Introducción a modelos gráficos probabilísticos y al razonamiento causal

Irving Gómez Méndez

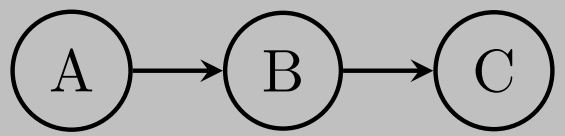
Enero, 2024

Teoría de grafos

Grafos

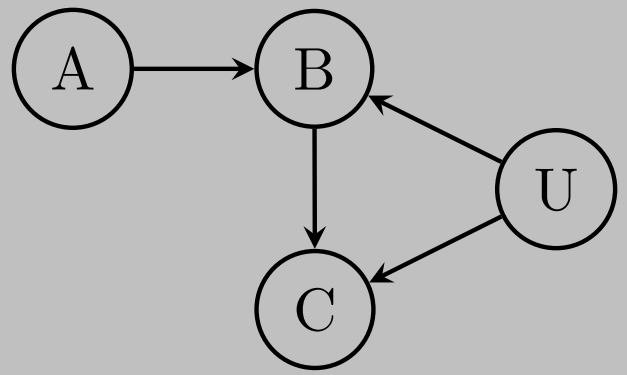
Un grafo $\mathcal{G}$ consiste de vértices (nodos) y de aristas (conexiones) entre los vértices. Las conecciones pueden ser dirigidas (tiene una flecha en una dirección) o no dirigidas. A las aristas también les podemos asociar un peso. Un grafo con todas sus aristas dirigidas se llama un grafo dirigido, y uno con todas sus aristas no dirigidas lo llamamos un grafo no dirigido.

Por ejemplo, podemos tener tres vértices: $\{A, B, C\}$, y un conjunto de dos aristas dirigidas entre estos tres vértices: $\{(A, B), (B, C)\}$. Por lo que, $\mathcal{G} = (\{A, B, C\}, \{(A, B), (B, C)\})$. Generalmente representamos los grafos mediante figuras:



Caminos

Un camino entre el nodo A y el nodo B es una secuencia de vértices $A_0 = A, A_1, \ldots, A_n = B$, con (A_k, A_{k+1}) una arista en el grafo, conectando A y B.



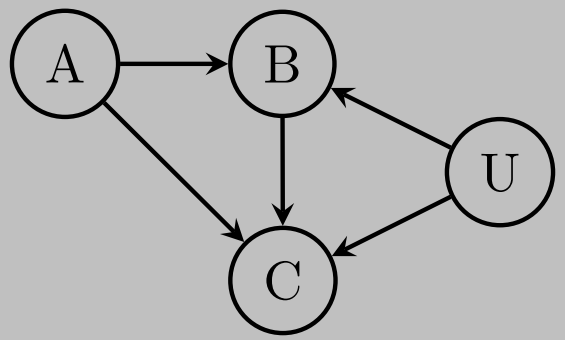
Por ejemplo, en el grafo anterior los caminos que conectan A y C son:

1. $A \to B \to C$.
2. $A \to B \leftarrow U \to C$.

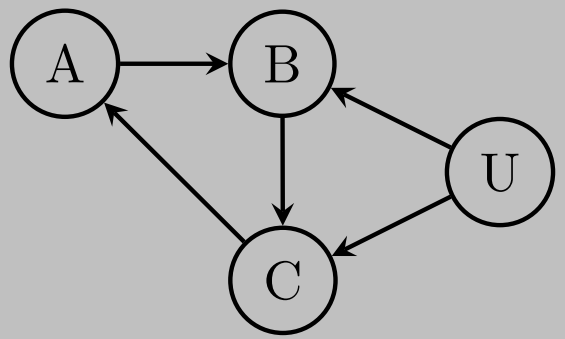
Grafos dirigidos acíclicos (DAGs)

Un grafo dirigido acíclico (DAG) es un grafo dirigido $\mathcal{G}$ tal que siguiendo cualquier camino de vértices respetando la dirección de las aristas, no existe ningún camino que regrese a un vértice anterior.

Este es un DAG:

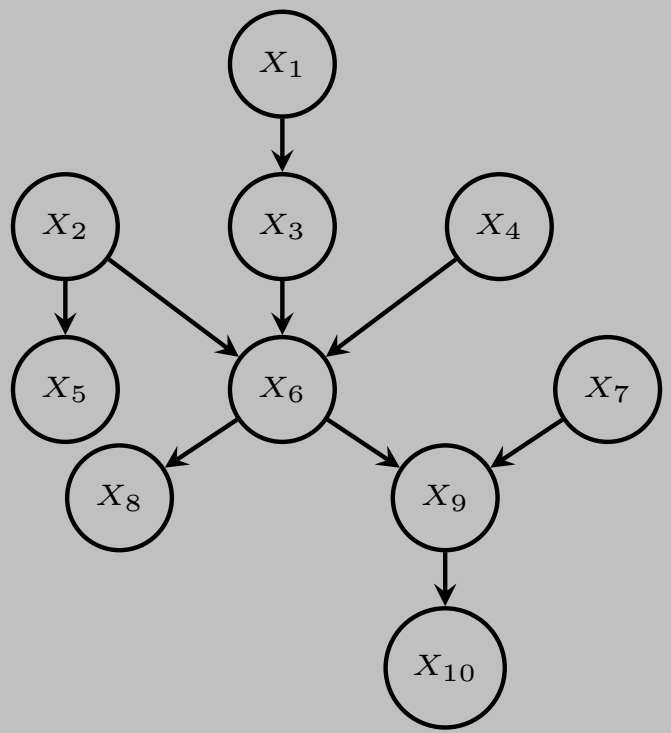


Este no es un DAG:



Relaciones en un DAG

En un DAG, los *ancestros* de A son todos aquellos nodos que tienen un camino dirigido que termina en A. Por otro lado, los *descendientes* de A son todos aquellos nodos que tiene un camino dirigido que comienza en A.



