention period and arrive at the real causal impact. That's where Difference in differences helps.

## Difference in differences (DD):

The underlying assumption of difference-in-differences (DD) design is that the trend of the control group provides an adequate proxy for the trend that would have been observed in the treatment group in the absence of treatment. Thus, the difference in change of slope would be the actual treatment effect. The assumption here is that there treatment group and control group must follow the same trend in the pre-period.

For this analysis, the control region was identified by spotting for the region that had the lowest variation in % difference of GDP across years between each region and Basque country. Alternatively, we can look for the control region by eyeballing the GDP trend for treatment and control groups when feasible. In this case, Cataluna region was recognized to be the best control region. Let's look at the GDP trend for test and control regions below:

```
#Picking the closest control group based on gdp
pre <- basq_clean %>%
  filter(post == 0) %>%
  left_join(dplyr::select(basq_clean[basq_clean$post==0 & basq_clean$treat == 1, ], gdpcap, year),
            by = c("year"= 'year')) %>%
  mutate(perc_diff = (gdpcap.y - gdpcap.x) / gdpcap.y) %>%
  group_by(regionname) %>%
  summarise(gdp_var = abs(var(perc_diff))) %>%
  arrange(gdp_var)
# Validating assumption
did_data <- basq_clean %>%
  filter(regionname %in% c("Basque Country (Pais Vasco)", "Cataluna"))
ggplot(did_data, aes(x=year, y=gdpcap, group = regionname)) +
  geom_line(aes(color = regionname)) +
  theme_classic() +
  geom_vline(xintercept=1975, color = "steelblue", linetype = "dashed") +
  labs(title="GDP trend over years for different regions",
       y="GDP per capita",x="Years", color = "Region") +
  scale_color_manual(labels = c("Basque (treated)", "Cataluna (control)"), values = c("#00BFC4", "#F876
```




GDP trend for Cataluna region goes hand in hand with Basque's GDP with an exception for a few years in the pre-period. Thus, there should be no problem in considering Cataluna to be our control region.

Let's go ahead and fit the regression with GDP as the dependent variable and treatment indicator and pre-post indicator as the independent variables. The key aspect here is to feed the interaction between treatment and pre-post indicator as we want the estimate to contain the effect of being treated along with being in post-period in comparison to not being treated and being in pre-period. After fitting, let's look at the regression results below:

```
# Difference in differences;  message=FALSE, warning=FALSE
did <- lm(data = did_data, gdpcap ~ treat*post)
stargazer(did, type="text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                gdpcap
## ------------------------------------------------
## treat                           0.139
##                                (0.376)
##
## post                          3.339***
##                                (0.371)
##
## treat:post                     -0.855
##                                (0.525)
```

```
##
## Constant                       5.244***
##                                 (0.266)
##
## ------------------------------------------------
## Observations                      86
## R2                               0.607
## Adjusted R2                      0.593
## Residual Std. Error         1.218 (df = 82)
## F Statistic           42.279*** (df = 3; 82)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Looking at the estimate of the interaction variable suggests that the GDP in Basque country reduced by 0.855 units because of the terrorist intervention that happened. Now there is a stark difference between estimates provided by First difference method and the DD method. Quantitatively, we can see how the First difference estimates could be deceiving and naive to look at.

If you're interested, you can read more about Difference in difference here. For now, let's move on to other causal inference methods.

## Causal Impact

This is a methodology developed by Google to estimate the causal impact of a treatment in the treated group. The official documentation can be found here.

The motivation to use Causal Impact methodology is that the Difference in differences in limited in the following ways:

- DD is traditionally based on a static regression model that assumes independent and identically distributed data despite the fact that the design has a temporal component
- Most DD analyses only consider two time points: before and after the intervention. In practice, we also have to consider the manner in which an effect evolves over time, especially its onset and decay structure

The idea here is to use the trend in the control group to forecast the trend in the treated group which would be the trend if the treatment had not happened. Then the actual causal estimate would be the difference in the actual trend vs the counter-factual trend of the treated group that we predicted. Causal Impact uses Bayesian structural time-series models to explain the temporal evolution of an observed outcome. Essentially, Causal Impact methodology is very close to the Synthetic control methodology we are going to see next.

Control region in this case is considered to be Cataluna again. With the treated and control region's GDP in place, let's feed them to the Causal Impact function in R and look at the results.

