Taller de Econometría Aplicada II

Diferencias en Diferencias

Prof. Manuel Lecuanda

5.2. Estimación mediante MCO

Objetivo

Aplicar la técnica cuasiexperimental de *diferencias en diferencias* mediante la estimación por MCO, y analizar e interpretar sus resultados en el contexto del problema.

Se revisará la estimación directa y por mínimos cuadrados ordinarios, así como la especificación más general con datos de panel con efectos fijos.

Referencia:

Card, D. and A. Krueger (1994) *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, The American Economic Review, Volume 84, Number 4, Pages 772-793

Diferencias en Diferencias (DiD)

En muchos casos la variable de respuesta cambia en el tiempo para un grupo de individuos o un corte transversal.

La fuente de sesgo de variable omitida en estos casos puede ser la presencia de variables no observables a nivel de categoría y tiempo.

En estos casos podemos utilizar la estrategia de identificación llamada de **Diferencias-en-Diferencias** (DiD o dif-en-dif).

Para estimar el efecto del tratamiento, simplemente se compara la variable de respuesta entre las unidades tratadas antes y después del tratamiento.

Dado que pueden existir otros factores que hayan cambiado durante el tratamiento se utiliza un grupo de control para descontar estos posibles factores y aislar el efecto del tratamiento.

Orígenes

La primera aplicación de DiD se le atribuye al médico Snow (1855) quien analizó la hipótesis de que el cólera se transmitía por beber agua contaminada.

El experimento que diseñó fue el siguiente: en 1849 dos compañías de agua, SVC y LC, obtenían el agua del contaminado río Támesis en el centro de Londres. En 1852 la compañía LC comenzó a obtener su agua de una parte más limpia del río.

En vez de sólo comparar las tasas de mortalidad en los distritos servidos por la compañía que cambió su fuente de agua, en 1852 vs 1849, utilizó a los distritos servidos por la compañía que no hizo cambios como grupo de control y comparó los cambios en las tasas de mortalidad por cólera entre 1854 y 1849 en distritos servidos por estas dos compañías. Encontró que las tasas de mortalidad en los distritos servidos por LC cayeron en comparación con el cambio producido en los distritos servidos por SVC.

Ejemplo

Para realizar esta estimación no se requiere ninguna librería en particular, haremos uso de librerías que ya se han utilizado para mostrar los cálculos.

```
# Activar las librerías

library(tidyverse)
library(knitr)
library(dplyr)
library(ggplot2)
library(kableExtra)
library(stargazer)
```

Ejemplo

Se dispone de la información de una variable de respuesta Y para cuatro individuos $I = \{1, 2, 3, 4\}$ en dos momentos del tiempo $T = \{0, 1\}$, antes y después de un tratamiento. Los dos primeros individuos reciben el tratamiento ($G = 1$) y los dos últimos no lo hacen, por lo que formarán nuestro grupo de control ($G = 0$).

Definimos las variables de la siguiente manera:

```
I<-c(1,1,2,2,3,3,4,4)
T<-c(0,1,0,1,0,1,0,1)
G<-c(1,1,1,1,0,0,0,0)
Y<-c(46,74,54,96,30,50,40,60)
```

Se busca estimar el efecto del tratamiento sobre la variable de respuesta.

Conjunto de datos

El conjunto de datos se ve de la siguiente manera:

```
datos<-data.frame(I,T,G,Y)
kbl(datos,align="cccc") %>%
kable_styling(position = "center")
```

I	T	G	Y
1	0	1	46
1	1	1	74
2	0	1	54
2	1	1	96
3	0	0	30
3	1	0	50
4	0	0	40
4	1	0	60

Conjunto de datos agrupados

```
tabla<-datos %>%
  group_by(T, G) %>%
  summarize(Media = mean(Y))
kbl(tabla,align="ccc") %>%
  kable_styling(position = "center")
```

T	G	Media
0	0	35
0	1	50
1	0	55
1	1	85

Cálculo de manera directa

Definimos los escalares de las diferencias ya calculadas por grupo:

```
T0G0<-tabla$Media[1]
T0G1<-tabla$Media[2]
T1G0<-tabla$Media[3]
T1G1<-tabla$Media[4]
```

Y las diferencias de interés son las siguientes:

```
Dif0<-T0G1-T0G0
Dif1<-T1G1-T1G0
DiD<-Dif1-Dif0
```

