Clase auxiliar 6

Aplicaciones de probabilidades y estadística en gestión

Departamento de Ingeniería Civil Industrial
Universidad de Chile

Ronald Leblebici Garo

16 de octubre de 2017

Antes de imprimir esta presentación, piense si es realmente necesario.

¿Qué vimos la clase pasada?

- Método LASSO (selección de modelos)
- Experimentos y sesgo de selección

¿Qué vimos la clase pasada?

Método LASSO

lugar	porcentaje	prediccion	residuo	residuo2
Ciudad Férrica	0,66	0,504814977	0,155185023	0,024082391
Ciudad Azuliza	0,63	0,475792931	0,154207069	0,02377982
Ciudad Malvalona	0,76	0,514192747	0,245807253	0,060421205
Ciudad Lavacal	0,6	0,276917708	0,323082292	0,104382167
Ciudad Petalia	0,43	0,208032856	0,221967144	0,049269413
Ciudad Arbora	0,71	0,525404661	0,184595339	0,034075439
Ciudad Algaria	0,65	0,550047419	0,099952581	0,009990518
Ciudad Polis	0,58	0,57639732	0,00360268	1,29793E-05
Ciudad Plateada	0,66	0,609719187	0,050280813	0,00252816
Ciudad Celeste	0,59	0,349129405	0,240870595	0,058018644
Ciudad Carmín	0,59	0,728144673	-0,138144673	0,019083951
Ciudad Azulona	0,83	0,560769637	0,269230363	0,072484988
Ciudad Fucsia	0,6	0,349729468	0,250270532	0,062635339
Ciudad Azafrán	0,76	0,681708659	0,078291341	0,006129534
Ciudad Canela	0,63	0,48969744	0,14030256	0,019684808
Ciudad Verde	0,71	0,631516706	0,078483294	0,006159627
Ciudad Malva	0,59	0,340196037	0,249803963	0,06240202
Ciudad Azalea	0,6	0,322156084	0,277843916	0,077197242
Ciudad Trigal	0,61	0,861559835	-0,251559835	0,063282351
Ciudad Iris	0,68	0,771447052	-0,091447052	0,008362563
Ciudad Orquídea	0,85	0,798486406	0,051513594	0,00265365
Ciudad Olivo	0,65	0,771481193	-0,121481193	0,01475768

β_0	β_local	β_desempleo	β_desempleo2
0,00150909	0,000150352	0,000304953	0,000141975

|β_0|	|β_local|	|β_desempleo|	|β_desempleo2|
0,00150909	0,000150352	0,000304953	0,000141975

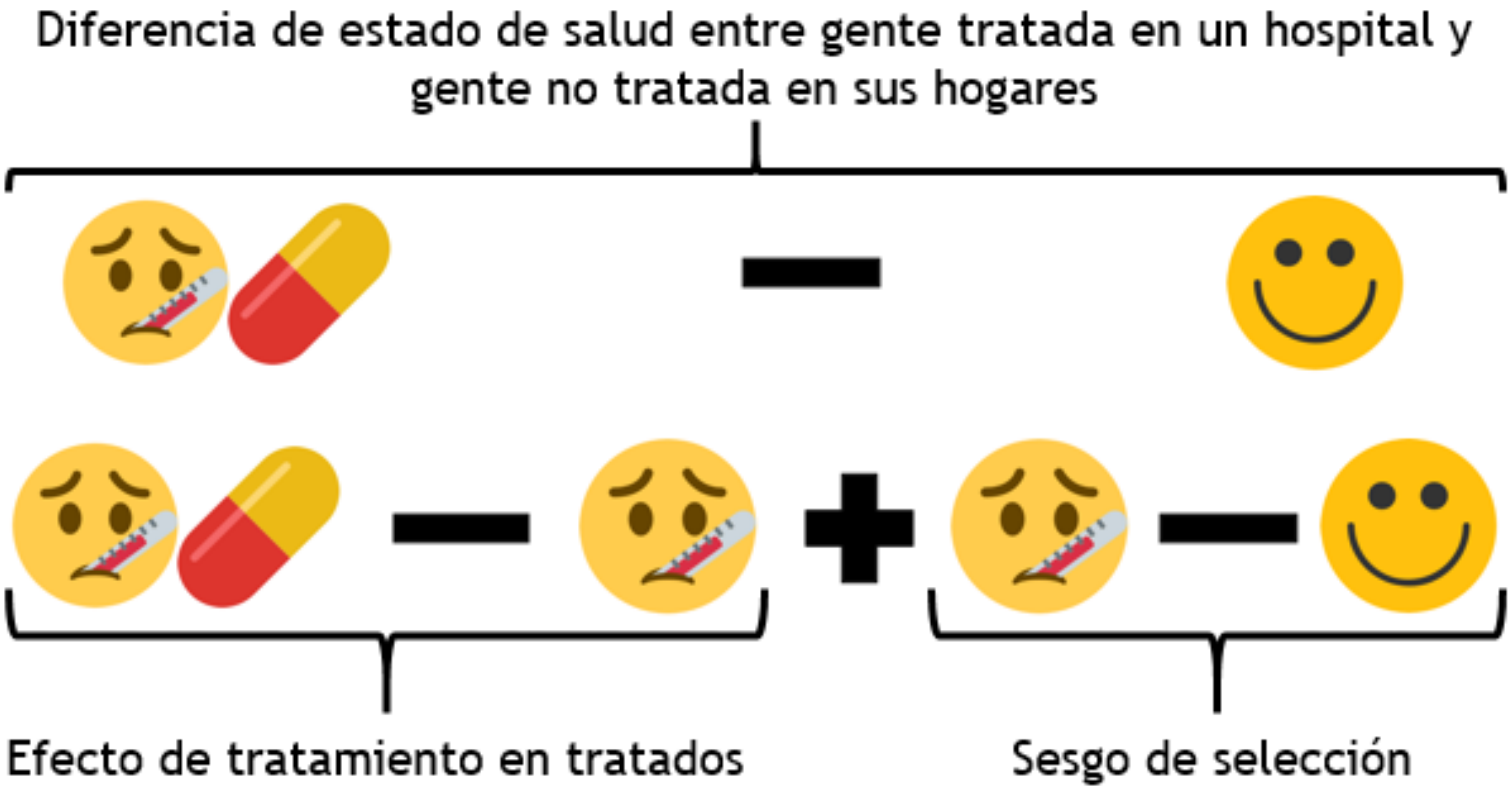
$$\sum_{n=1}^{N}(Y_n - \widehat{\beta}_0 - \sum_{k=1}^{K}\widehat{\beta}_k X_{nk})^2 + \lambda \sum_{k=1}^{K}\left|\widehat{\beta}_k\right|$$