En el DAG anterior, los ancestros de X_6 son $\{X_1, X_2, X_3, X_4\}$, y sus descendientes son $\{X_8, X_9, X_{10}\}$.

Redes bayesianas

Suponga que se cuenta con una distribución p definida sobre n v.a., las cuales hemos ordenado de manera arbitraria como $X_1, X_2, \ldots, X_n$. La regla de la cadena del cálculo de probabilidades nos permite descomponer p como el producto de n distribuciones condicionales:

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j | X_1, \ldots, X_{j-1}).$$

Suponga que la probabilidad condicional de la variable X_j no depende realmente de todos los predecesores de X_j, sino únicamente de un subconjunto de estos. En otras palabras, suponga que X_j es independiente de todos los otros predecesores, una vez que conocemos el valor de un subconjunto de ellos, los cuales denotaremos como PA_j. Entonces,

$$p(X_j | X_1, \ldots, X_{j-1}) = p(X_j | \mathrm{PA}_j).$$

Padres markovianos

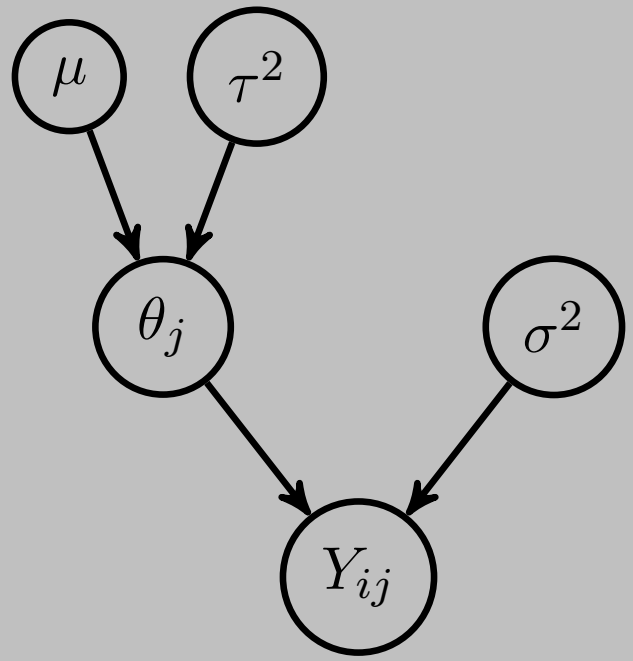
Al conjunto PA_j se le llama los padres markovianos de X_j, o simplemente sus padres. Esta condición puede ser representada mediante un DAG en la que las variables representan nodos y las flechas son dibujadas desde cada nodo en el conjunto de padres PA_j hacia el nodo hijo X_j.

Modelos jerárquicos bayesianos

Por ejemplo, considere el siguiente modelo jerárquico bayesiano:

$$
\begin{aligned}
Y_{ij}|\theta_j, \sigma^2 &\sim \mathcal{N}(\theta_j, \sigma^2), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \\
\theta_j|\mu, \tau^2 &\sim \mathcal{N}(\mu, \tau^2), \quad j = 1, \ldots, J \\
p(\mu) &\propto \mathbb{1}_{\mathbb{R}}(\mu) \\
p(\tau^2) &\propto \mathbb{1}_{(0,\infty)}(\tau^2) \\
p(\sigma^2) &\propto \frac{1}{\sigma^2} \mathbb{1}_{(0,\infty)}(\sigma^2).
\end{aligned}
$$

El cual puede ser representado graficamente como:



Entonces,

$$\begin{aligned} p(Y_{ij}, \theta_j, \mu, \tau^2, \sigma^2) =& p(Y_{ij}|\theta_j, \mu, \tau^2, \sigma^2) p(\theta_j|\mu, \tau^2, \sigma^2) \\ & \times p(\mu|\tau^2, \sigma^2) p(\tau^2|\sigma^2) p(\sigma^2) \end{aligned}$$

$$= p(Y_{ij}|\theta_j, \sigma^2) p(\theta_j|\mu, \tau^2) p(\mu) p(\tau^2) p(\sigma^2)$$

Compatibilidad markoviana

Si una distribución de probabilidad p admite la factorización de la forma

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j | \mathrm{PA}_j)$$

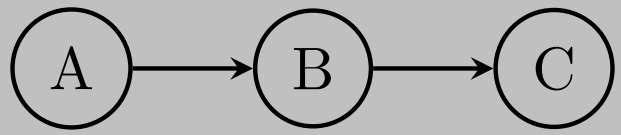
relativa al DAG $\mathcal{G}$, decimos que $\mathcal{G}$ representa p, que $\mathcal{G}$ y p son compatibles, o que p es Markov relativa a $\mathcal{G}$.

