Causalidad y regresion

Walter Sosa Escudero
wsosa@udesa.edu.ar
19 de octubre de 2023

Preliminares: regresion en una dummy

$$Y_i = \alpha + \beta D_i + u_i, \quad i = i, \ldots, N$$

Notacion

- T = total de observaciones con $D_i = 1$ ('tratados')
- $N - T$ = 'no tratados'.
- $\bar{Y}_T$, $\bar{Y}_{N-T}$, promedios tratados y no tratados.

Resultado: $\hat{\beta} = \bar{Y}_T - \bar{Y}_{N-T}$

Demostracion: ver Apendice a esta clase.

Efecto y causa

$$Y_i = \alpha + \beta D_i + u_i$$

Paraguas, lluvia. Fertilizante, altura. AUH, asiste al secundario.

- En que sentido β mide el efecto que D tiene sobre Y?
- En que sentido $\hat{\beta}$ en base a $(D_i, Y_i), i = 1, \ldots, n$ estima el efecto que D tiene sobre Y?

Todavia no tenemos una definicion clara de causa.

Causa y efecto en base a observables

- $D = 0, 1$, 'causa', 'tratamiento'. Notacion $D_1 \equiv (D = 1)$, $D_0 \equiv (D = 0)$.
- Y es un resultado ('efecto').
- $Y|D_1 =$ resultado observable si hubo tratamiento. $Y|D_0$ si no hubo tratamiento.

Tratados y no tratados

Resulta tentador pensar que el efecto causal es la diferencia entre 'tratados y no tradados':

$$Y|D_1 - Y|D_0$$

Ej: comparar personas que hicieron / no hicieron dieta, recibieron o no la AUH, estudian o no.

Problema?

Antes y despues

Tampoco funciona comparar 'antes y despues'

$$Y|D_1 - Y|D_0$$

- Peso antes y despues de hacer dieta.

Problema?

Causa y resultados potenciales

- Cuestion filosofica delicada. Aproximacion simple.
- Resultados potenciales: Y_{1i}, Y_{0i} independientemente de si hubo o no tratamiento.
- Ej: Y_{1i} nota si *estudiases*. Son 'promesas'. Y_{0i} ingreso si no *recibieses* la AUH.
- Efecto causal: $\beta = Y_{1i} - Y_{0i}$.

Causalidad en terminos de diferencias entre resultados potenciales.

Estas variables 'existen' mas alla del tratamiento.

Inobservabilidad de contrafacticos

- Problema: para una persona i, se observa Y_{1i} o Y_{0i} *pero nunca ambos*.
- D implica haber eliminado una ruta observable. Ambas rutas potenciales 'existen'.
- '*El tiempo se bifurca perpetuamente hacia innumerables futuros. En uno de ellos soy su enemigo*'. (J.L. Borges, en 'El jardin de senderos que se bifurcan)

`Esa trama de tiempos que se aproximan, se bifurcan, se cortan o que secularmente se ignoran, abarca todas la posibilidades. No existimos en la mayoría de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En éste, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravesar el jardín, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma'

Tratamiento independiente (aleatorizado)

- Recordar: Si Z y D son independientes, $E(Z|D) = E(Z)$.
- $\tau_i = Y_{1i} - Y_{0i}$, *efecto causal*.
- Si Y_{1i} y Y_{0i}, τ_i fuesen observable, τ_i es trivialmente estimable. Borges.

¿Es posible estimar τ_i con la informacion observable, Y_i?

Llamemos $\tau = E(\tau_i) = E(Y_{1i} - Y_{0i})$, efecto causal promedio.

Resultado: *cuando D es independiente de Y_1 y Y_0, es posible estimar τ en base a Y.*

$$\begin{aligned} \tau = E(Y_{1i} - Y_{01}) &= E(Y_{1i}) - E(Y_{0i}) \\ &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \qquad \text{(¿por que?)} \\ &= E(Y_i|D_i = 1) - E(Y_i|D_i = 0) \end{aligned}$$

$$\tau = E(Y_{1i} - Y_{01}) = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

Entonces:

$$\hat{\tau} = \bar{Y}_T - \bar{Y}_{N-T}$$

O, alternativamente, $\hat{\tau}$ es el coeficiente de regresion en:

$$Y_i = \alpha + \tau D_i + u_i$$

Estimable: el efecto causal (en base a inobservables, Y_{1i}, Y_{0i}) se puede estimar en base a magnitudes observables (Y_i).

