

Intervención en causalidad

Matías Reyes

Facultad de Ciencias
Udelar

22 de Setiembre de 2022

- **Formula de ajuste:**

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

- **Regla 1** (La regla del efecto causal):

Dado un grafo G el efecto causal de X sobre Y esta dado por

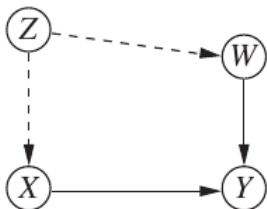
$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z)$$

Tambien tenemos la forma

$$P(y|do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)}$$

El factor $P(X = x|PA = z)$ se le llama *propensity score*

- **Back-Door Criterion (BDC):** En un *DAG* G decimos que un conjunto de variables Z satisface el *BDC* relativo a un par ordenado de variables (X, Y) si Z no contiene descendientes de X y Z bloquea todo camino entre X y Y que contenga una flecha entrante a X

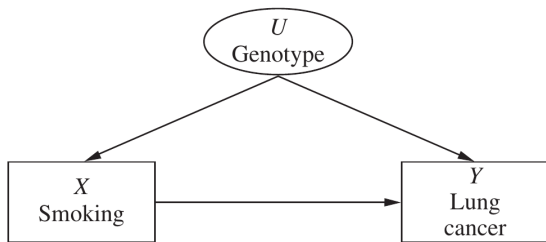


Ej.: Tratamos de calcular el efecto de consumir una droga X sobre el porcentaje de recuperación de una enfermedad Y donde también medimos el peso de la persona W y además sabemos que el estado socioeconómico afecta a W y X

¿Qué pasa cuando no tenemos un conjunto de variables que cumple el BDC?

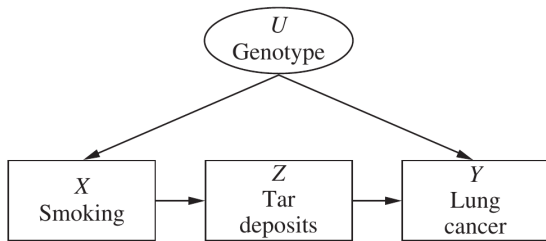
Podemos utilizar el operador *do* para identificar efectos causales sobre conjuntos de variables con una configuración mas compleja. por ejemplo un conjunto de variables que cumplan el **Front-Door Criterion**

Ejemplo introductorio: Se trabaja bajo la hipótesis de que existe un genotipo cancerígeno U que a su vez induce una adicción a la nicotina.



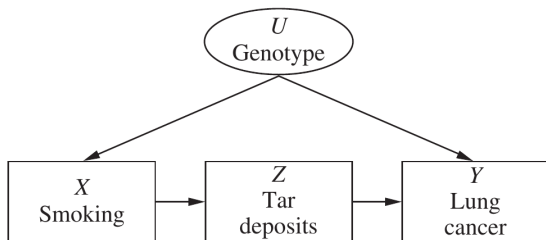
(a)

Asumiendo que la variable U no puede ser medida tenemos que en el modelo (a) no podemos usar *BDC* ya que el camino $X \leftarrow U \rightarrow Y$ no puede ser bloqueado, eso es, el efecto causal de fumar sobre el cáncer de pulmón no es identificable en este modelo



(b)

En cambio considerando el modelo modificado (b) podemos llegar mas lejos. En esta caso tenemos una variable más Z la cual mide si la persona tiene depósitos de alquitrán en los pulmones. De todas formas no tenemos forma de bloquear el camino $X \leftarrow U \rightarrow Y$, pero tenemos una forma de identificar el efecto causal de X sobre Y (eso es $P(Y = y|do(X = x))$) mediante el uso repetido de *BDC*.



(b)

Primero analicemos los datos. De una selección al azar de 800.000 personas las cuales tienen un alto riesgo de padecer de cáncer se mide

- 1 Si fuma
- 2 Cantidad de alquitrán en los pulmones
- 3 Si se le detecto cáncer de pulmón

Los resultados son presentados en la siguiente tabla

| | Tar 400 | | No tar 400 | | All subjects 800 | |
|-----------|--------------|------------|---------------|-------------|---------------------|---------------|
| | Smokers | Nonsmokers | Smokers | Nonsmokers | Smokers | Nonsmokers |
| No cancer | 380 | 20 | 20 | 380 | 400 | 400 |
| | 323 (85%) | 1 (5%) | 18 (90%) | 38 (10%) | 341 (85%) | 39 (9.75%) |
| Cancer | 57 | 19 | 2 | 342 | 59 | 361 |
| | (15%) | (95%) | (10%) | (90%) | (15%) | (90.25%) |

Estos datos han sido interpretados de dos maneras diferentes, resultando en inferencias contrarias. Uno puede discutir que los datos prueban los beneficios de fumar apuntando a que solo el 15% de los fumadores desarrollaron cáncer de pulmón comparado con el 92,25% de las personas que no fuman. Es más dentro de los subconjuntos de Tar y No tar los fumadores tienen un menor porcentaje de cáncer de pulmón.