Reglas de *d*-separación

Reglas de *d*-separación

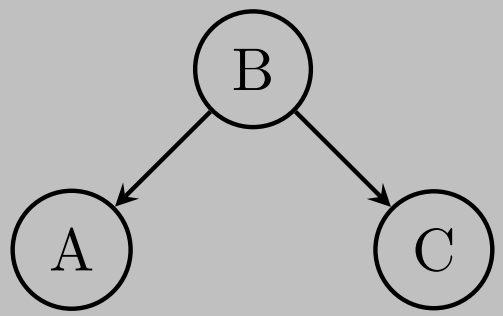
Para cualquier DAG, existen cuatro relaciones básicas entre las variables, las cuales satisfacen ciertas condiciones de independencia.

Tubería, cadena o mediador



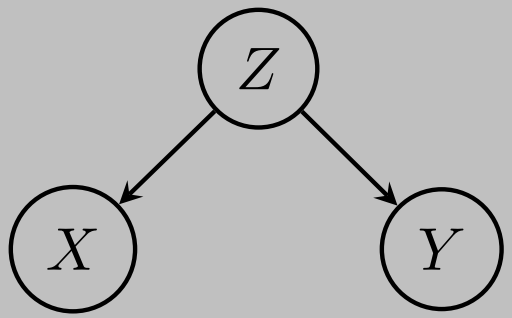
Esta relación es una cadena o mediación, en la que B es el mecanismo o *mediador* que transmite el efecto de A a C. Este DAG crea una relación entre A y C, por lo que $A \not\perp\!\!\!\perp C$. Pero, una vez que se inclye B la relación desaparece. Es decir, condicionado en B, A y C son independientes $A \perp\!\!\!\perp C|B$. En otras palabras, una vez que conocemos B, A no agrega más información sobre C.

Tenedor



Esta relación es llamada *tenedor*, y B es llamada causa común de A y C. Como en la cadena, este DAG crea una relación entre A y C, $A \not\perp\!\!\!\perp C$. Pero una vez que se incluye B la relación desaparece, $A \perp\!\!\!\perp C|B$.

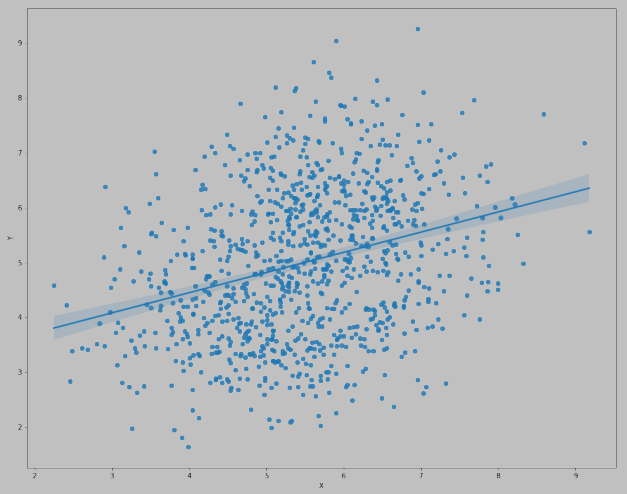
Ejemplo de un tenedor

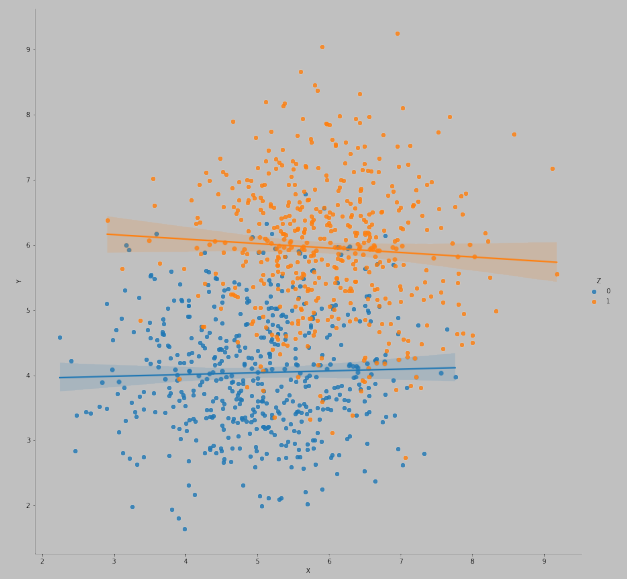


$$Z \sim \text{Ber}(p).$$

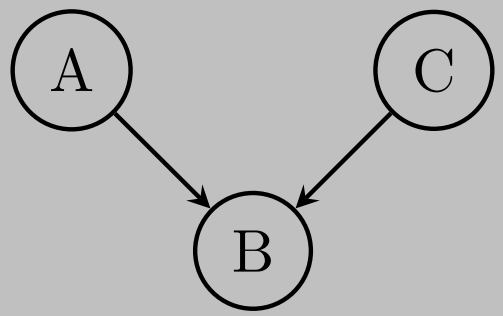
$$X \sim \begin{cases} \mathcal{N}(5,1) & \text{if } Z = 0, \\ \mathcal{N}(6,1) & \text{if } Z = 1. \end{cases}$$

$$Y \sim \begin{cases} \mathcal{N}(4,1) & \text{if } Z = 0, \\ \mathcal{N}(6,1) & \text{if } Z = 1. \end{cases}$$