$$= \quad 32{,}51825 \quad + \quad 0{,}01$$

$$= \quad 32{,}52453$$

¿Qué vimos la clase pasada?

Experimentos y sesgo de selección



¿Qué veremos hoy?

- Introducción a datos de corte transversal, series de tiempo y datos de panel
- Dif-in-Dif
- Test Reset de Ramsey (selección de modelos)

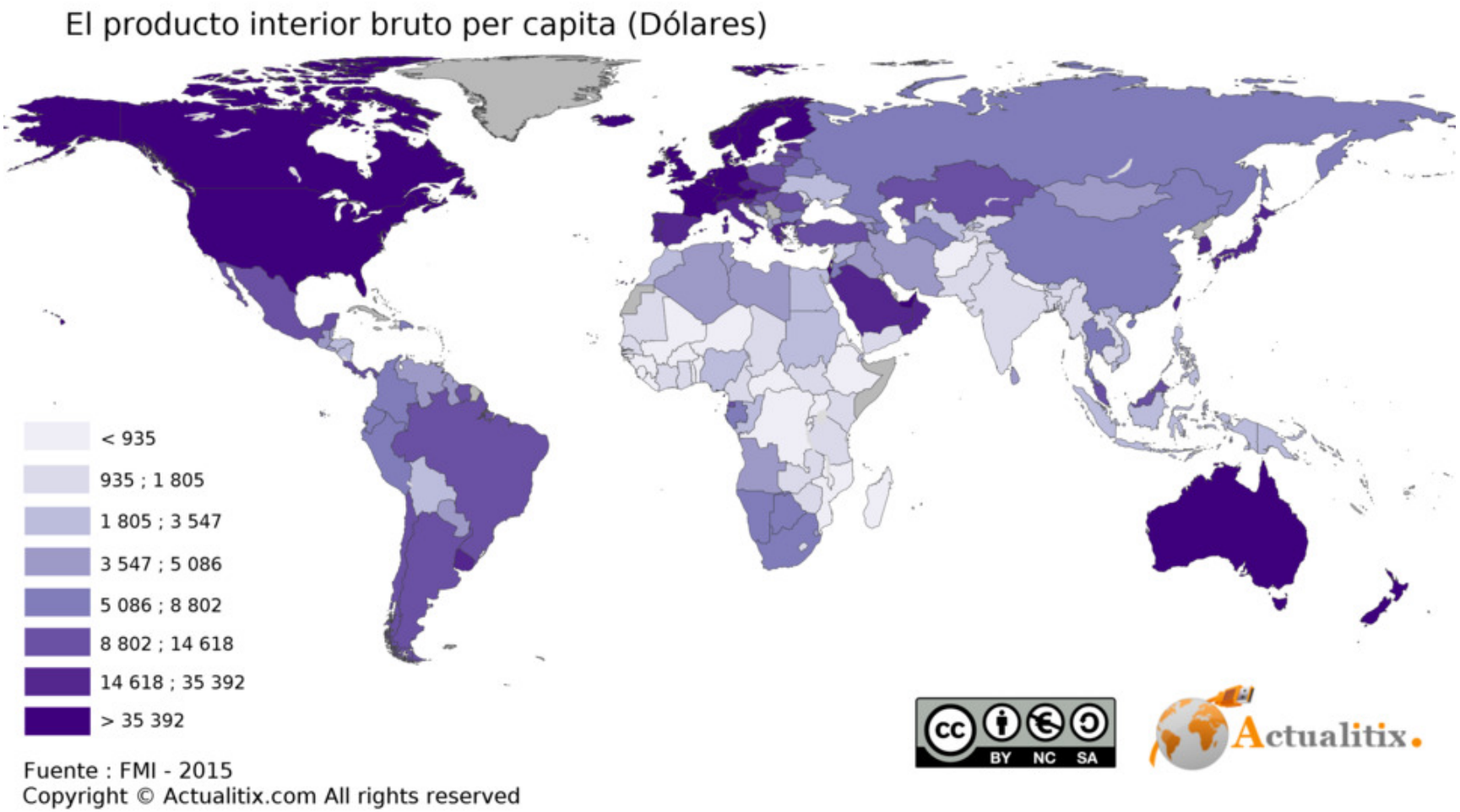
Datos de corte transversal

Contienen información de distintos **individuos**.

