

## **Trabajo Práctico N° 3:** **Diferencias en Diferencias.**

### **Ejercicio 1: Evaluación de Impacto con DiD.**

Una práctica común en la evaluación de un programa cuando se tienen datos de panel para dos períodos es la siguiente: sea  $y_{it}$  el resultado observado para  $i$  en el período  $t$ . En  $t=1$ , nadie está en el programa. En  $t=2$ , algunos están en el grupo de control y otros en el grupo de tratamiento. Sea  $prog_{it}$  una variable que vale 1 si el individuo  $i$  está en el grupo de tratamiento en el período  $t$  y cero en caso contrario. Notar que  $prog_{i1}=0$  para todo  $i$ . Se puede plantear el siguiente modelo:

$$y_{it} = \theta_1 + \theta_2 d2_t + \delta_1 prog_{it} + c_i + u_{it},$$

con  $E(u_{it} / prog_{i2}, c_i) = 0$ . En el que  $d2_t$  es una variable dummy que vale 1 si  $t=2$ , 0 si  $t=1$ ;  $c_i$  es el efecto no observado. Usando el método de primeras diferencias, mostrar que  $\hat{\theta}_2 = \overline{\Delta y_t}$  y  $\hat{\delta}_1 = \overline{\Delta y_t} - \overline{\Delta y_c}$ , donde  $\overline{\Delta y_c}$  es el cambio promedio en  $y$  a lo largo de los dos períodos para el grupo con  $prog_{i2}=0$ , y  $\overline{\Delta y_t}$  es el cambio promedio en  $y$  a lo largo de los dos períodos para el grupo con  $prog_{i2}=1$ .

El modelo en  $t=1$  es:

$$y_{i1} = \theta_1 + \theta_2 * 0 + \delta_1 * 0 + c_i + u_{i1}$$

$$y_{i1} = \theta_1 + c_i + u_{i1}.$$

El modelo en  $t=2$  es:

$$y_{i2} = \theta_1 + \theta_2 * 1 + \delta_1 prog_{i2} + c_i + u_{i2}$$

$$y_{i2} = \theta_1 + \theta_2 + \delta_1 prog_{i2} + c_i + u_{i2}.$$

Tomando primeras diferencias, se tiene:

$$\Delta y_i = y_{i2} - y_{i1}$$

$$\Delta y_i = (\theta_1 + \theta_2 + \delta_1 prog_{i2} + c_i + u_{i2}) - (\theta_1 + c_i + u_{i1})$$

$$\Delta y_i = \theta_1 + \theta_2 + \delta_1 prog_{i2} + c_i + u_{i2} - \theta_1 - c_i - u_{i1}$$

$$\Delta y_i = \theta_2 + \delta_1 prog_{i2} + \Delta u_i.$$

Los estimadores OLS de  $\theta_2$  y  $\delta_1$  son:

$$\hat{\delta}_1 = \frac{Cov(prog_{i2}, \Delta y_i)}{Var(prog_{i2})}, \quad \hat{\theta}_2 = \overline{\Delta y} - \hat{\delta}_1 \overline{prog_{i2}}.$$

Sea  $N_1$  la cantidad de veces que  $prog_{i2}=1$ , se tiene:

$$Cov(prog_{i2}, \Delta y_i) = \sum_{i=1}^N prog_{i2} \Delta y_i - N \overline{prog_{i2}} \overline{\Delta y}$$

$$Cov(prog_{i2}, \Delta y_i) = \sum_{prog_{i2}=1} \Delta y_i - N \frac{\sum_{i=1}^N prog_{i2}}{N} \overline{\Delta y}$$

