Diferencias en Diferencias

Maestría en Economía Uninorte

Carlos Andrés Yanes

2024-08-24

Introducción

Vamos a trabajar con la base de datos >¿Un mejor seguro médico aumenta el consumo de atención sanitaria?

El experimento rand de 1970 sobre el **seguro** médico es un gran experimento social Asigno aleatoriamente distintos niveles de seguro médico a distintas familias.

- Las familias participan durante 3-5 años.
- Para garantizar que la participación de los participantes no les perjudicara en comparación con su seguro médico habitual en comparación con su seguro médico habitual
- Todas las pólizas tenían un límite anual (MDE) a partir del cual la asistencia sanitaria era gratuita

Preambulo

El método de **Diferencias en Diferencias (DID)** es una técnica econométrica ampliamente utilizada para evaluar el **impacto causal** de políticas públicas, programas o intervenciones que se implementan en ciertas unidades (como individuos, empresas, regiones, etc.) pero no en otras. Este método se basa en comparar la evolución de un resultado de interés entre un grupo tratado (que recibe la intervención) y un grupo de control (que no la recibe) *antes* y *después* del tratamiento.

El enfoque DID asume que, en ausencia del tratamiento, la diferencia en las tendencias del resultado entre los grupos tratado y de control habría permanecido constante. De esta manera, cualquier desviación de esta tendencia paralela se atribuye al efecto causal del tratamiento.

En economía, el método de Diferencias en Diferencias es particularmente útil cuando los experimentos aleatorios no son factibles o éticos, y permite controlar por factores no observados que podrían influir en los resultados, siempre que estos factores sean constantes en el tiempo. Esta técnica ha sido utilizada en estudios que analizan desde el impacto de cambios en las políticas fiscales y laborales hasta los efectos de programas educativos y de salud, convirtiéndose en una herramienta clave para el análisis de políticas públicas.

Preparación

Antes de implementar el código de estimación, es crucial preparar la base de datos asegurando que las variables relevantes estén correctamente definidas y limpiadas. Esto implica verificar que las variables de tratamiento y resultado estén codificadas adecuadamente, que las covariables no presenten valores faltantes, y que los pesos muestrales, si son aplicables, estén correctamente asignados. Además, es esencial que los datos estén en el formato adecuado para ser utilizados en los modelos estadísticos, lo que incluye transformar variables según sea necesario y asegurarse de que todas las observaciones relevantes sean incluidas en el análisis.

Limpiar el entorno de R

```
rm(list = ls())
```

Estipulación de la base

Vamos a cargar los paquetes a utilizar en esta ocasión

```
library(pacman)
p_load(dplyr, lmtest, sandwich, plm, fixest, haven)
```

El paso a seguir es cargar la base de datos (formato stata) a R.

```
datos <- read_dta("health.dta")
```

Resumen estadistico

En las posibilidades siempre es bueno mirar que contiene y dicen nuestros datos, para ello no queda demas ir mirando primero que etiquetas traen consigo

```
glimpse(datos)
```

```
Rows: 1,071
Columns: 26
$ hhid93      <dbl> 590020, 590100, 590100, 590110, 590070, 590090, 590090, 59~
$ pcode       <dbl> 41, 41, 9, 40, 5, 44, 45, 43, 40, 9, 9, 4, 6, 5, 3, 42, 41~
$ idcommunity <dbl> 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 70, 70, 70~
$ year        <dbl> 98, 98, 93, 98, 93, 98, 98, 98, 98, 93, 93, 93, 93, 93, 93~
$ hightreat   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ post        <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1~
$ postXhigh   <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ waz         <dbl> 0.73, 0.64, -0.17, 0.59, -0.62, 3.48, 1.67, 1.08, 0.30, 2.~
$ whz         <dbl> 1.18, 2.72, -0.44, 4.62, 0.74, 6.02, 4.59, 0.93, 1.37, 2.1~
$ haz         <dbl> -0.22, -2.60, -0.04, -5.68, -2.06, -1.36, -2.76, 0.98, -1.~
$ fedu        <dbl> 0, 0, 0, 0, 0, 0, 0, 9, 10, 8, 10, 0, 0, 2, 0, 5, 0, 0, 0,~
$ medu        <dbl> 7, 0, 10, 7, 6, 10, 8, 0, 0, 6, 0, 6, 2, 4, 4, 0, 7, 2, 0,~
$ hhsizep     <dbl> 6, 7, 9, 6, 5, 16, 16, 16, 10, 9, 9, 5, 6, 5, 3, 4, 7, 4, ~
$ lntotminc   <dbl> 8.558975, 7.699238, 7.939818, 7.711250, 4.941642, 8.782922~
$ immuniz     <dbl> 2, 2, 0, 2, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1~
$ nonclinic   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ male        <dbl> 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1~
$ age         <dbl> 3.1068494, 3.0191782, 0.0000000, 3.0876713, 1.5000000, 1.0~
$ age93_0     <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0~
$ age93_1     <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
$ age93_2     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ age93_3     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0~
$ age98_0     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ age98_1     <dbl> 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ age98_2     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1~
$ age98_3     <dbl> 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
```