	edad	escolaridad	ln_ingreso	genero
1	15	10	.	1
2	45	8	13.33133	1
3	55	8	12.69645	1
4	10	.	.	1
5	19	15	.	0
6	10	.	.	0
7	23	12	.	1
8	30	5	12.11355	1
9	22	14	.	0
10	10	.	.	0
11	33	12	12.52906	1
12	63	6	.	1
13	3	.	.	1
14	67	3	10.91962	0

Datos de corte transversal

Contienen información de distintos **individuos**.



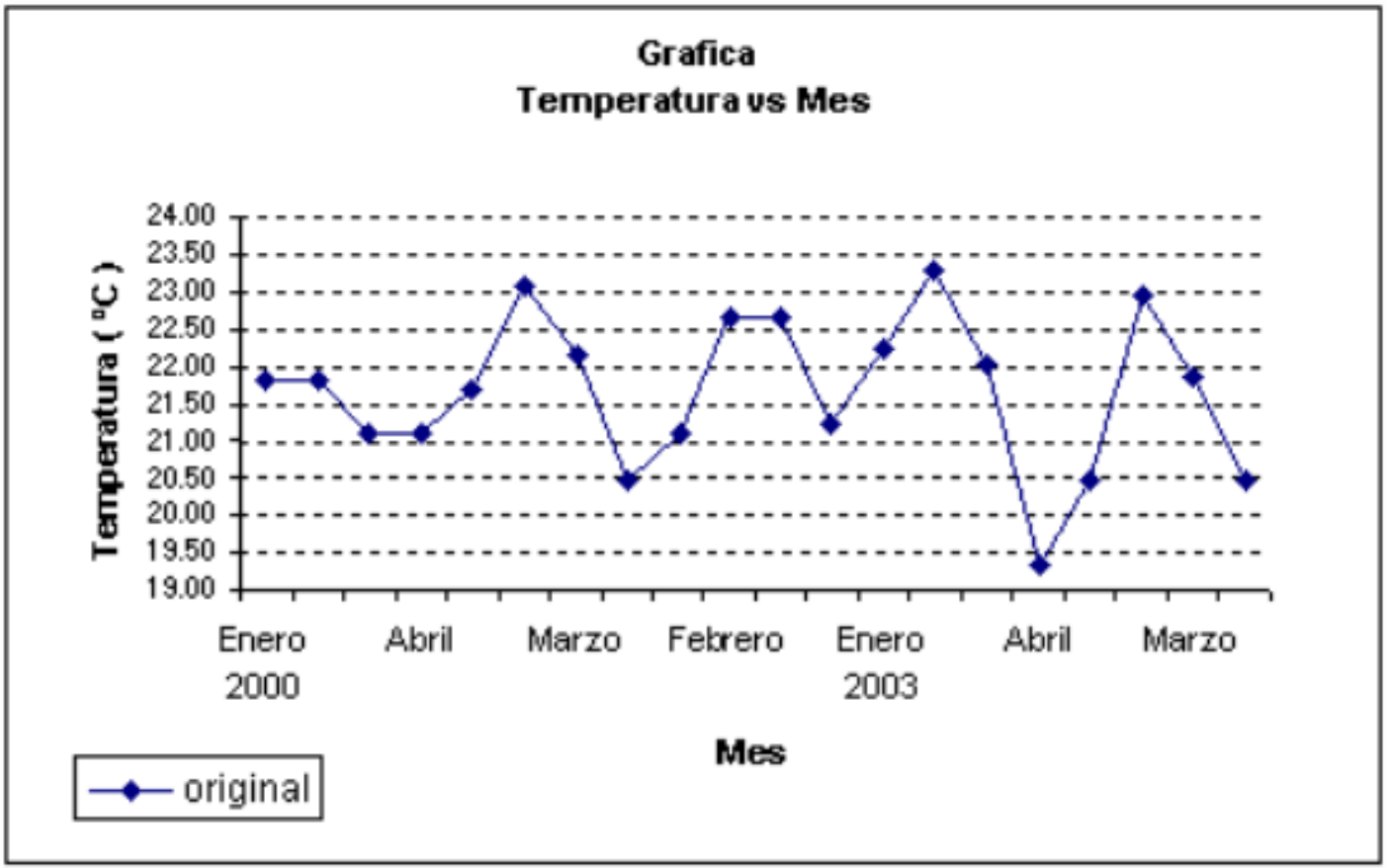
Series de tiempo

Contienen información de **un individuo** en distintos **periodos**.