$$Cov(prog_{i2}, \Delta y_i) = N_1 \frac{\sum_{prog_{i2}=1} \Delta y_i}{N_1} - \sum_{i=1}^N prog_{i2} \overline{\Delta y}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = N_1 \overline{\Delta y_t} - N_1 \overline{\Delta y_c}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = N_1 (\overline{\Delta y_t} - \overline{\Delta y_c})$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = N_1 \left( \frac{\sum_{\text{prog}_{i2}=1} \Delta y_i}{N_1} - \frac{\sum_{i=1}^N \Delta y_i}{N} \right)$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = N_1 \frac{N \sum_{\text{prog}_{i2}=1} \Delta y_i - N_1 \sum_{i=1}^N \Delta y_i}{N_1 N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{N \sum_{\text{prog}_{i2}=1} \Delta y_i - N_1 \sum_{i=1}^N \Delta y_i}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{N \sum_{\text{prog}_{i2}=1} \Delta y_i - N_1 (\sum_{\text{prog}_{i2}=0} \Delta y_i + \sum_{\text{prog}_{i2}=1} \Delta y_i)}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{N \sum_{\text{prog}_{i2}=1} \Delta y_i - N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i - N_1 \sum_{\text{prog}_{i2}=1} \Delta y_i}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{(N - N_1) \sum_{\text{prog}_{i2}=1} \Delta y_i - N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{(N - N_1) N_1 \overline{\Delta y_t} - (N - N_1) N_1 \overline{\Delta y_c}}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = \frac{(N - N_1) N_1 (\overline{\Delta y_t} - \overline{\Delta y_c})}{N}$$

$$\text{Cov}(\text{prog}_{i2}, \Delta y_i) = (1 - \frac{N_1}{N}) N_1 (\overline{\Delta y_t} - \overline{\Delta y_c}).$$

$$\text{Var}(\text{prog}_{i2}) = \sum_{i=1}^N \text{prog}_{i2}^2 - N (\overline{\text{prog}_{i2}})^2$$

$$\text{Var}(\text{prog}_{i2}) = N_1 - N \left( \frac{N_1}{N} \right)^2$$

$$\text{Var}(\text{prog}_{i2}) = N_1 - N \frac{N_1^2}{N^2}$$

$$\text{Var}(\text{prog}_{i2}) = N_1 - \frac{N_1^2}{N}$$

$$\text{Var}(\text{prog}_{i2}) = (1 - \frac{N_1}{N}) N_1.$$

Por lo tanto, se tiene:

$$\hat{\delta}_1 = \frac{(1 - \frac{N_1}{N}) N_1 (\overline{\Delta y_t} - \overline{\Delta y_c})}{(1 - \frac{N_1}{N}) N_1}$$

$$\hat{\delta}_1 = \overline{\Delta y_t} - \overline{\Delta y_c}.$$

$$\hat{\theta}_2 = \overline{\Delta y} - (\overline{\Delta y_t} - \overline{\Delta y_c}) \overline{\text{prog}_{i2}}$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^N \Delta y_i}{N} - \frac{N_1}{N} \left( \frac{\sum_{\text{prog}_{i2}=1} \Delta y_i}{N_1} - \frac{\sum_{\text{prog}_{i2}=0} \Delta y_i}{N - N_1} \right)$$

$$\hat{\theta}_2 = \frac{\sum_{i=1}^N \Delta y_i}{N} - \frac{\sum_{\text{prog}_{i2}=1} \Delta y_i}{N} + N_1 \frac{\sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{(N - N_1) \sum_{i=1}^N \Delta y_i - (N - N_1) \sum_{\text{prog}_{i2}=1} \Delta y_i + N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{(N - N_1) (\sum_{\text{prog}_{i2}=0} \Delta y_i + \sum_{\text{prog}_{i2}=1} \Delta y_i) - (N - N_1) \sum_{\text{prog}_{i2}=1} \Delta y_i + N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{(N - N_1) \sum_{\text{prog}_{i2}=0} \Delta y_i + (N - N_1) \sum_{\text{prog}_{i2}=1} \Delta y_i - (N - N_1) \sum_{\text{prog}_{i2}=1} \Delta y_i + N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{(N - N_1) \sum_{\text{prog}_{i2}=0} \Delta y_i + N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{N \sum_{\text{prog}_{i2}=0} \Delta y_i - N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i + N_1 \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{N \sum_{\text{prog}_{i2}=0} \Delta y_i}{N(N - N_1)}$$

$$\hat{\theta}_2 = \frac{\sum_{\text{prog}_{i2}=0} \Delta y_i}{N - N_1}$$

$$\theta_2^{OLS} = \overline{\Delta y_c}.$$

**Ejercicio 2: DiD Simple en Stata.**

(a) Abrir la base “Panel101.dta” y generar las siguientes variables de resultado, de tiempo y de tratamiento. Los “países” tratados son los países 5 a 7 y el tratamiento se otorgó en 1994.