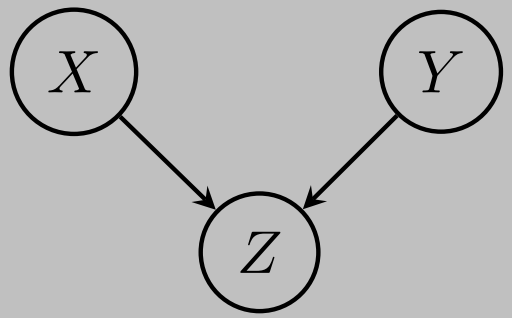


Colisionador



Esta relación es llamada *colisionador*. A diferencia de las otras dos relaciones, en un colisionador no existe asociación entre A y C a menos que se condicione en B. Es decir, $A \perp\!\!\!\perp C$, pero $A \not\perp\!\!\!\perp C|B$.

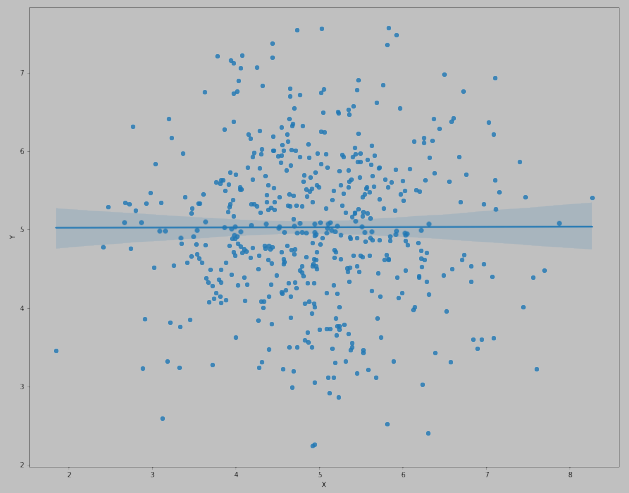
Ejemplo de colisionador

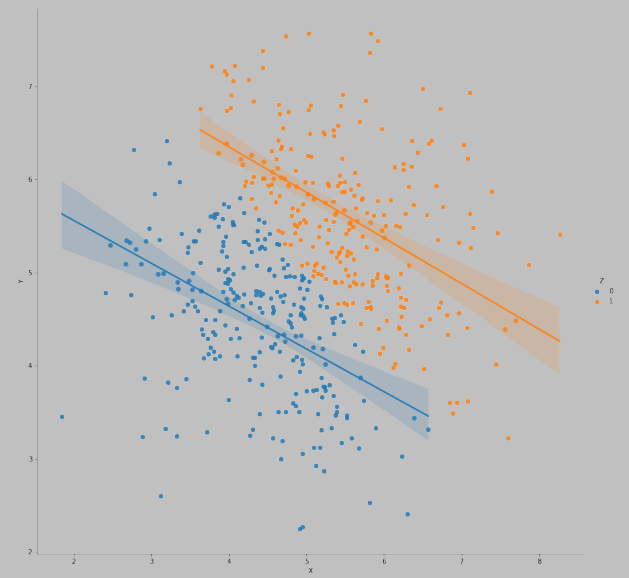


$$X \sim \mathcal{N}(5, 1).$$

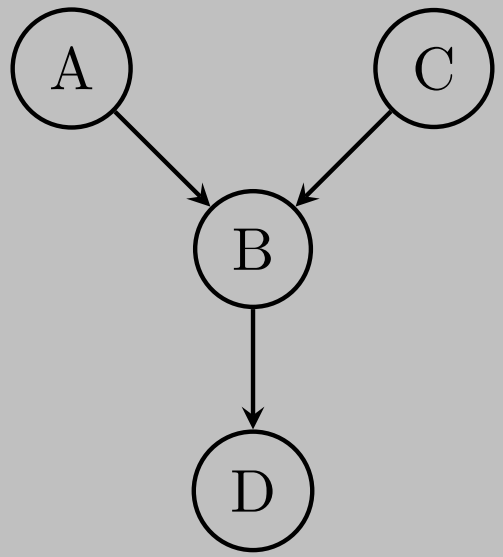
$$Y \sim \mathcal{N}(5, 1).$$

$$Z = \begin{cases} 1 & \text{if } 5X + 5Y > 50, \\ 0 & \text{otherwise.} \end{cases}$$





Descendiente



Condicionar en un descendiente condiciona, hasta cierto punto, en sus padres. Condicionar en D condiciona, en menor medida, en B. La razón es que D posee algo de información sobre B. Los descendientes son comunes, como no siempre podemos medir alguna variable directamente usamos un *proxy*.