Luego de esto entonce si hacemos una descripción muy rapida de las estadísticas mas importantes de los datos. Siempre le piden a los investigadores establecer una tabla o descriptores para ello

```
summary(datos)
```

```
     hhid93            pcode        idcommunity         year      
 Min.   : 590010   Min.   : 2.00   Min.   : 59.0   Min.   :93.00  
 1st Qu.:2050180   1st Qu.: 9.00   1st Qu.:205.0   1st Qu.:93.00  
 Median :2230180   Median :16.00   Median :223.0   Median :93.00  
 Mean   :2066044   Mean   :24.51   Mean   :206.6   Mean   :95.33  
 3rd Qu.:2340490   3rd Qu.:41.00   3rd Qu.:234.0   3rd Qu.:98.00  
 Max.   :2440170   Max.   :57.00   Max.   :244.0   Max.   :98.00  
   hightreat           post          postXhigh           waz         
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-5.8800  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:-1.2600  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :-0.2400  
 Mean   :0.4276   Mean   :0.4669   Mean   :0.1979   Mean   :-0.2059  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 0.7600  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   : 4.9400  
      whz               haz               fedu             medu       
 Min.   :-9.890   Min.   :-9.9800   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:-0.480   1st Qu.:-2.0650   1st Qu.: 0.000   1st Qu.: 1.000  
 Median : 0.480   Median :-0.9900   Median : 0.000   Median : 5.000  
 Mean   : 0.639   Mean   :-0.9498   Mean   : 1.758   Mean   : 4.728  
 3rd Qu.: 1.385   3rd Qu.: 0.0750   3rd Qu.: 3.000   3rd Qu.: 8.000  
 Max.   : 9.990   Max.   : 9.9900   Max.   :12.000   Max.   :14.000  
    hhsizep        lntotminc         immuniz          nonclinic      
 Min.   : 2.00   Min.   :3.549   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 8.00   1st Qu.:6.179   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :10.00   Median :6.854   Median :0.0000   Median :0.0000  
 Mean   :11.17   Mean   :6.829   Mean   :0.4911   Mean   :0.1204  
```

```
 3rd Qu.:14.00   3rd Qu.:7.470   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :34.00   Max.   :9.846   Max.   :2.0000   Max.   :3.0000
      male             age           age93_0          age93_1
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:1.232   1st Qu.:0.0000   1st Qu.:0.0000
 Median :1.0000   Median :2.167   Median :0.0000   Median :0.0000
 Mean   :0.5145   Mean   :2.146   Mean   :0.1083   Mean   :0.1382
 3rd Qu.:1.0000   3rd Qu.:3.083   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :3.992   Max.   :1.0000   Max.   :1.0000
    age93_2          age93_3          age98_0           age98_1
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000
 Mean   :0.1401   Mean   :0.1466   Mean   :0.06162   Mean   :0.1307
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.0000
    age98_2          age98_3
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000
 Mean   :0.1447   Mean   :0.1298
 3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000
```