**Stata.**

(b) Computar e interpretar el estimador de diferencias en diferencias utilizando una regresión lineal.

Source	SS	df	MS	Number of obs	=	70
Model	51.8976208	3	17.2992069	F(3, 66)	=	1.98
Residual	575.397385	66	8.71814219	Prob > F	=	0.1249
Total	627.295005	69	9.09123196	R-squared	=	0.0827
				Adj R-squared	=	0.0410
				Root MSE	=	2.9527

  

Y	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
time	2.289455	.9529637	2.40	0.019	.3868012	4.192108
treated	1.77597	1.127562	1.58	0.120	-.4752802	4.02722
did	-2.519512	1.455676	-1.73	0.088	-5.425863	.3868395
_cons	.358144	.7381625	0.49	0.629	-1.115645	1.831933

(c) Computar e interpretar el estimador de diferencias en diferencias utilizando una especificación con efectos fijos.

Fixed-effects (within) regression	Number of obs	=	70
Group variable: country	Number of groups	=	7
R-squared:	Obs per group:		
Within = 0.0968	min =		10
Between = 0.0116	avg =		10.0
Overall = 0.0341	max =		10
corr(u_i, Xb) = -0.3880	F(2, 61)	=	3.27
	Prob > F	=	0.0448

Y	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
time	2.289455	.8986787	2.55	0.013	.4924355	4.086474
treated	0 (omitted)					
did	-2.519512	1.372754	-1.84	0.071	-5.264504	.2254806
_cons	1.119274	.5262124	2.13	0.037	.0670471	2.1715
sigma_u	1.6151513					
sigma_e	2.7844542					
rho	.25176021	(fraction of variance due to u_i)				
F test that all u_i=0: F(6, 61) = 2.67				Prob > F = 0.0230		

(d) *Computar e interpretar el estimador de diferencias en diferencias utilizando el paquete de Stata diff.*

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 70

	Before	After		
Control:	16	24	40	
Treated:	12	18	30	
	28	42		
-----				
Outcome var.	Y	S. Err.	t	P> t
-----				
Before				
Control	0.358			
Treated	2.134			
Diff (T-C)	1.776	1.128	1.58	0.120
After				
Control	2.648			
Treated	1.904			
Diff (T-C)	-0.744	0.921	0.81	0.422
Diff-in-Diff	-2.520	1.456	1.73	0.088*
-----				

R-square: 0.08

\* Means and Standard Errors are estimated by linear regression

\*\*Inference: \*\*\* p<0.01; \*\* p<0.05; \* p<0.1

### Ejercicio 3: Card & Krueger (1994).

Este ejercicio se basa en el artículo “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”.

(a) ¿Qué efecto intentan estimar los autores en el artículo?

(b) ¿Cuál es la estrategia de identificación?

(c) Utilizar el archivo “CardKrueger1994.dta”. Utilizando diff, computar el estimador de diferencias en diferencias.

#### DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 801

	Before	After		
Control:	78	77	155	
Treated:	326	320	646	
	404	397		
-----				
Outcome var.	fte	S. Err.	t	P> t
-----+-----+-----+-----+-----				
Before				
Control	19.949			
Treated	17.065			
Diff (T-C)	-2.884	1.135	-2.54	0.011**
After				
Control	17.542			
Treated	17.573			
Diff (T-C)	0.030	1.143	0.03	0.979
Diff-in-Diff	2.914	1.611	1.81	0.071*
-----				

R-square: 0.01

\* Means and Standard Errors are estimated by linear regression

\*\*Inference: \*\*\* p<0.01; \*\* p<0.05; \* p<0.1

(d) Repetir el inciso (c) utilizando errores estándar de bootstrap.