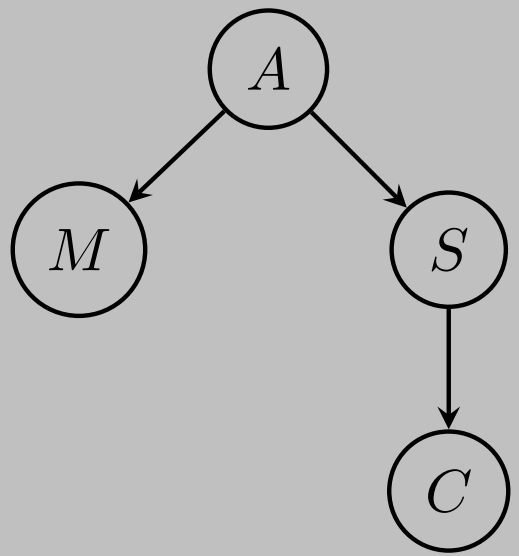
Acertijos con datos reales

Estos son dos acertijos que he encontrado analizando datos reales:

1. Las personas viudas son menos riesgosas según las calificaciones crediticias.
2. Las mujeres tienden a estar solteras o divorciadas, mientras que los hombres tienden a estar casados.

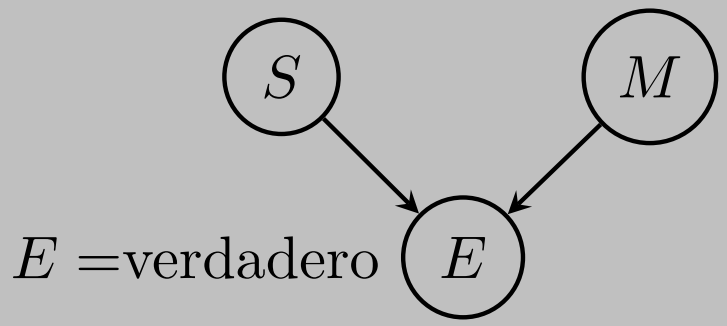
En ambos casos existe una relación entre dos variables que es difícil de explicar, hasta que consideramos una tercer variable.

Para solucionar el primer acertijo, debemos considerar la edad. Las personas viudas tienden a ser mayores que las personas solteras o casadas. Además, las personas mayores tienden a tener mayor estabilidad económica y tener menor deuda que las personas jóvenes. Por lo tanto, podemos inferir que el siguiente DAG explica la relación entre el estado civil de una persona y su calificación crediticia.



donde las variables son, A: edad, M: estado civil, S: estabildiad económica, C: calificación crediticia.

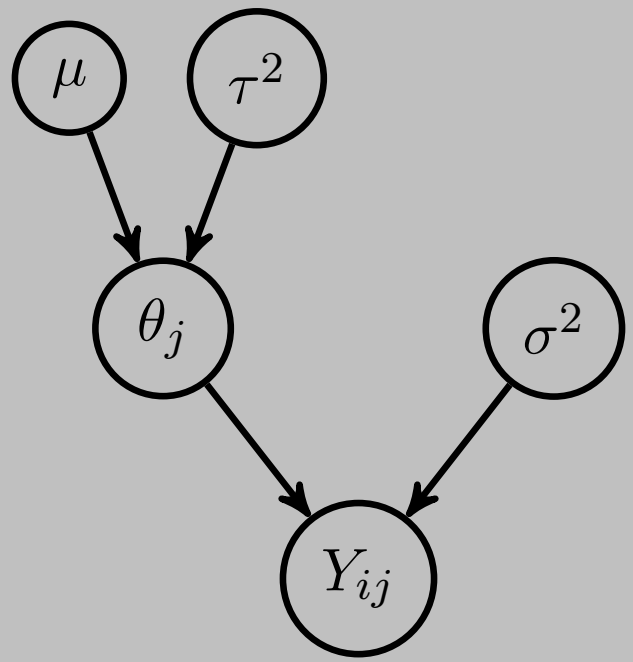
Algo que no mencioné en el segundo acertijo fue que sólo se consideraron empleados. En la sociedad mexicana todavía es común que, en parejas casadas, sean los hombres los que trabajan mientras las mujeres se quedan al cuidado del hogar. Entonces, las mujeres que trabajan son aquellas que no están casadas. Es decir, ser empleado es un colisionador del sexo y estado civil de una persona.



donde las variables son, S: sexo, M: estado civil, E: empleado. Y estamos condicionando en el valor $E =$verdadero.

d-separación en el modelo jerárquico

Recuerde que nuestro modelo jerárquico podía ser representado gráficamente como



En estadística bayesiana estamos interesados en simular una muestra de la distribución posterior. Un método para hacerlo es el muestreador Gibbs, para el cual necesitamos todas las distribuciones posteriores condicionales.