Semana	Ventas
1	17
2	21
3	19
4	23
5	18
6	16
7	20
8	18
9	22
10	20
11	15
12	22

Series de tiempo

Contienen información de **un individuo** en distintos **periodos**.



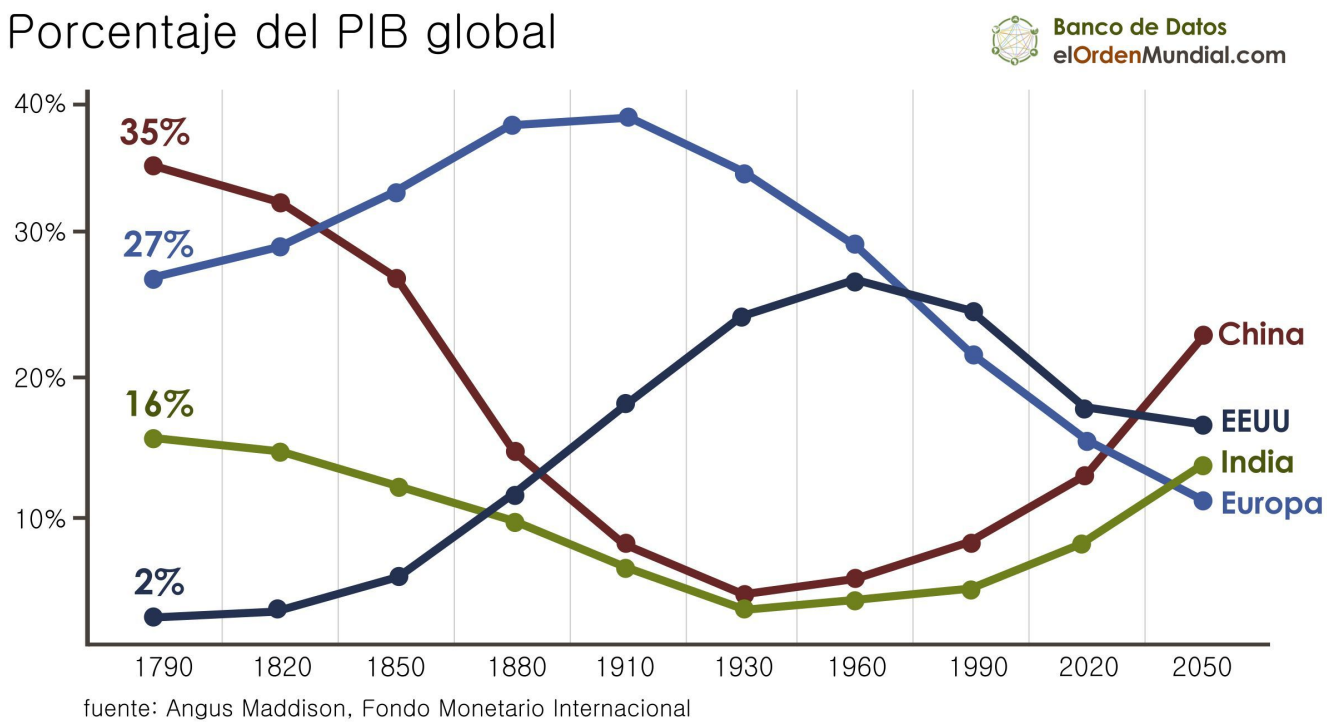
Datos de panel

Contienen información de distintos **individuos** en distintos **periodos**.