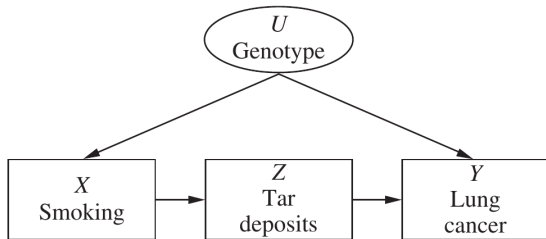
Front-Door Criterion

| | Smokers 400 | | Nonsmokers 400 | | All subjects 800 | |
|-----------|----------------|-------------|-------------------|-------------|---------------------|-------------|
| | Tar | No tar | Tar | No tar | Tar | No tar |
| No cancer | 380 | 20 | 20 | 380 | 400 | 400 |
| | 323 (85%) | 18 (90%) | 1 (5%) | 38 (10%) | 324 (81%) | 56 (19%) |
| Cancer | 57 | 2 | 19 | 342 | 76 | 344 |
| | (15%) | (10%) | (95%) | (90%) | (9%) | (81%) |

Por otro lado ordenando los datos como en la tabla de arriba podemos discutir que fumar aumenta el riesgo de cáncer de pulmón, esto es porque si uno elige fumar entonces la chance de que se generen depósitos de alquitrán en los pulmones es del 95% comparado con el 5% de los no fumadores. Luego fijándonos en la tabla vemos que los depósitos de alquitrán son perjudiciales para ambos fumadores y no fumadores.

Así que sin importar si existe una tendencia natural por fumar inducida por U uno debería evitar la acumulación de alquitrán en los pulmones y no fumar es una forma muy efectiva de evitar dicha acumulación.

Volviendo al modelo (b)



(b)

Pensando en lo que acabamos de hacer en términos del modelo (b) tenemos que primero analizamos en efecto causal de X en Z el cual es identificable porque no hay ninguna puerta trasera de X a Z

$$P(Z = z | do(X = x)) = P(Z = z | X = x)$$

Luego observemos que el efecto de Z en Y también es identificable ya que la única puerta trasera puede ser bloqueada condicionando en X

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x)P(X = x)$$

Por ultimo y razonando que si de elije asignar $Z = z$ entonces la probabilidad de Y sería $P(Y = y|do(Z = z))$. Pero la probabilidad de que asignar ese valor a Z dado que elegimos asignar $X = x$ es $P(Z = z|do(X = x))$. Luego sumando sobre todos los valores posibles de Z tenemos

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))P(Z = z|do(X = x))$$

Sustituyendo con las ecuaciones anteriores tenemos la *front-door formula*

$$P(Y = y|do(X = x))$$

||

$$\sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x)$$

Generalizando para configuraciones donde hay mas que un camino de X a Y

Definicion: (Front-Door Criterion) Se dice que un conjunto de variables Z satisface el criterio de la puerta delantera (*FDC*) relativo a un par ordenado de variables (X, Y) si

- 1 Z intercepta a todos los caminos directos de X a Y
- 2 Todos los caminos de X a Z están bloqueados
- 3 Todas las puertas traseras de Z a Y están bloqueadas por X

Teorema:(Front-Door Adjustment) Si Z satisface *FDC* relativo a (X, Y) y si $P(x, z) > 0$ entonces el efecto causal de X en Y es identificable y esta dado por la formula

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x')$$

Las intervenciones que consideramos hasta ahora estuvieron limitadas a forzar un grupo de variables X a tomar cierto valor x , en general las intervenciones hacen que un grupo de variables X responda a otro conjunto de variables Z de alguna manera específica por ejemplo por una relación funcional $x=g(z)$.

Ej.: Supongamos que un doctor decide administrar una droga solo a los pacientes con una temperatura Z mayor a $Z = z$, en este caso la acción va a estar condicionada por el valor de Z lo que se escribe como $do(X = g(Z))$ donde $g(Z) = 1$ si $Z > z$ y $g(Z) = 0$ de otro modo (donde $X = 0$ significa que no se administra la droga). Entonces el valor tomado por X va a ser una variable aleatoria que a su vez depende de Z . Luego tenemos que $P(Y = y | do(X = g(Z)))$ depende solamente de Z y g

Conditional Interventions and Covariate-Specific Effects

Para estimar el efecto causal de este tipo de relaciones veremos el concepto de *z-specific effect* de X , esto es $P(Y = y|do(X = x), Z = z)$ lo que representa o mide a la distribución de Y en una subpoblación de Z donde $Z = z$ luego de la intervención. Razonando de una manera similar a *BDC* obtenemos

Regla 2: El efecto z-especifico $P(Y = y|do(X = x), Z = z)$ es identificable cuando podemos medir un conjunto de variables S tal que $S \cup Z$ satisfacen *BDC*. En este caso el efecto z-especifico esta dado por

$$P(Y = y|do(X = x), Z = z) \\ \parallel \\ \sum_s P(Y = y|X = x, S = s, Z = z)P(S = s)$$

Ahora estamos en condiciones de calcular $P(Y = y|do(X = g(Z)))$. Condicionando $Z = z$ tenemos

$$\begin{aligned} & P(Y = y|do(X = g(Z))) \\ = & \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z))) \\ & = \sum_z P(Y = y|do(X = g(z)), Z = z)P(Z = z) \\ & = \sum_z P(Y = y|do(X = x), Z = z)|_{x=g(Z)}P(Z = z) \end{aligned}$$

Todos nuestros procedimientos de intervención nos dicen que es posible obtener resultados teóricos a partir de datos observacionales. Lo único que tenemos que hacer es identificar un conjunto de covariables Z que satisfagan algún criterio y usar una fórmula pertinente.

En la práctica ajustar Z puede ser difícil (tamaño de dimensiones, filtrar valores y tamaño de población). Para evadir trataremos de obtener una aproximación de $g(x, z) = P(X = x, Z = z)$, esto es el propensity score. Esto se puede hacer ajustando parámetros de una función g lo suficientemente flexible a nuestro conjunto de observaciones.

Luego de tener esa g podemos generar muestras de la probabilidad post intervención P_m mediante la forma especial de la formula de ajuste mostrada en 3.3

$$P(y|do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)}$$

Esto ya lo hicimos para la paradoja de Simpson