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 801

	Before	After	
Control:	78	77	155
Treated:	326	320	646
	404	397	

Bootstrapped Standard Errors

Outcome var.	fte	S. Err.	t	P> t
Before				
Control	19.949			
Treated	17.065			
Diff (T-C)	-2.884	1.329	-2.17	0.030**
After				
Control	17.542			
Treated	17.573			
Diff (T-C)	0.030	1.080	0.03	0.977
Diff-in-Diff	2.914	1.620	1.80	0.072*

R-square: 0.01

\* Means and Standard Errors are estimated by linear regression

\*\*Inference: \*\*\* p<0.01; \*\* p<0.05; \* p<0.1

(e) Repetir el inciso (c) utilizando la cadena de restaurantes como variables explicativas.

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 801

	Before	After	
Control:	78	77	155
Treated:	326	320	646
	404	397	

Report - Covariates and coefficients:

Variable(s)	Coeff.	Std. Err.	z	P> z
bk	0.917	1.060	0.865	0.387
kfc	-9.205	0.961	-9.575	0.000
roys	-0.897	1.136	-0.790	0.430

Bootstrapped Standard Errors

Outcome var.	fte	S. Err.	t	P> t
Before				
Control	21.161			
Treated	18.837			
Diff (T-C)	-2.324	1.276	-1.82	0.069*
After				
Control	18.758			
Treated	19.369			
Diff (T-C)	0.611	0.946	0.65	0.518
Diff-in-Diff	2.935	1.563	1.88	0.060*

R-square: 0.19

\* Means and Standard Errors are estimated by linear regression

\*\*Inference: \*\*\* p<0.01; \*\* p<0.05; \* p<0.1



**Ejercicio 4: didregress.**

Hasta ahora, se estimó utilizando la especificación:

$$Y_{it} = \beta_0 + \beta_1 D_{it} + \beta_2 T_{it} + \beta_3 (D \times T)_{it} + z_{it} \theta + u_{it} \quad (1)$$

O, en el caso de datos longitudinales, se amplía la ecuación con efectos fijos. En Stata, se usa los comandos `regress` y/o `xtreg`. Stata 17 incluyó nuevos comandos para estimar modelos con la siguiente forma:

$$Y_{ist} = \gamma_s + \gamma_t + z_{ist} \beta + \delta D_{st} + u_{ist}$$

con el comando `didregress` o incluyendo efectos fijos por individuo en el caso de datos longitudinales con el comando `didregress`.

**(a)** Explicar en qué difieren estos comandos con respecto al setup usual de DiD.

**(b)** ¿Se puede replicar las regresiones de los ejercicios anteriores con estos comandos?

**(c)** Un proveedor de salud está interesado en estudiar el efecto de un nuevo procedimiento de ingreso hospitalario en la satisfacción de los pacientes. El proveedor dispone de datos mensuales de pacientes de enero a julio. El nuevo procedimiento de admisiones fue implementado en abril por hospitales que estaban bajo nueva administración. De los 46 hospitales del estudio, 18 implementaron el nuevo procedimiento. El proveedor de salud utilizará una regresión DID para analizar el efecto del nuevo procedimiento de admisiones en los hospitales que participaron en el programa. El resultado de interés es la satisfacción del paciente, `satis`, que se registra como un promedio de las respuestas a un conjunto de cuatro preguntas realizadas a los pacientes. `satis` puede tomar valores entre 0 y 10, donde 10 es el mayor nivel de satisfacción posible y 0 es la decepción total. La variable `procedimiento` marca las observaciones tratadas; es 1 si una persona encuestada ingresó al hospital utilizando el nuevo procedimiento después de marzo y 0 en caso contrario. Los datos están en la base "hospdd.dta". Evaluar el impacto del nuevo procedimiento sobre la satisfacción de los pacientes.