Cálculo de manera directa

```
tablaDiD<-tibble(T = c("Después","Antes","Diferencia"),
                 G0 = c(T1G0,T0G0,T1G0-T0G0),
                 G1 = c(T1G1,T0G1,T1G1-T0G1),
                 D =c(Dif1,Dif0,Dif1-Dif0))
kbl(tablaDiD,col.names=c("Tiempo",
           "Control","Tratamiento","Diferencia"),
    align="lccc") %>%
    kable_styling(position = "center")
```

Tiempo	Control	Tratamiento	Diferencia
Después	55	85	30
Antes	35	50	15
Diferencia	20	35	15

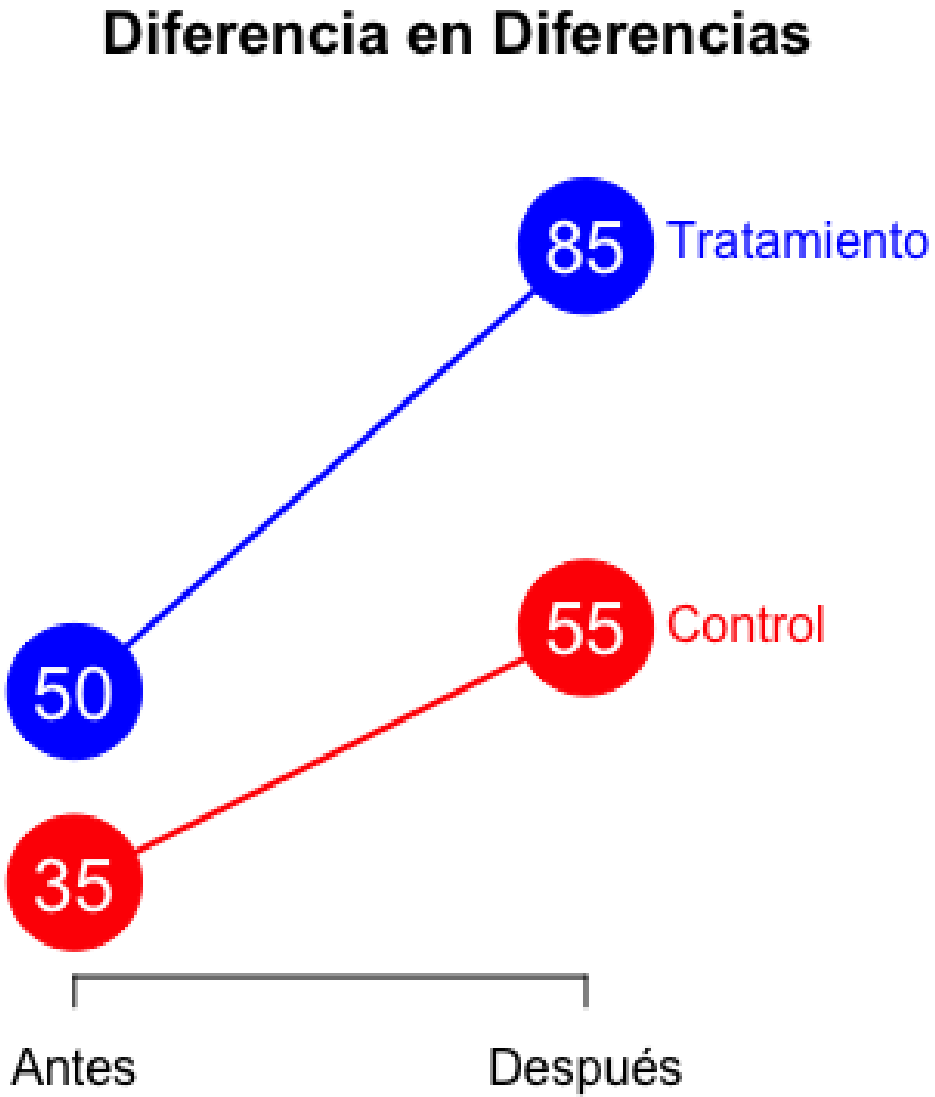
Gráficamente

```
# Definir paleta de colores
palette(c(adjustcolor("blue",alpha.f=1),
          adjustcolor("red",alpha.f=1),
          adjustcolor("darkgreen",alpha.f = 1)))

# Gráficar puntos
plot( c(0, 1, 0, 1), c(T0G1, T1G1, T0G0, T1G0), bty = "n",
      xlim = c(-0.5, 2), ylim = c(30, 90),
      col = c(1,1,2,2), pch=19, cex=6,
      xaxt = "n", yaxt = "n",
      xlab = "", ylab = "",
      main = "Diferencia en Diferencias")
```