```r
# Causal Impact
basq_CI <- basq_clean %>%
  filter(regionname %in% c("Basque Country (Pais Vasco)", "Cataluna")) %>%
  mutate(date = as.Date(paste0(year, "-01", "-01"), format = "%Y-%m-%d")) %>%
  dplyr::select(date, regionname, gdpcap) %>%
  spread(regionname, gdpcap)
names(basq_CI) <- c("date", "Basque", "another")
pre.period <- as.Date(c("1955-01-01", "1975-01-01"))
post.period <- as.Date(c("1976-01-01", "1997-01-01"))
impact <- CausalImpact(basq_CI, pre.period, post.period)
summary(impact)
```

```
## Posterior inference {CausalImpact}
##
##                          Average          Cumulative
## Actual                   7.9              173.1
## Prediction (s.d.)        8.6 (0.32)       189.5 (6.93)
## 95% CI                   [8, 9.2]         [175, 203.4]
##
## Absolute effect (s.d.)   -0.75 (0.32)     -16.46 (6.93)
## 95% CI                   [-1.4, -0.099]   [-30.3, -2.169]
##
## Relative effect (s.d.)   -8.7% (3.7%)     -8.7% (3.7%)
## 95% CI                   [-16%, -1.1%]    [-16%, -1.1%]
##
## Posterior tail-area probability p:   0.01542
## Posterior prob. of a causal effect:  98.458%
##
## For more details, type: summary(impact, "report")
```

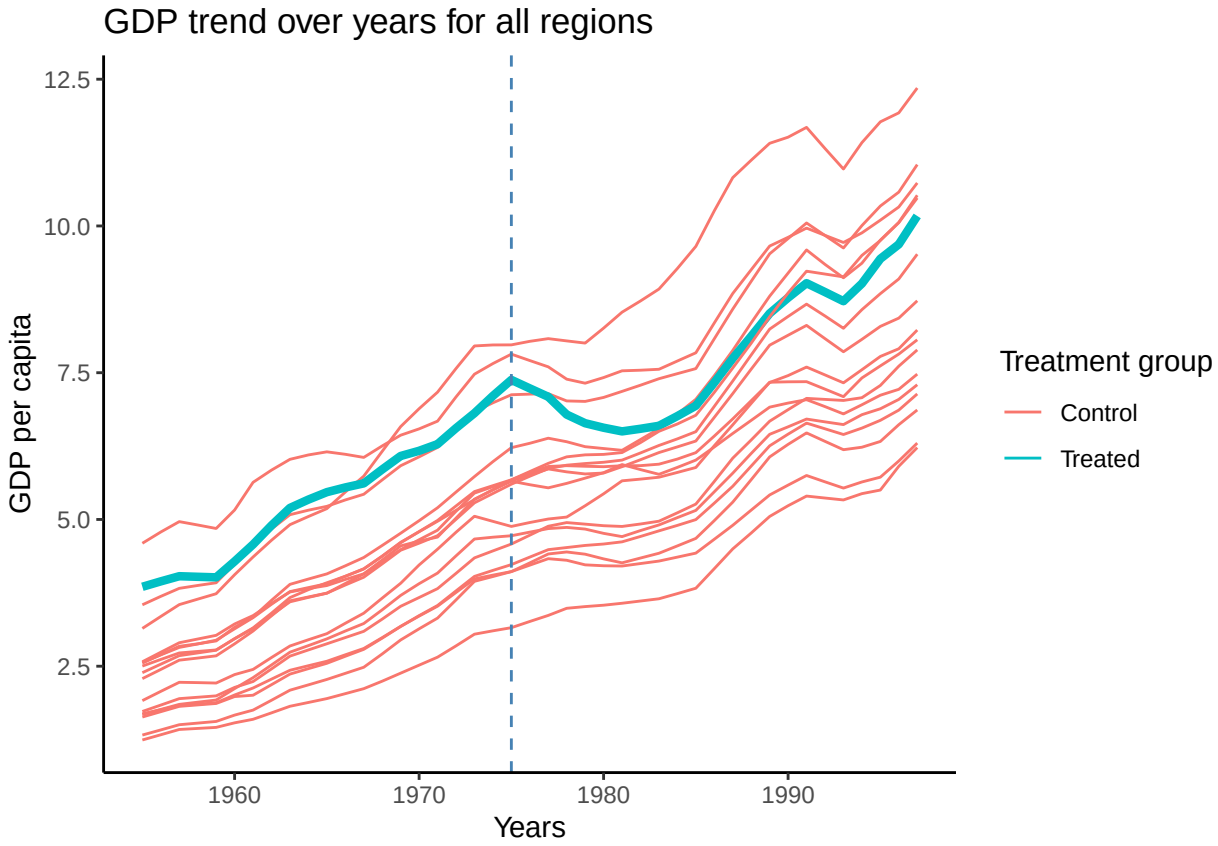
The Absolute effect is the difference in GDP per capita between what the actual GDP was after the treatment and what the GDP would have been if the treatment had not occurred. From the results, we can see the absolute effect gives us a value of -0.76 which means that the GDP per capita reduced by 0.76 units i.e. 8.8% because of the terrorist conflict that happened in Basque country. This is almost equal to what we saw using Difference in Differences method. For those interested about Causal Impact, an exhaustive explanation of the method is given by the authors of this method that can be found here.

## Synthetic Control:

Synthetic control is a technique which is very similar to Causal Impact in estimating the true impact of a treatment. Both the methods use the help of control groups to construct a counter-factual of the treated group giving us as idea of what the trend is if the treatment had not happened. The counter-factual GDP of the treated group would be predicted by the GDP of the control groups and also other possible covariates in the control group. The synth algorithm predicts the counter-factual by assigning weights to the regressors in the control groups which is helpful in identifying individual regressors and their influence in prediction. Ultimately, the true causal impact is the difference in GDP between actual GDP and the counter-factual GDP if the treatment had not happened.

The difference between Synthetic Control and Causal Impact is that Synthetic Control uses only pre-treatment variables for matching while Causal Impact uses the full pre and post-treatment time series of predictor variables for matching. Let's look at a plot for GDP trend in all the 17 other regions in the data.