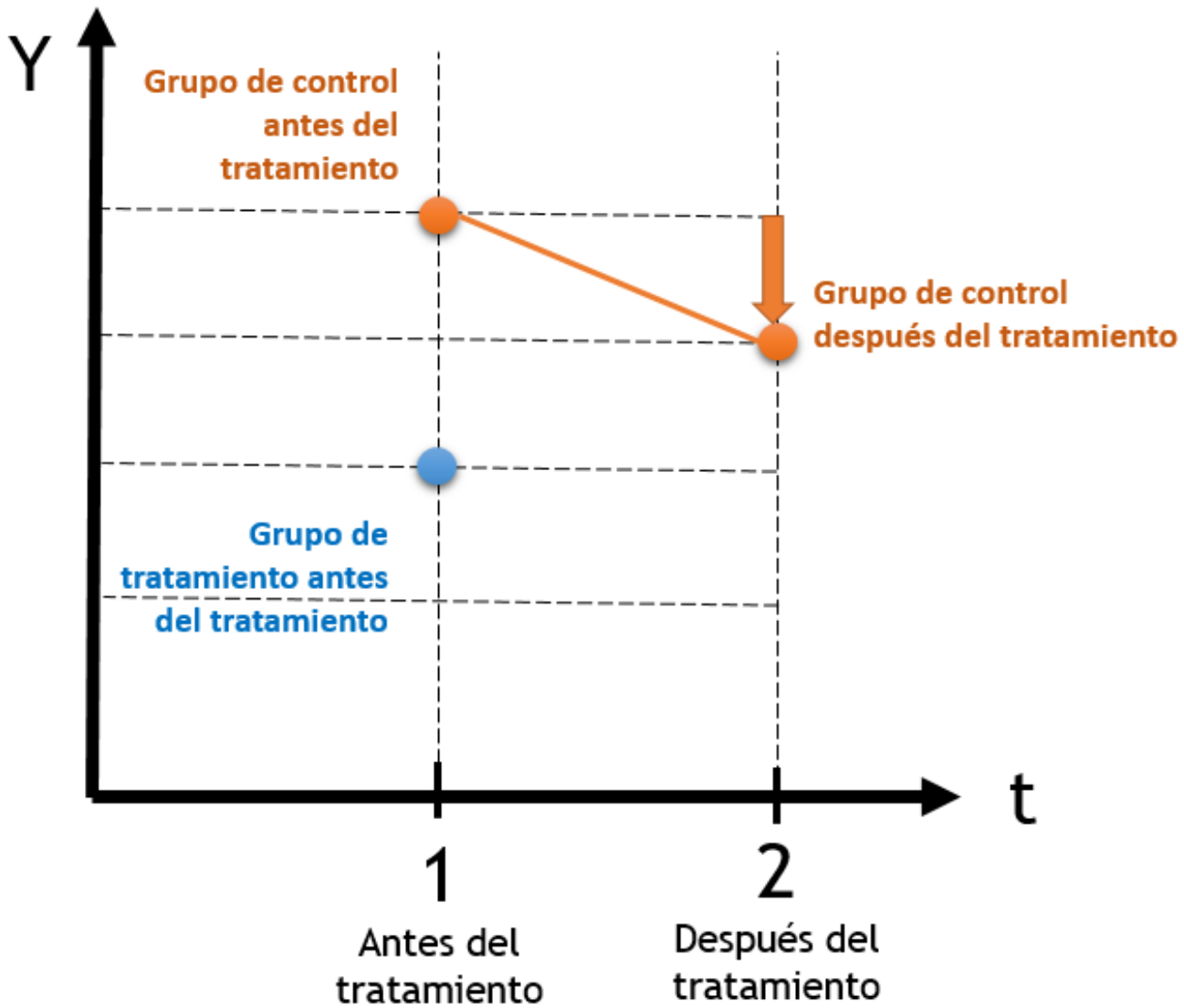
individuo	*periodo*	*ingresos*	*edad*	*sexo*
1	2003	1500	27	1
1	2004	1700	28	1
1	2005	2000	29	1
2	2003	2100	41	2
2	2004	2100	42	2
2	2005	2200	43	2