Gráficamente

```
# Gráficar lineas
segments( x0=0, x1=1, y0=T0G0, y1=T1G0, col=2, lwd=2 )
segments( x0=0, x1=1, y0=T0G1, y1=T1G1, col=1, lwd=2 )

# Ejes y leyendas
text( c(0, 1, 0, 1), c(T0G1, T1G1, T0G0, T1G0),
      c(T0G1, T1G1, T0G0, T1G0), cex = 1.5, col =" white" )
axis( side = 1, at=c(0, 1),
      labels=c("Antes","Después" ) )
text( 1.1, T1G1, "Tratamiento",
      col = "blue", pos = 4, cex = 1 )
text( 1.1, T1G0, "Control",
      col = "red", pos = 4, cex = 1)
```

Gráficamente



Estimación mediante MCO

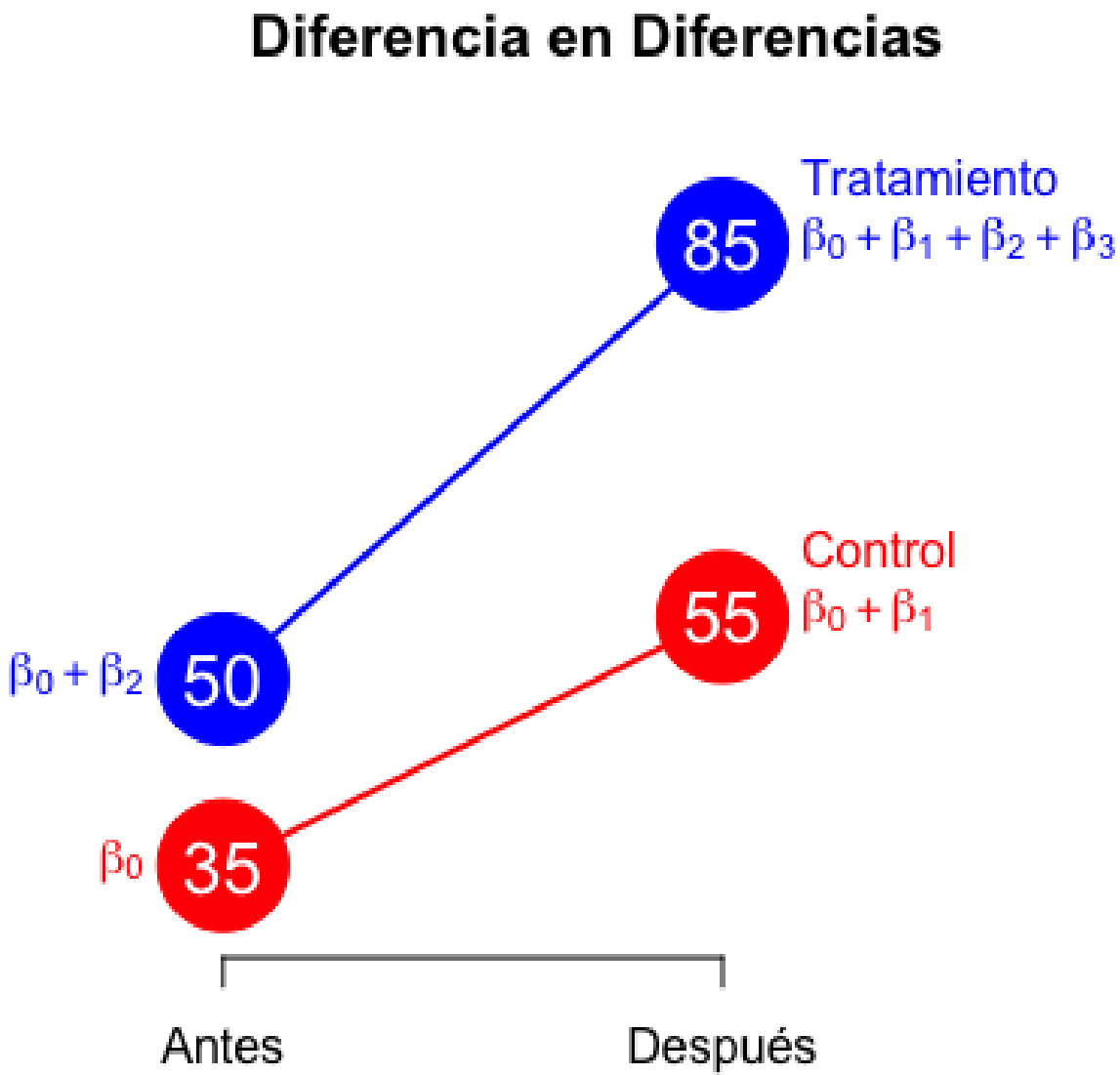
```
reg<-lm(Y~T*G,datos)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ T * G, data = datos)
##
## Residuals:
##   1   2   3   4   5   6   7   8
##  -4 -11   4  11  -5  -5   5   5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.000      6.837   5.119  0.00689 **
## T             20.000      9.670   2.068  0.10743
## G             15.000      9.670   1.551  0.19578
## T:G           15.000     13.675   1.097  0.33428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.67 on 4 degrees of freedom
## Multiple R-squared:  0.8758, Adjusted R-squared:  0.7827
## F-statistic: 9.403 on 3 and 4 DF,  p-value: 0.02769
```