	Potencial		Observable			
	y0	y1	D	y0	y1	y
1	5.7	6.7	0	5.7		5.7
2	3.2	4.2	1		4.2	4.2
3	3.8	4.8	1		4.8	4.8
4	9.3	10.3	1		10.3	10.3
5	9.3	10.3	0	9.3		9.3
6	1.5	2.5	1		2.5	2.5
7	.3	1.3	0	.3		.3
8	6.7	7.7	1		7.7	7.7
9	4.7	5.7	0	4.7		4.7
10	6.1	7.1	1		7.1	7.1

Promedio	5.0	6.1
Efecto causal estimado		1.1
Efecto causal		1

- Tratamiento aletorio: independiente de los resultados potenciales.
- $E(Y_{1i}|D_i) = E(Y_{1i})$, $E(Y_{0i}|D_i) = E(Y_{0i})$
- Ejemplos?

Regresion, correlacion y causalidad

Idea: tratamiento independiente implica que en:

$$Y_i = \alpha + \tau D_i + u_i$$

$E(u_i|D_i) = 0$.

Notar que $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})\, D_i$

Supondremos $Y_{1i} - Y_{0i} = \tau + \nu_i$, ν_i independiente de D_i y $E(\nu_i) = 0$

τ es el efecto causal promedio.

$$
\begin{aligned}
Y_i &= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
&= Y_{0i} + (\tau + \nu_i) D_i \\
&= E(Y_{0i}) + (\tau + \nu_i) D_i + Y_{0i} - E(Y_{0i}) \\
&= E(Y_{0i}) + \tau D_i + Y_{0i} - E(Y_{0i}) + \nu_i D_i \\
Y_i &= \alpha + \tau D_i + u_i
\end{aligned}
$$

con $u_i \equiv Y_{0i} - E(Y_{01}) + \nu_i D_i$

$$Y_i = \alpha + \tau D_i + u_i, \qquad u_i \equiv Y_{0i} - E(Y_{01}) + \nu_i D_i$$

Notar que $E(\nu_i D_i | D_i) = D_i E(\nu_i | D_i) = D_i E(\nu_i) = 0$. Entonces:

$$\begin{aligned} E(Y_i | D_i = 1) &= \alpha + \tau + E(Y_{0i} | D_i = 1) - E(Y_{01}) \\ E(Y_i | D_i = 0) &= \alpha + E(Y_{0i} | D_i = 0) - E(Y_{01}) \end{aligned}$$

$$E(Y_i | D_i = 1) - E(Y_i | D_i = 0) = \tau + E(Y_{0i} | D_i = 1) - E(Y_{0i} | D_i = 0))$$

$$\underbrace{E(Y_i|D_i=1) - E(Y_i|D_i=0)}_{\text{Correlacion}} = \underbrace{\tau}_{\text{Causalidad}} + \underbrace{E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0))}_{\text{Sesgo de seleccion}}$$

Bajo tratamiento aleatorio:

$$\begin{aligned} \text{Sesgo} &= E(Y_{0i}|D_i=1) - E(Y_{0i}|D_i=0)) \\ &= E(Y_{0i}) - E(Y_{0i}) \qquad (\text{por?}) \\ &= 0 \end{aligned}$$

O, alternativamente, en

$$Y_i = \alpha + \tau D_i + u_i$$

$$\begin{aligned} E(u_i|D_i) &= E\left(Y_{0i} - E(Y_{0i}) + \nu_i D_i|\, D_i\right) \\ &= E(Y_{0i}) - E(Y_{0i)} - E(\nu_i D_i|D_i) \\ &= 0 \end{aligned}$$

Idea: bajo tratamiento aleatorio, el coeficiente de regresar Y en D admite una interpretacion causal.

Tratamiento aleatorio?

- Tratamiento aleatorio: eleccion de tratamiento sin mirar resultados potenciales.
- Experimento o cuasi experimento.
- D se mueve en forma exogena ('causa').
- Datos observacionales: la gente no hace dieta porque si, ni toma aspirinas al azar sino porque inicialmente tenia fiebre.
- Auge de la aproximacion experimental en medicina. Economia?
- Experimento: control de la variabilidad exogena.

- Lluvia y paraguas con datos observacionales: $\beta = Y_1 - Y_2$, obviamente cero. Puro sesgo: $Y|D_1 - Y|D_0 = S$
- Lluvia y paraguas en un experimento: β estima correctamente $\beta = 0$ (sesgo nulo).

Resumiendo

- Causalidad: relacion entre potenciales. Uno no es observable.
- Bajo aleatorizacion de tratamiento, $Y = \alpha + \beta D + u$ tiene una interpretacion causal. $\hat{\beta}$ es insesgado.
- Rol de $E(u|D) = 0$: D varia en forma exogena.
- Relevancia del razonamiento experimental.
- Cuestion muy importante en las ciencias sociales en los ultimos tiempos.