Datos de panel

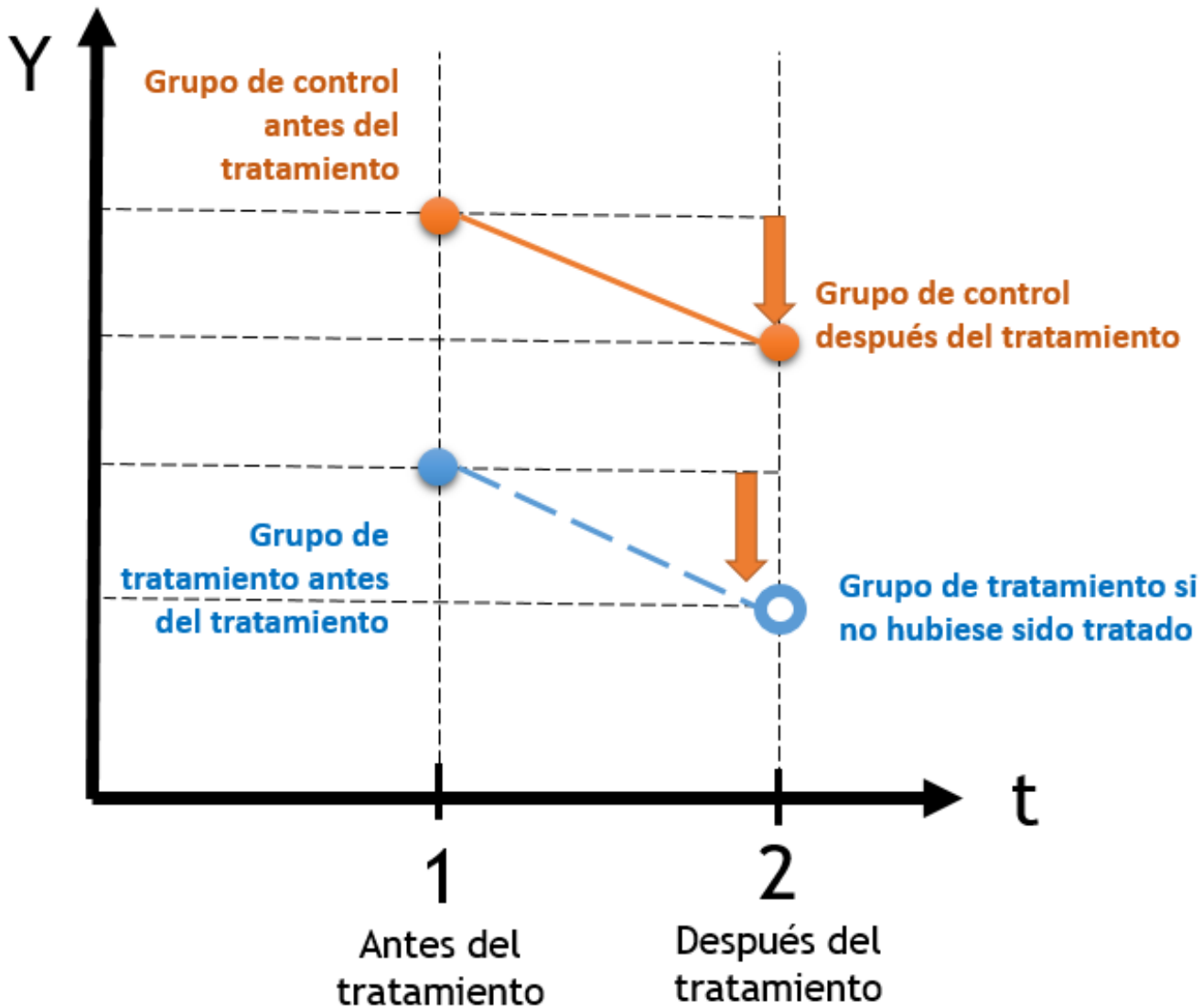
Contienen información de distintos **individuos** en distintos **periodos**.