Gráficamente

```
# Leyendas
text( 1.1, T1G1+5, "Tratamiento",
      col = "blue", pos = 4, cex = 1 )
text( 1.1, T1G0+5, "Control",
      col = "red", pos = 4, cex = 1)

# Interpretación
text( -0.1, T0G1, expression(paste(beta[0] + beta[2])),
      col="blue", pos=2, cex = 1 )
text( 1.1, T1G1, expression(paste(beta[0] + beta[1] + beta[2]
      + beta[3])),col="blue", pos=4, cex=1 )
text( -0.1, T0G0, expression(paste(beta[0])),
      col="red", pos=2, cex = 1 )
text( 1.1, T1G0, expression(paste(beta[0] + beta[1])),
      col="red", pos=4, cex=1 )
```

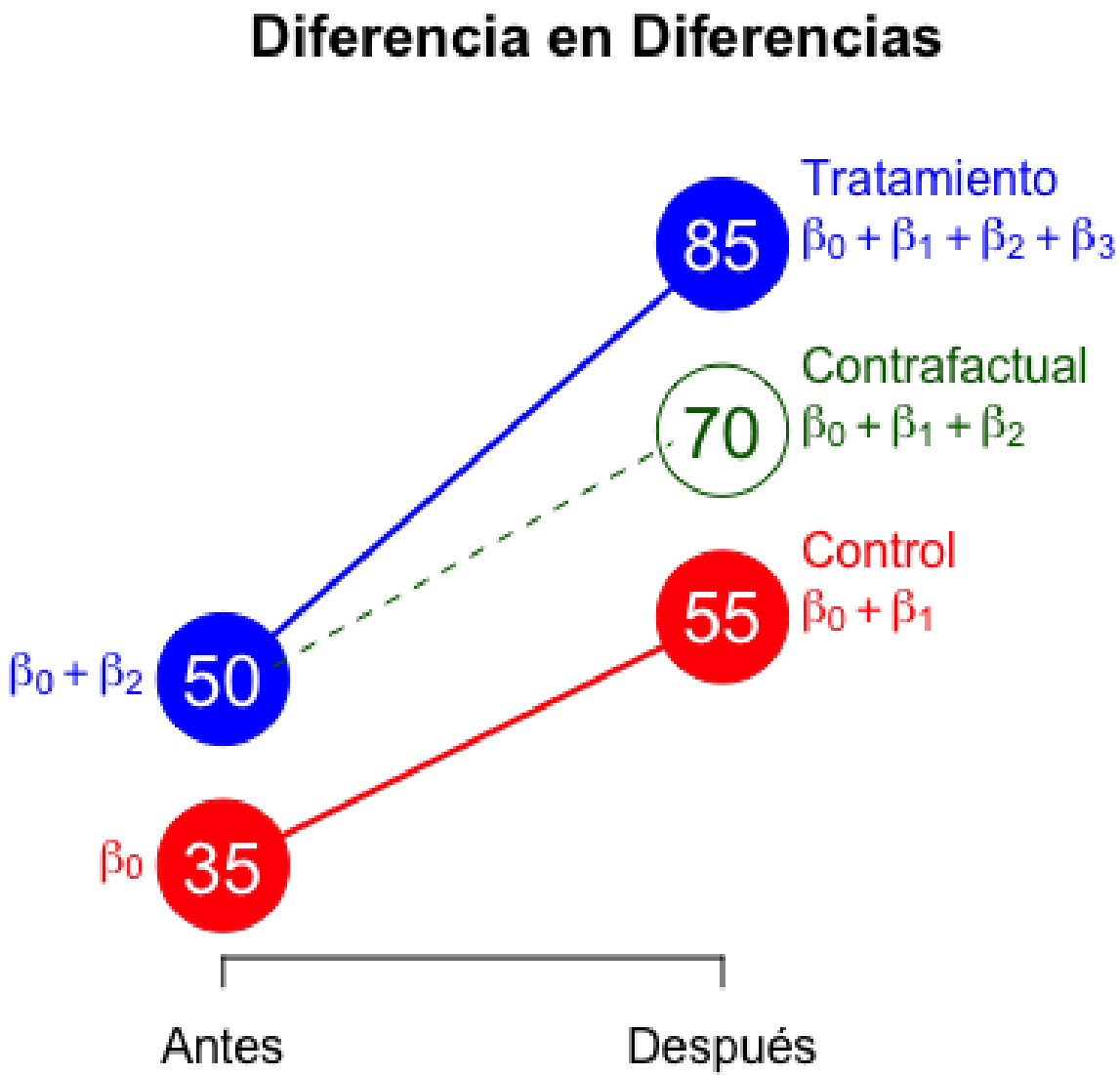
Gráficamente



Contrafactual

```
# Contrafactual
CF<-T0G1-T0G0+T1G0
points( 1, CF, cex=6, col=3 )
segments( x0=0.1, x1=0.9, y0=T1+1, y1=CF-1,
          col=3, lwd=1, lty=2 )
text( 1, CF, CF, col=3, cex=1.5 )
text( 1.1, CF+5, "Contrafactual", col=3, pos=4, cex=1 )
text( 1.1, CF, expression(paste(beta[0]+beta[1]+beta[2])),
      col=3, pos=4, cex=1 )
```

Gráficamente



Interpretación

Así, la significancia de cada uno de los coeficientes de la regresión estimada es informativa:

- β_0: ¿Es significativo el valor promedio de la variable de respuesta en el grupo de control antes del tratamiento?
- β_1: ¿Es significativa la diferencia entre los valores promedio del grupo de tratamiento y control después del tratamiento?
- β_2: ¿Es significativa la diferencia entre los valores promedio del grupo de control antes y después del tratamiento?