```
basq_synth <- basq_clean %>%
  rename(Y = gdpcap) %>%
  mutate(regionname = as.character(regionname))
ggplot(basq_synth, aes(x=year,y=Y,group=regionno)) +
  geom_line(aes(color=as.factor(treat), size=as.factor(treat))) +
  geom_vline(xintercept=1975,linetype="dashed", color = "steelblue") + theme_classic() +
  labs(title="GDP trend over years for all regions",
       y="GDP per capita",x="Years", color = "Treatment group") +
  scale_color_manual(labels = c("Control", "Treated"), values = c("#F8766D", "#00BFC4")) +
  scale_size_manual(values = c(0.5,1.5), guide = 'none')
```




As seen from the plot above, all the control regions have a similar upward trend in GDP as Basque country's in the pre-period. This suggests that GDP in Basque could be constructed fairly accurately using the data from other regions.

The implementation of synthetic control on this problem statement has already been given in the official documentation of the package found here.

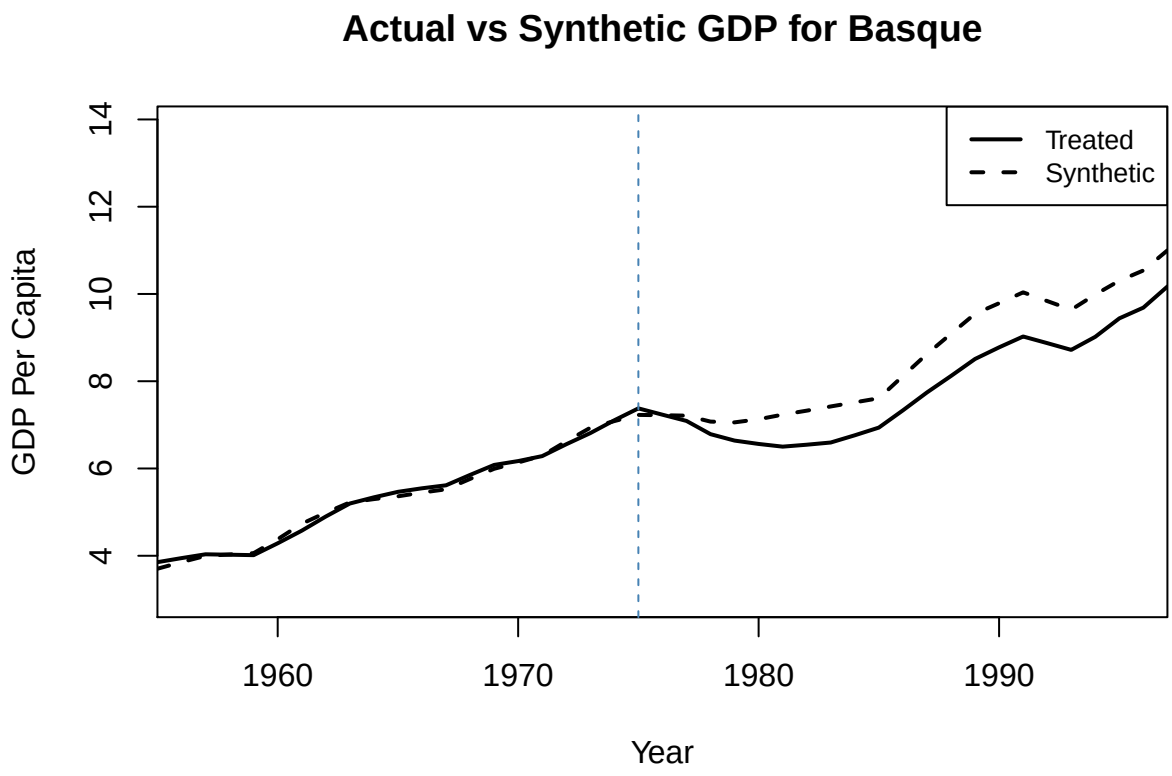
```
# synth; message=FALSE, warning=TRUE, include=FALSE
dataprep.out <-
  dataprep(
  foo = basque
  ,predictors= c("school.illit",
                 "school.prim",
                 "school.med",
                 "school.high",
                 "school.post.high"
                 ,"invest"
                 )
   ,predictors.op = c("mean")
   ,dependent     = c("gdpcap")
   ,unit.variable = c("regionno")
   ,time.variable = c("year")
   ,special.predictors = list(
    list("gdpcap",1960:1969,c("mean")),
    list("sec.agriculture",seq(1961,1969,2),c("mean")),
    list("sec.energy",seq(1961,1969,2),c("mean")),
    list("sec.industry",seq(1961,1969,2),c("mean")),
    list("sec.construction",seq(1961,1969,2),c("mean")),
```