Selección clave

Vamos a intentar solo mirar un grupo de variables, no todas las que tenemos o buscamos tener en nuestro cuestionario se hacen necesarias

Subgrupo

Entonces hacemos una selección mas acorde

```
datos %>%
  select(idcommunity, year, hightreat, post, postXhigh, waz, whz,
         fedu, medu, hhsizep, lntotminc, immuniz, nonclinic, age) %>%
  glimpse()
```

```
Rows: 1,071
Columns: 14
$ idcommunity <dbl> 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 59, 70, 70, 70~
$ year        <dbl> 98, 98, 93, 98, 93, 98, 98, 98, 98, 93, 93, 93, 93, 93, 93~
$ hightreat   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ post        <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1~
$ postXhigh   <dbl> 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ waz         <dbl> 0.73, 0.64, -0.17, 0.59, -0.62, 3.48, 1.67, 1.08, 0.30, 2.~
$ whz         <dbl> 1.18, 2.72, -0.44, 4.62, 0.74, 6.02, 4.59, 0.93, 1.37, 2.1~
$ fedu        <dbl> 0, 0, 0, 0, 0, 0, 0, 9, 10, 8, 10, 0, 0, 2, 0, 5, 0, 0, 0,~
$ medu        <dbl> 7, 0, 10, 7, 6, 10, 8, 0, 0, 6, 0, 6, 2, 4, 4, 0, 7, 2, 0,~
$ hhsizep     <dbl> 6, 7, 9, 6, 5, 16, 16, 16, 10, 9, 9, 5, 6, 5, 3, 4, 7, 4, ~
$ lntotminc   <dbl> 8.558975, 7.699238, 7.939818, 7.711250, 4.941642, 8.782922~
$ immuniz     <dbl> 2, 2, 0, 2, 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1~
$ nonclinic   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ age         <dbl> 3.1068494, 3.0191782, 0.0000000, 3.0876713, 1.5000000, 1.0~
```

Estadisticas de ellas

```
datos %>%
  select(idcommunity, year, hightreat, post, postXhigh, waz, whz,
         fedu, medu, hhsizep, lntotminc, immuniz, nonclinic, age) %>%
  summary()
```

```
  idcommunity         year         hightreat           post        
 Min.   : 59.0   Min.   :93.00   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:205.0   1st Qu.:93.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :223.0   Median :93.00   Median :0.0000   Median :0.0000  
 Mean   :206.6   Mean   :95.33   Mean   :0.4276   Mean   :0.4669  
 3rd Qu.:234.0   3rd Qu.:98.00   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :244.0   Max.   :98.00   Max.   :1.0000   Max.   :1.0000  
   postXhigh           waz                whz              fedu       
 Min.   :0.0000   Min.   :-5.8800   Min.   :-9.890   Min.   : 0.000  
 1st Qu.:0.0000   1st Qu.:-1.2600   1st Qu.:-0.480   1st Qu.: 0.000  
 Median :0.0000   Median :-0.2400   Median : 0.480   Median : 0.000  
 Mean   :0.1979   Mean   :-0.2059   Mean   : 0.639   Mean   : 1.758  
 3rd Qu.:0.0000   3rd Qu.: 0.7600   3rd Qu.: 1.385   3rd Qu.: 3.000  
 Max.   :1.0000   Max.   : 4.9400   Max.   : 9.990   Max.   :12.000  
```

```
      medu            hhsizep         lntotminc         immuniz
 Min.   : 0.000   Min.   : 2.00   Min.   :3.549   Min.   :0.0000
 1st Qu.: 1.000   1st Qu.: 8.00   1st Qu.:6.179   1st Qu.:0.0000
 Median : 5.000   Median :10.00   Median :6.854   Median :0.0000
 Mean   : 4.728   Mean   :11.17   Mean   :6.829   Mean   :0.4911
 3rd Qu.: 8.000   3rd Qu.:14.00   3rd Qu.:7.470   3rd Qu.:1.0000
 Max.   :14.000   Max.   :34.00   Max.   :9.846   Max.   :2.0000
   nonclinic           age
 Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:1.232
 Median :0.0000   Median :2.167
 Mean   :0.1204   Mean   :2.146
 3rd Qu.:0.0000   3rd Qu.:3.083
 Max.   :3.0000   Max.   :3.992
```