- β_3: ¿Es significativo el impacto del tratamiento en la variable de respuesta? **(estimador de diferencias en diferencias)**

Especificación alternativa

Alternativamente, esta estimación coincide con la que se puede obtener mediante un modelo de datos de panel con efectos fijos para cada individuo.

Para ello se construyen las variables dicotómicas para cada categoría del corte trasnversal y se estima el modelo, omitiendo el intercepto y la variable dicotómica de grupo para evitar la multicolinealidad perfecta.

```
reg2<-lm(Y~factor(I)+T*G-G)
```

Al visualizar la estimación, los resultados coinciden, en particular para el efecto del impacto del tratamiento.

```
summary(reg2)
```

Especificación alternativa

```
summary(reg2)
```

```
##
## Call:
## lm(formula = Y ~ factor(I) + T * G - G)
##
## Residuals:
##          1          2          3          4          5          6          7
##  3.500e+00 -3.500e+00 -3.500e+00  3.500e+00 -3.220e-15  2.109e-15  8.882e-16
##          8
## -1.332e-15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   42.500      4.287   9.915   0.0100 *
## factor(I)2    15.000      4.950   3.030   0.0938 .
## factor(I)3   -12.500      6.062  -2.062   0.1753
## factor(I)4    -2.500      6.062  -0.412   0.7201
## T             20.000      4.950   4.041   0.0561 .
## T:G           15.000      7.000   2.143   0.1654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```