Number of groups and treatment time

Time variable: month

Control: procedure = 0

Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression

Number of obs = 7,368

Data type: Repeated cross-sectional

(Std. err. adjusted for 46 clusters in hospital)

	Coefficient	Robust std. err.	t	P> t	[95% conf. interval]
satis					
ATET					
procedure					
(New vs Old)	.8479879	.0321121	26.41	0.000	.7833108 .912665

Note: ATET estimate adjusted for covariates, group effects, and time effects.

**(d)** ¿Cómo se interpreta el coeficiente obtenido? ¿Se cumple el supuesto de tendencias paralelas?

El coeficiente obtenido de la regresión DID (0,8479879) representa el efecto promedio del nuevo procedimiento de admisión hospitalaria sobre la satisfacción de los pacientes en los hospitales que implementaron el nuevo procedimiento, en comparación con aquellos que no lo hicieron. Específicamente, se estima que el nuevo procedimiento incrementa la satisfacción de los pacientes en, aproximadamente, 0,85 puntos en la escala de 0 a 10.

El supuesto de tendencias paralelas es crucial para la validez de la estimación del modelo DID. Este supuesto implica que, en ausencia del tratamiento (nuevo procedimiento), las tendencias de la satisfacción de los pacientes habrían sido similares en los hospitales tratados y no tratados. Sin embargo, la evaluación de este supuesto no se puede realizar directamente a partir de los coeficientes del modelo; se requiere un análisis gráfico o pruebas formales previas.

Para verificar si este supuesto se cumple, se podría graficar las tendencias de la satisfacción en ambos grupos (tratados y no tratados) antes de la implementación del nuevo procedimiento (es decir, de enero a marzo) y observar si las tendencias son similares. Si las tendencias son paralelas antes de la implementación, entonces, es razonable suponer que el supuesto se cumple.

Si los datos muestran que las tendencias de satisfacción entre los grupos tratados y no tratados divergen antes de abril, entonces, el supuesto de tendencias paralelas no se cumple y la interpretación causal del coeficiente podría ser sesgada. En tal caso, se deberían considerar ajustes adicionales o utilizar métodos alternativos para analizar el efecto del nuevo procedimiento.

(e) Comentar sobre los errores estándar utilizados y estudiar las distintas opciones que el comando tiene pre-configuradas para usar. ¿Hay diferencias en la inferencia?

### Errores estándar robustos:

Number of groups and treatment time

Time variable: month

Control: procedure = 0

Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression  
Data type: Repeated cross-sectional

Number of obs = 7,368

	Coefficient	Robust std. err.	z	P> z	[95% conf. interval]
satis					
ATET					
procedure					
(New vs Old)	.8479879	.0353158	24.01	0.000	.7787701 .9172056

Note: ATET estimate adjusted for covariates, group effects, and time effects.

### Errores estándar hc2:

Number of groups and treatment time

Time variable: month

Control: procedure = 0

Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression

Number of obs = 7,368

Data type: Repeated cross-sectional

No. of clusters = 46

	Coefficient	Robust HC2 std. err.	t	P> t	[95% conf. interval]
satis					
ATET					
procedure					
(New vs Old)	.8479879	.0325552	26.05	0.000	.7819941 .9139816

Note: ATET estimate adjusted for covariates, group effects, and time effects.

### Errores estándar bootstrap:

Number of groups and treatment time

Time variable: month

Control: procedure = 0

Treatment: procedure = 1

	Control	Treatment
Group		
hospital	28	18
Time		
Minimum	1	4
Maximum	1	4

Difference-in-differences regression

Number of obs = 7,368

Replications = 50

Wald chi2(1) = 888.21

Prob > chi2 = 0.0000

Data type: Repeated cross-sectional

(Replications based on 46 clusters in hospital)

	Observed coefficient	Bootstrap std. err.	z	P> z	Normal-based [95% conf. interval]
satis					
ATET					
procedure					
(New vs Old)	.8479879	.0284533	29.80	0.000	.7922204 .9037553

Note: ATET estimate adjusted for covariates, group effects, and time effects.

Por lo tanto, se puede observar que no hay grandes diferencias en la inferencia, ya que el ATET es estadísticamente significativo al 1%.