Posterior condicional de μ y τ^2

La siguiente tabla muestra los caminos que conectan μ con las otras variables:

Camino	Contiene colisionadores	Abierto/Cerrado	Cómo abrir/cerrar
$\mu \to \theta_j \leftarrow \tau^2$	Sí	Cerrado	Para abrir: condicionar en θ_j
$\mu \to \theta_j \to Y_{ij} \leftarrow \sigma^2$	Sí	Cerrado	Para abrir: condicionar en Y_{ij} Para cerrar: condicionar en θ_j

Por lo tanto,

$$p(\mu|\boldsymbol{\theta}, \tau^2, \sigma^2, \mathbf{Y}) = p(\mu|\boldsymbol{\theta}, \tau^2).$$

Similarmente,

$$p(\tau^2|\boldsymbol{\theta}, \mu, \sigma^2, \mathbf{Y}) = p(\tau^2|\boldsymbol{\theta}, \mu).$$

Posterior condicional de σ^2

La siguiente tabla muestra los caminos que conectan σ^2 con las otras variables:

Camino	Contiene colisionadores	Abierto/Cerrado	Cómo abrir/cerrar
$\sigma^2 \to Y_{ij} \leftarrow \theta_j$	Sí	Cerrado	Para abrir: condicionar en Y_{ij}
$\sigma^2 \to Y_{ij} \leftarrow \theta_j \leftarrow \mu$ $\sigma^2 \to Y_{ij} \leftarrow \theta_j \leftarrow \tau^2$	Sí	Cerrado	Para abrir: condicionar en Y_{ij} Para cerrar: condicionar en θ_j

Entonces,

$$p(\sigma^2|\boldsymbol{\theta}, \mu, \tau^2, \mathbf{Y}) = p(\sigma^2|\boldsymbol{\theta}, \mathbf{Y}).$$

Conditional posterior of θ_j

The next table shows the paths that connect θ_j with the other variables:

Camino	Contiene colisionadores	Abierto/Cerrado	Cómo abrir/cerrar
$\theta_j \rightarrow Y_{ij} \leftarrow \sigma^2$	Sí	Cerrado	Para abrir: condicionar en Y_{ij}

Thus,

$$p(\theta_j|\mu, \tau^2, \sigma^2, \mathbf{Y}) = p(\theta_j|\mu, \tau^2, \sigma^2, \mathbf{Y}).$$

Descendiente en el modelo jerárquico

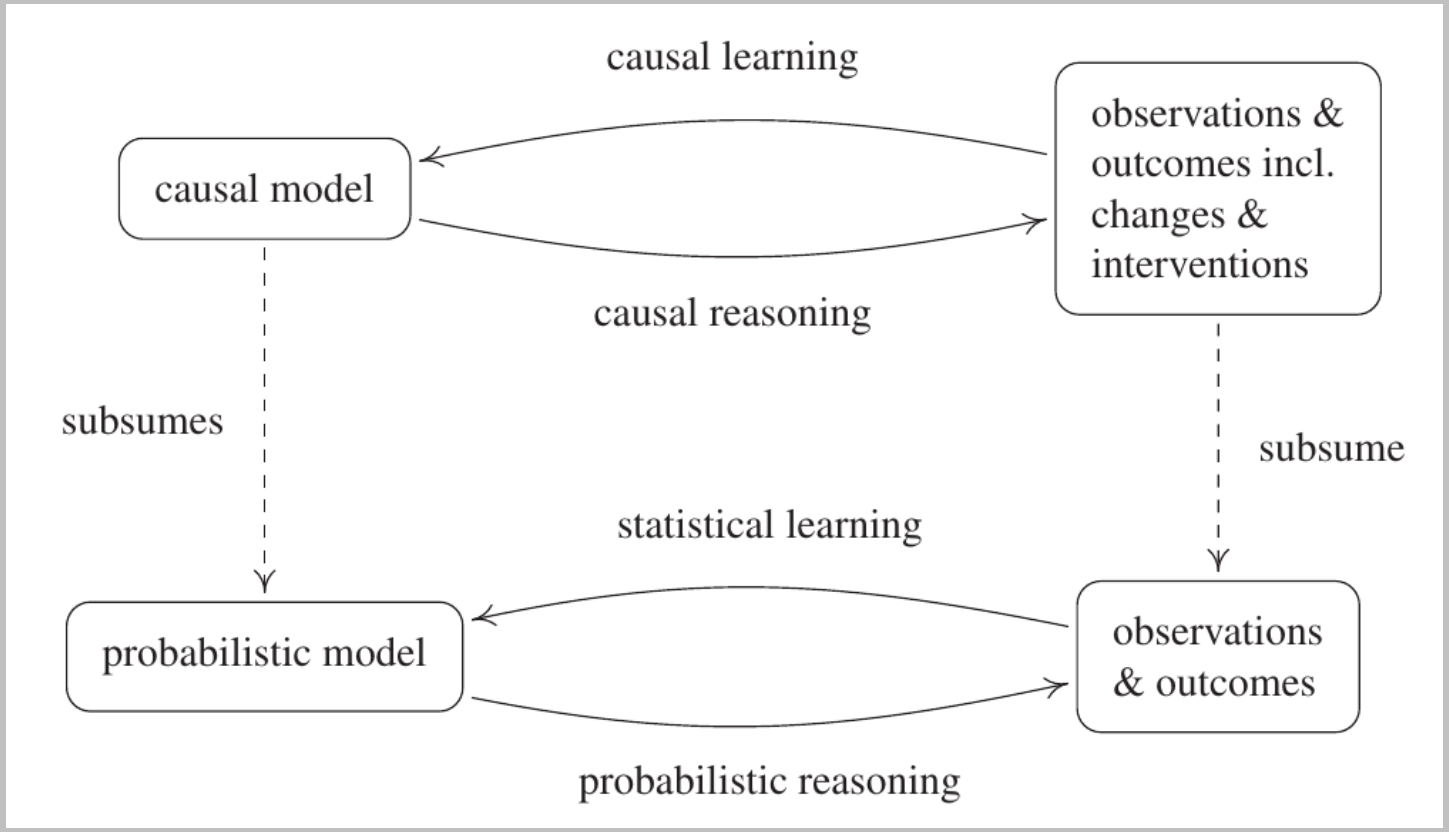
Como Y_{ij} es un descendiente de θ_j, entonces condicionar en Y_{ij} crea una relación entre μ y τ^2. Pero, como Y_{ij} es un colisionar entre θ_j y σ^2, condicionar en Y_{ij} también crea una relación entre θ_j y σ^2. De hecho, se puede demostrar que