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019

© Nobel Media. Photo: A. Mahmoud
Abhijit Banerjee
Prize share: 1/3

© Nobel Media. Photo: A. Mahmoud
Esther Duflo
Prize share: 1/3

© Nobel Media. Photo: A. Mahmoud
Michael Kremer
Prize share: 1/3

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2019 was awarded jointly to Abhijit Banerjee, Esther Duflo and Michael Kremer "for their experimental approach to alleviating global poverty"

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

© Nobel Prize Outreach. Photo: Clement Morin
David Card
Prize share: 1/2

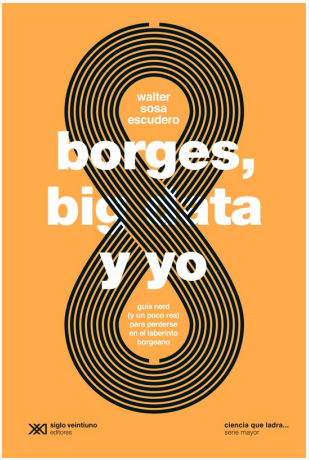
© Nobel Prize Outreach. Photo: Stefan Bladh
Joshua D. Angrist
Prize share: 1/4

© Nobel Prize Outreach. Photo: Stefan Bladh
Guido W. Imbens
Prize share: 1/4

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 was divided, one half awarded to David Card "for his empirical contributions to labour economics", the other half jointly to Joshua D. Angrist and Guido W. Imbens "for their methodological contributions to the analysis of causal relationships"

Referencias

- Angrist, J. y Pischke, J., 2014, *Mastering Metrics: the Path from Cause to Effect*, Cap. 2, Princeton University Press, Princeton.
- Sosa Escudero, W., 2022, *Que es (y que no es) la Estadistica*, Siglo XXI Editores, Buenos Aires. Capitulo 3: El huevo y la gallina: causalidades y casualidades.
- Sosa Escudero, W., 2020, *Borges, big data y yo, Siglo XXI Editores*, Buenos Aires.
- Borges, J.L., 1944, El jardin de senderos que se bifurcan, en *Ficciones*, Sudamericana, Buenos Aires.
- Hernán M. y Robins J., 2020, *Causal Inference: What If*. Boca Raton: Chapman-Hall/CRC.


walter
sosa
escudero
borges,
y yo
guía nerd
(y un poco real)
para perderse
en el laberinto
borgeano
siglo veintiuno
editores
ciencia que ladra...
serie mayor

Apendice: $\hat{\beta}$ como diferencia de medias

$$Y_i = \alpha + \beta D_i + u_i, \quad i = i, \ldots, N$$

Notacion

- $T =$ tratados, $N - T =$ no tratados.
- $\bar{Y}_T$, $\bar{Y}_{N-T}$, promedios tratados y no tratados.
- $\sum_T Y_i \equiv \sum D_i Y_i, \quad \sum_{N-T} \equiv \sum (1 - D) Y_i$

Resultado: $\hat{\beta} = \bar{Y}_T - \bar{Y}_{N-T}$

Prueba

Recordar

$$\hat{\beta} = \frac{\sum d_i Y_i}{\sum d_i^2}, \quad d_i \equiv D_i - \bar{D}$$

Denominador:

$$\begin{aligned}
\sum d_i^2 &= \sum (D_i - \bar{D})^2 \\
&= \sum D_i^2 - N\bar{D}^2 \\
&= \sum D_i - N\ T^2/N^2 \\
&= T - T^2/N \\
&= T\ (1 - T/N)
\end{aligned}$$

Numerador:

$$\begin{aligned}
\sum d_i Y_i &= \sum (D_i - \bar{D}) Y_i \\
&= \sum D_i Y_i - \bar{D} \sum Y_i \\
&= \sum_T Y_i - T/N \left(\sum_T Y_i + \sum_{N-T} Y_i \right) \\
&= T\ \bar{Y}_T - T/N \Big(T\ \bar{Y}_T + (N-T)\ \bar{Y}_{N-T} \Big) \\
&= \bar{Y}_T\ T(1 - T/N)\ -\ \bar{Y}_{N-T}\ T(1 - T/N) \\
&= T(1 - T/N)\ \left(\bar{Y}_T - \bar{Y}_{N-T} \right)
\end{aligned}$$

Reemplazando y simplificando se obtiene el resultado.

Ejercicio: derivar $\hat{\alpha}$ para este caso.