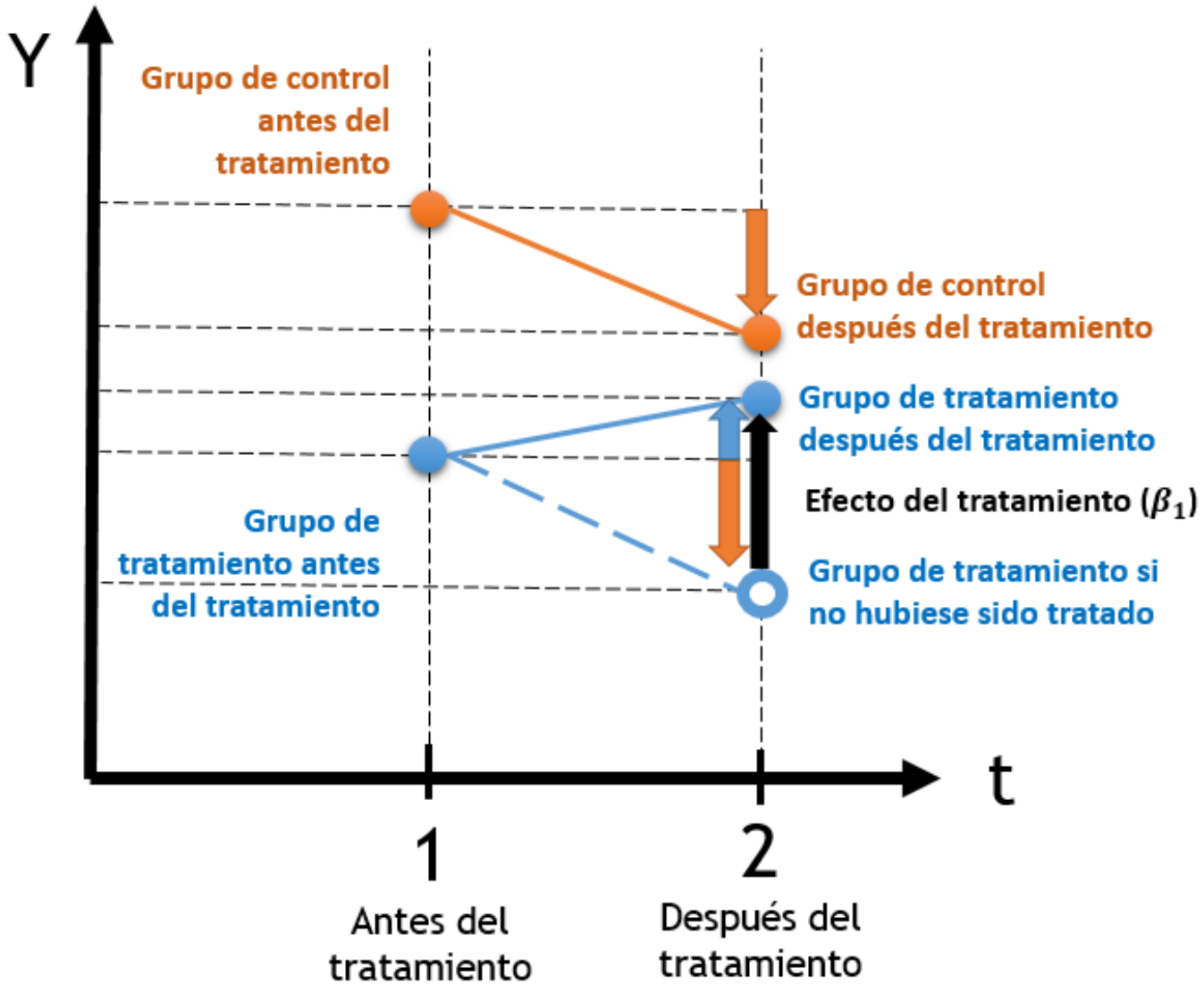
Dif-in-Dif


Y
Grupo de control antes del tratamiento
Grupo de control después del tratamiento
Grupo de tratamiento antes del tratamiento
t
1
2
Antes del tratamiento
Después del tratamiento

Dif-in-Dif


Y
Grupo de control antes del tratamiento
Grupo de control después del tratamiento
Grupo de tratamiento antes del tratamiento
Grupo de tratamiento si no hubiese sido tratado
t
1
Antes del tratamiento
2
Después del tratamiento

Dif-in-Dif


Y
Grupo de control antes del tratamiento
Grupo de control después del tratamiento
Grupo de tratamiento después del tratamiento
Efecto del tratamiento (β_1)
Grupo de tratamiento antes del tratamiento
Grupo de tratamiento si no hubiese sido tratado
t
1
Antes del tratamiento
2
Después del tratamiento

Dif-in-Dif

El efecto β_1 de la diapositiva anterior corresponde al que se muestra en el siguiente modelo:

$$Y_{nt} = \beta_0 + \beta_1 D_{nt} + \alpha_n + \delta_t + \epsilon_{nt}$$

Donde D_{nt} es una variable dummy que indica si es que el individuo n en el momento t fue sometido al tratamiento o no.

Después de un poco de álgebra y supuestos, se obtiene el efecto del tratamiento y su estimación:

$$\beta_1 = \mathbb{E}(\Delta Y_{n2} | D_{n2} = 1) - \mathbb{E}(\Delta Y_{n2} | D_{n2} = 0)$$

Dif-in-Dif

Por Ley de los Grandes Números, sabemos que los promedios son estimadores consistentes de las esperanzas.

$$\beta_1 = \mathbb{E}(\Delta Y_{n2}|D_{n2} = 1) - \mathbb{E}(\Delta Y_{n2}|D_{n2} = 0)$$

$$\hat{\beta}_1 = [\bar{Y}_{2,D_{n2}=1} - \bar{Y}_{1,D_{n2}=1}] - [\bar{Y}_{2,D_{n2}=0} - \bar{Y}_{1,D_{n2}=0}]$$

Como podemos ver, nuestro estimador no es más que una diferencia de diferencias. De ahí viene el nombre de *dif-in-dif*.

P1.1: Dif-in-Dif

	Tratamiento	Control
t=1	6	5,9
t=2	6,3	5,7

$$\hat{\beta}_1 = \underbrace{\underbrace{[\underbrace{\bar{Y}_{2,D_{n2}=1}}_{\square} - \underbrace{\bar{Y}_{1,D_{n2}=1}}_{\square}]}_{\square} - \underbrace{[\underbrace{\bar{Y}_{2,D_{n2}=0}}_{\square} - \underbrace{\bar{Y}_{1,D_{n2}=0}}_{\square}]}_{\square}}_{\square}$$

P1.1: Dif-in-Dif

	Tratamiento	Control
t=1	6	5,9
t=2	6,3	5,7

$$\hat{\beta}_1 = \underbrace{\underbrace{[\underbrace{\bar{Y}_{2,D_{n2}=1}}_{\boxed{6,3}} - \underbrace{\bar{Y}_{1,D_{n2}=1}}_{\boxed{6,0}}]}_{\boxed{0,3}} - \underbrace{[\underbrace{\bar{Y}_{2,D_{n2}=0}}_{\boxed{5,7}} - \underbrace{\bar{Y}_{1,D_{n2}=0}}_{\boxed{5,9}}]}_{\boxed{-0,2}}}_{\boxed{0,5}}$$

P1.2: Dif-in-Dif

Para esta pregunta, utilizaremos la base de datos de la Universidad de Princeton contenida en la página:
`http://dss.princeton.edu/training/Panel101.dta`.