```
    list("sec.services.venta",seq(1961,1969,2),c("mean")),
    list("sec.services.nonventa",seq(1961,1969,2),c("mean")),
    list("popdens",1969,c("mean")))
    ,treatment.identifier  = 17
    ,controls.identifier   = c(2:16,18)
    ,time.predictors.prior = c(1964:1969)
    ,time.optimize.ssr     = c(1960:1969)
    ,unit.names.variable   = c("regionname")
    ,time.plot            = c(1955:1997)
    )
synth.out = synth(dataprep.out)
```

```
##
## X1, X0, Z1, Z0 all come directly from dataprep object.
##
##
## ****************
##  searching for synthetic control unit
##
##
## ****************
## ****************
## ****************
##
## MSPE (LOSS V): 0.008864606
##
## solution.v:
##  0.02773094 1.194e-07 1.60609e-05 0.0007163836 1.486e-07 0.002423908 0.0587055 0.2651997 0.02851006 0
##
## solution.w:
##  2.53e-08 4.63e-08 6.44e-08 2.81e-08 3.37e-08 4.844e-07 4.2e-08 4.69e-08 0.8508145 9.75e-08 3.2e-08 5
```

After execution, Let's look at the plot between actual GDP and the counter-factual GDP.

```
# Two native plotting functions.
# Path.plot() plots the synthetic against the actual treated unit data.
path.plot(dataprep.res = dataprep.out, synth.res = synth.out,Xlab="Year",Ylab="GDP Per Capita")
abline(v=1975,lty=2,col="steelblue")
title("Actual vs Synthetic GDP for Basque")
```




```
# Gaps.plot() shows the deviation between the synthetic and the actual over time.
#gaps.plot(dataprep.res = dataprep.out, synth.res = synth.out,Xlab="Year",Ylab="Diff in GDP Per Capita",
#abline(v=1975,lty=2,col="red")
```

The path-graph shows how the smooth relationship is between the synthetic and actual trend in the pre-period and how it deviates gradually once the treatment happens. That difference in trends in the post-period would be our average treatment effect.

The root mean squared error of the Actual and Synthetic trends were found to be 0.57 units.

```
# Calculating root mean squared error between actual and synthetic
round(sqrt(mean((dataprep.out$Y1plot - (dataprep.out$Y0plot %*% synth.out$solution.w))^2)), 2)
```

```
## [1] 0.57
```

We can conclude that the true causal impact of terrorist conflict on Basque country is reduction of GDP by 0.57 units calculated using Synthetic control method.

Let's look at the comparison of results from all three methods below:

```
labels <- c("DiD", "Causal Impact", "Synthetic Control")
values <- c(-0.85, -0.76, -0.57)
values_df <- data.frame(labels,values)
names(values_df) <- c("Method", "Change in GDP per captia")
kable(values_df)
```

| Method | Change in GDP per captia |
|---|---|
| DiD | -0.85 |
| Causal Impact | -0.76 |
| Synthetic Control | -0.57 |

The magnitude of the causal impact differs only by a small margin between the three methods and there is no method which will give us the "correct answer". Most times, the approaches we use will be restricted by the nature of the experiment and what causal threats we are trying to address.

Some other techniques useful for inferring causal impact are Propensity score Matching, Fixed Effects Regression, Instrumental variables and Regression Discontinuity.