Regresion, correlacion y causalidad

Walter Sosa Escudero

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.

Problema?

Antes y despues

Tampoco funciona comparar 'antes y despues'

$$Y|D_1 - Y|D_0$$

- Peso antes y despues de hacer dieta.
- Nuevamente, comparacion de peras y manzanas.
- Ceteris paribus?

Causa y efecto en base a contrafacticos

- Cuestion filosofica delicada. Aproximacion simple.
- Resultados potenciales.

$$\begin{array}{ll} Y_0 & \text{si } D = 0 \\ Y_1 & \text{si } D = 1 \end{array}$$

  independientemente de si hubo o no tratamiento.
- Ej: Y_1 temperatura si *tomases* un analgesico. Son 'promesas'. Y_0 salario si no *recibieses* la AUH
- Efecto causal: $\beta = Y_1 - Y_0$ (caida en la fiebre si tomases una aspirina con respecto a que no la tomes).

Causalidad en terminos de diferencias entre resultados potenciales (contrafacticos)

Inobservabilidad de contrafacticos

- Problema: se observa Y_1 o Y_0 *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)

Observables

En la practica se observa Y

$$Y = \begin{cases} Y_1 & \text{si } D = 1 \\ Y_0 & \text{si } D = 0 \end{cases}$$

O, alternativamente:

$$Y = Y_0 + (Y_1 - Y_0)\, D$$

Inobservancia de contrafacticos: Si a una persona le di una droga, observo la temperatura de la persona habiendole dado la droga, pero no veo a la misma persona en la circunstancia de no haberle dado la droga. Y viceversa!

Regresion y causalidad

- El problema de medir el efecto causal parece no tener solucion (inobservabilidad de contrafactuales)
- El modelo de regresion $Y_i = \alpha + \beta D_i + u_i$ estimado por MCO tiene un problema (sesgo por seleccion) y una solucion (aleatorizacion)

Sesgo de seleccion

- Notacion $D_1 \equiv (D = 1), \;\; D_0 \equiv (D = 0)$
- Comparacion personas tratadas y no tratadas

$$Y \mid D_1 \; - \; Y \mid D_0$$

  Verbalizacion: peso de gente que hizo dieta con gente que no hizo dieta.
- Problema? (peras con manzanas)

$$\begin{aligned} Y|D_1 - Y|D_0 &= \left[Y|D_1 - Y_0|D_1\right] + \left[Y_0|D_1 - Y|D_0\right] \\ &= \left[Y_1|D_1 - Y_0|D_1\right] + \left[Y_0|D_1 - Y_0|D_0\right] \end{aligned}$$

$$Y|D_1 - Y|D_0 = \beta + S$$

con $S \equiv Y_0|D_1 - Y_0|D_0$

S es el *sesgo por seleccion*.

$$\underbrace{Y|D_1 - Y|D_0}_{\text{Dif Observables}} = \underbrace{\beta}_{\text{Efecto causal}} + \underbrace{S}_{\text{Sesgo}}$$

Sesgo:

$$S \equiv Y_0|D_1 - Y_0|D_0$$

- Diferencia en peso potencial sin tratamiento, entre tratados y no tratados.
- En la practica? Quien hace dieta / toma analgesicos?
- Con datos observacionales $S \neq 0$.
- Sesgo de seleccion: la comparacion entre tratados y no tratados estima el efecto causal MAS el sesgo.

Correlacion no es causalidad

$$\begin{array}{ccccc} Y|D_1 - Y|D_0 & = & \beta & + & S \\ \text{Correlacion} & = & \text{Causalidad} & + & \text{Sesgo} \end{array}$$

Ej: paraguas y lluvia en datos observacionales: correlacion positiva, causalidad nula. Puro sesgo.

Aleatorizacion al rescate

Tratamiento aleatorio: D es independiente de Y_1 y Y_0

$$\begin{aligned} Y|D_1 - Y|D_0 &= \beta + \big[Y_0|D_1 - Y_0|D_0\big] \\ E\big[Y|D_1 - Y|D_0\big] &= \beta + E\big[Y_0|D_1\big] - E\big[Y_0|D_0\big] \\ &= \beta + E\big[Y_0|D_1\big] - E\big[Y_0|D_1\big] \\ &= \beta \end{aligned}$$

El paso clave es que bajo tratamiento aleatorio $E\big[Y_0|D_1\big] = E\big[Y_0|D_0\big]$

Resultado: *el tratamiento aleatorio elimina el sesgo de seleccion*.

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).

Aleatorizacion y exogeneidad

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

Que implica la aleatorizacion en el modelo causal en terminos del modelo lineal bajo los supuestos clasicos?

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= E(Y_0) + \beta D + \left[Y_0 - E(Y_0)\right] \\ Y &= \alpha + \beta D + u \end{aligned}$$

con $\alpha \equiv E(Y_0)$ y $u \equiv Y_0 - E(Y_0)$

- Supongamos que tenemos una muestra $(Y_i, D_i), i = 1, \ldots, n$
- Para que $\hat{\beta}$ sea insesgado necesitamos $E(u_i|D_i) = 0$.

$$
\begin{aligned}
E(u_i|D_i) &= E\left[Y_0 - E(Y_0) \mid D_i\right] \\
&= E(Y_0|D_i) - E(Y_0) \\
&= E(Y_0) - E(Y_0) \\
&= 0,
\end{aligned}
$$

ya que bajo aleaotorizacion $E(Y_0) = E(Y_0|D_i)$, de modo que $\hat{\beta}$ en base a datos observables es insesgado para el efecto causal.

Conclusion: *Bajo aleatorizacion de tratamiento,* $Y = \alpha + \beta D + u$ *tiene una interpretacion causal.* $\hat{\beta}$ *es insesgado para los datos observacionales (no hace falta ver los potenciales).*

Resumiendo

- Causalidad: relacion entre contrafacticos. 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.

Referencias

- Angrist, J. y Pischke, J., 2014, *Mastering Metrics: the Path from Cause to Effect*, Cap. 2, Princeton University Press, Princeton.
- Sosa Escudero, W., 2014, *Que es (y que no es) la Estadistica*, Siglo XXI Editores, Buenos Aires. Capitulo 3: El huevo y la gallina: causalidades y casualidades.
- Borges, J.L., 1944, El jardin de senderos que se bifurcan, en *Ficciones*, Sudamericana, Buenos Aires.

"A diferencia de Newton y de Schopenhauer, su antepasado no creia en un tiempo uniforme, absoluto. Creia en infinitas series de tiempos, en una red creciente y vertiginosa de tiempos divergentes, convergentes y paralelos. Esa trama de tiempos que se aproximan, se bifurcan, se cortan o que secularmente se ignoran, abarca todas la posibilidades. No existimos en la mayoria de esos tiempos; en algunos existe usted y no yo; en otros, yo, no usted; en otros, los dos. En este, que un favorable azar me depara, usted ha llegado a mi casa; en otro, usted, al atravezar el jardn, me ha encontrado muerto; en otro, yo digo estas mismas palabras, pero soy un error, un fantasma."

J.L. Borges, 1944, El jardin de senderos que se bifurcan

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.