En dicha base se tiene información sobre una variable `y`, para años entre 1990 y 1999 y ciudades enumeradas desde el 1 hasta el 7. Nos interesa conocer el efecto sobre la variable `y` de un tratamiento que se llevó a cabo desde 1994 para las ciudades 5, 6 y 7.

1. Plantee un modelo de regresión que permita estimar el efecto del tratamiento.
2. Ejecute esta regresión en *Stata* y estime el efecto del tratamiento β_{did}.
3. Instale la extensión `diff` y estime nuevamente el efecto de tratamiento mediante este comando. Compare los resultados con los de la parte anterior.

Test Reset de Ramsey

Por simplicidad consideremos el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + U_i$$

Supongamos que queremos saber si las combinaciones no lineales de X_{i1} y X_{i2} tienen poder explicativo sobre Y_i.

¿A qué nos referimos con combinaciones no lineales?

En este caso: X_{i1}^2, X_{i2}^2 y $X_{i1}X_{i2}$.

Si estas variables resultan ser importantes y no las consideramos, tendremos un problema de omisión de variables relevantes, lo cual no es deseable porque $\hat{\beta}_{MCO}$ se vuelve sesgado e inconsistente.

Test Reset de Ramsey

Claramente $\hat{Y}_i$ es función de $Xi1$ y X_{i2}:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$$

Si elevamos la ecuación al cuadrado, veremos que $\hat{Y}_i^2$ es de la siguiente forma:

$$\hat{Y}_i^2 = a + b \cdot X_{i1} + c \cdot X_{i2} + d \cdot X_{i1}^2 + e \cdot X_{i2}^2 + f \cdot X_{i1} \cdot X_{i2}$$

Ahora podemos ver que si agregamos $\hat{Y}_i^2$ como variable explicativa en nuestra regresión original, estaremos agregando implícitamente X_{i1}^2, X_{i2}^2 y $X_{i1}X_{i2}$.

Luego, el modelo propuesto para el test es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_2 \hat{Y}_i^2 + V_i$$

Si $\hat{\gamma}_2$ es estadísticamente significativo, es una señal de que alguno(s) de los términos contenidos en $\hat{Y}_i^2$ podría serlo también.

Test Reset de Ramsey

Hemos visto que podemos testear la significancia de factores cuadráticos (grado 2).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_2 \hat{Y}_i^2 + V_i$$

Podemos generalizar el mismo razonamiento para factores de grado P.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_2 \hat{Y}_i^2 + ... + \gamma_P \hat{Y}_i^P + V_i$$

Sin más preámbulo, mostremos cómo se realiza el test...

Test Reset de Ramsey

1. Especificar $Y = X\beta + U$ y estimar $\hat{\beta}_{MCO}$ e $\hat{Y}$.

```
reg y x_1 [...] x_K
predict y_estimado
```

2. Construir $\hat{Y}^p$ para $p \in \{1, ..., P\}$

```
gen y_estimado_2 = y_estimado ^ 2
[...]
gen y_estimado_P = y_estimado ^ P
```

3. Regresionar $Y = X\beta + \gamma_2 \hat{Y}^2 + ... + \gamma_P \hat{Y}^P + V$

```
reg y x_1 [...] x_K y_estimado_2 [...] y_estimado_P
```

4. Testear $H_0 : \gamma_2 = ... = \gamma_P = 0$ (restricciones múltiples).

```
test (y_estimado_2 = 0) [...] (y_estimado_P = 0)
```

P2: Test Reset de Ramsey

En la base de datos `Auxiliar 06 Test Reset Ramsey BD.dta` se encuentran las variables `y`, `x` y `z`.

Considere el modelo $y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + u_i$.

1. ¿Qué modelo plantearía para realizar el test Reset de Ramsey con potencias $p = 2, 3$?
2. Plantee un test de hipótesis (hipótesis nula, hipótesis alternativa, estadístico y región de rechazo) para dicho test.
3. Realice dicho test en *Stata*. ¿Qué puede concluir?

Resumen de la clase

- Existen distintos tipos de bases de datos: de **corte transversal** (muchos individuos en un periodo), **series de tiempo** (un individuo en muchos periodos) y **datos de panel** (muchos individuos en muchos periodos).

- Los **datos de panel** pueden ser útiles para aplicar el método **Dif-in-Dif**, el cual permite estimar el efecto de un tratamiento que se aplica para una parte de la muestra en un intervalo de tiempo.

- **Dif-in-dif en múltiples periodos** nos permite controlar heterogeneidades entre individuos que no varían en el tiempo.

- El **test Reset de Ramsey** permite evaluar si es "correcta" la forma funcional de nuestro modelo. Nos da una señal de si es que son significativas –en su conjunto– algunas combinaciones no lineales de los regresores ya considerados.