$$\mu|\tau^2, \sigma^2, \mathbf{Y} \sim \mathcal{N}(\hat{\mu}, V_\mu),$$

donde $\hat{\mu}$ y V_μ son funciones de $\tau^2, \sigma^2, \mathbf{Y}$.

Causalidad

Razonamiento causal e inferencia causal



El objetivo de la inferencia estadística es ajustar de la mejor manera posible un modelo probabílistico a partir de datos observados del fenómeno que estamos estudiando. Por otro lado, también podemos partir ded un modelo probabilístico dado, y deducir el comportamiento de observaciones bajo dicho modelo, lo que llamaremos razonamiento probabilístico.

De manera análoga, también podemos distinguir una inferencia causal, en la que a partir de datos observados buscamos inferir un modelo causal que los explique. Mientras que el razonamiento causal lo entenderemos como el proceso en el que buscamos deducir comportamientos causales a partir de un modelo causal dado.

DAGs

La interpretación de los DAGs como portadores de supuestos de independencia no implica necesariamante una causalidad. Sin embargo, es justamente este interpretación la que potencia el uso de DAGs.

do-cálculo

Causalidad

David Hume propuso la siguiente definición de causalidad:

> *Podemos definir una causa como un objeto seguido de por otro, donde todos los objetos, similares al primero, son seguidos por objetos similares al segundo. O, en otras palabras, donde,* ***si el primer objeto no hubiera estado, el segundo nunca hibiera existido.***

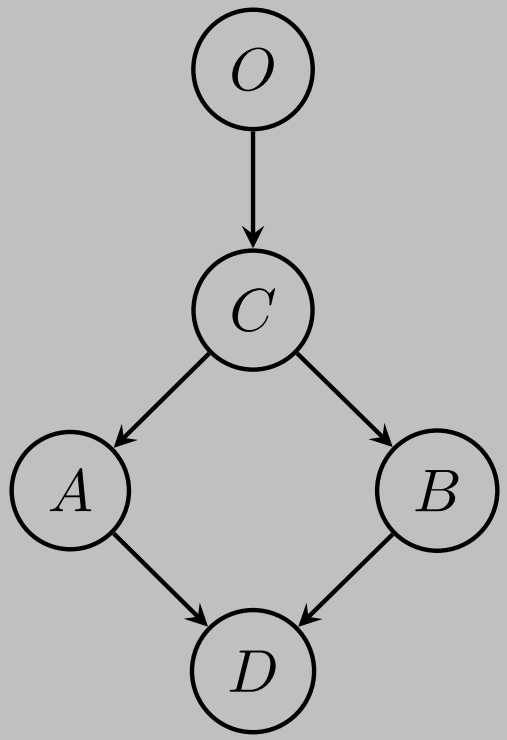
La escalera de la causalidad

Podemos distinguir tres "niveles" de inferencia causal.

1. El nivel más básico es la **asociación**, que corresponde con la actividad de **mirar**. El objetivo de mirar u observar es detectar regularidades en nuestro entorno. En este nivel sólo observamos si un conjunto de variables están estadísticamente relacionadas.
2. El segundo nivel es la **intervención**. La actividad de **hacer** corresponde a este nivel. **El hacer buscar predecir los efectos de alteraciones deliberadas** del entorno y escoger entre estas alteraciones la que produce el resultado deseado.
3. El nivel más alto es el de los **contrafactuales**, que corresponden con la actividad de **imaginar**. Es inútil preguntarse por las causas de las cosas a menos que podamos imaginar las consecuencias.

Ejemplo, pelotón de fusilamiento

Suponga que un prisionero está a punto de ser ejecutado por un pelotón de fusilamiento. Una serie de eventos deben de suceder opara que esto ocurra. Primero, una corte ordena la ejecución. La orden llega al capitań, quien da la orden a los soldados del pelotón (A y B) de disparar. Supondremos que los soldados son obedientes y expertos tiradores, por lo que sólo disparan bajo la orden dada, y si cualquiera de ellos dispara, el prisionero muere.



O: Corte, C: Capitán, A, B: Soldados, D: Estado del prisionero.