Mas integrado

```
datos %>%
  select(waz, hightreat, post, postXhigh, whz) %>%
  summary()
```

```
      waz             hightreat          post          postXhigh
 Min.   :-5.8800   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:-1.2600   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :-0.2400   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :-0.2059   Mean   :0.4276   Mean   :0.4669   Mean   :0.1979
 3rd Qu.: 0.7600   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   : 4.9400   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      whz
 Min.   :-9.890
 1st Qu.:-0.480
 Median : 0.480
 Mean   : 0.639
 3rd Qu.: 1.385
 Max.   : 9.990
```

Diferencias en diferencias

```
highpre <- mean(datos$waz[datos$hightreat == 1 & datos$post == 0], na.rm = TRUE)
highpost <- mean(datos$waz[datos$hightreat == 1 & datos$post == 1], na.rm = TRUE)
lowpre <- mean(datos$waz[datos$hightreat == 0 & datos$post == 0], na.rm = TRUE)
lowpost <- mean(datos$waz[datos$hightreat == 0 & datos$post == 1], na.rm = TRUE)

highdiff <- highpost - highpre
lowdiff <- lowpost - lowpre
diffindiff <- highdiff - lowdiff

cat("highpre  =", highpre, "\n")
```

```
highpre  = -0.5452439
```

```
cat("highpost =", highpost, "\n")
```

```
highpost = 0.3214623
```

```
cat("highdiff =", highdiff, "\n")
```

```
highdiff = 0.8667062
```

```
cat("lowpre   =", lowpre, "\n")
```

```
lowpre   = -0.4141846
```

```
cat("lowpost  =", lowpost, "\n")
```

```
lowpost  = -0.06909722
```

```
cat("lowdiff  =", lowdiff, "\n")
```

```
lowdiff  = 0.3450874
```

```
cat("diffindiff =", diffindiff, "\n")
```

```
diffindiff = 0.5216188
```

Diff-in-diff - no controls and cluster-robust standard errors

model1 <- lm(waz ~ postXhigh + post + hightreat, datos = datos) coeftest(model1, vcov = vcovCL, cluster = ~idcommunity)

Same with heteroskedastic-robust standard errors

coeftest(model1, vcov = vcovHC)

D in D with fixed effects for community and individual controls

datos <- pdatos.frame(datos, index = "idcommunity")

model2 <- plm(waz ~ postXhigh + fedu + medu + hhsizep + lntotminc + immuniz + nonclinic + age, datos = datos, model = "within") coeftest(model2, vcov = vcovHC)

D in D with fixed effects for community and individual controls using lm

model3 <- lm(waz ~ postXhigh + post + hightreat + factor(idcommunity) + fedu + medu + hhsizep + lntotminc + immuniz + nonclinic, datos = datos) coeftest(model3, vcov = vcovCL, cluster = ~idcommunity)

Same results now if drop post and hightreat

model4 <- lm(waz ~ postXhigh + factor(idcommunity) + fedu + medu + hhsizep + lntotminc + immuniz + nonclinic, datos = datos) coeftest(model4, vcov = vcovCL, cluster = ~idcommunity)

Multiple periods of datos example from Stata documentation

Loading example datosset

datos(hospital, package = "fixest")

Describe and summarize the datos

glimpse(hospital) summary(hospital)

Difference-in-Differences with multiple periods using fixest

did_model <- feols(satis ~ procedure | hospital + month, datos = hospital, cluster = ~hospital) summary(did_model)

Heteroskedastic-robust standard errors

coeftest(did_model, vcov = vcovHC)