Asociación

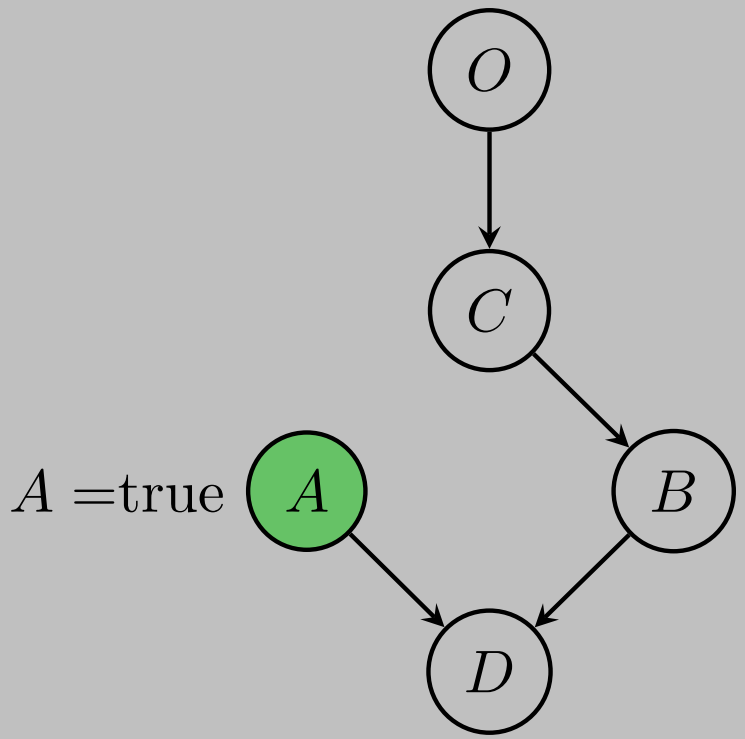
- Si el prisionero está muerto, ¿entonces la orden fue dada por la corte? Sí.
- Suponga que observamos que A disparó. ¿Qué nos dice esto sobre B? A no hubiera disparado si el capitán no lo hubiera ordenado, por lo tanto B también debió de disparar. Note que en este caso A y B están perfectamente correlacionadas, aunque A no causa B ni viceversa.

Intervención

- ¿Que sucede si A decide disparar por iniciativa propia? ¿El prisionero estaría vivo o muerto?

Si sólo usamos las reglas de la lógica, como nuestras computadoras lo hacen, la pregunta carece de sentido. Si queremos que las computadoras entiendan sobre causalidad, tenemos que enseñarles cómo romper las reglas. Tenemos que enseñarles la diferencia entre meramente observar un evento y hacer que éste suceda.

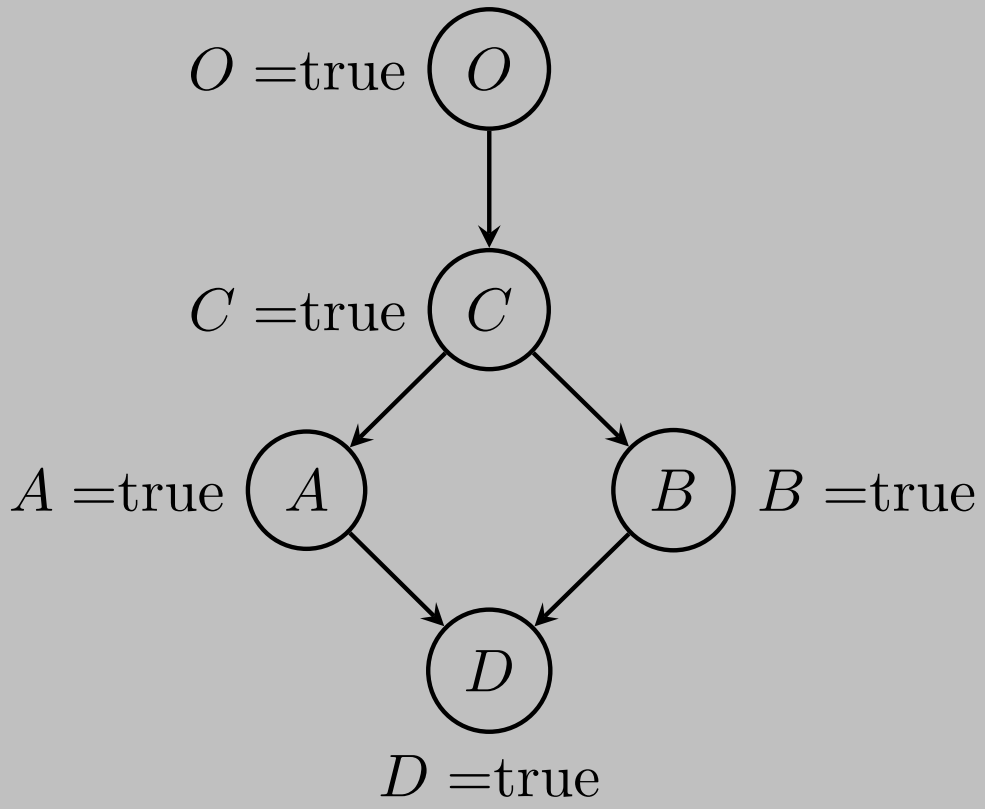
Hacer que un evento ocurra significa que lo libramos de todo lo demás que lo influye. Es decir, borramos todas las flechas que apuntan a la variable que intervenida (A), establecemos esa variable igual al valor prescrito (verdadero) y continuamos el análisis con la lógica usual. Esto es porque una vez que **hacemos** la intervención sólo resta **observar** su efecto.



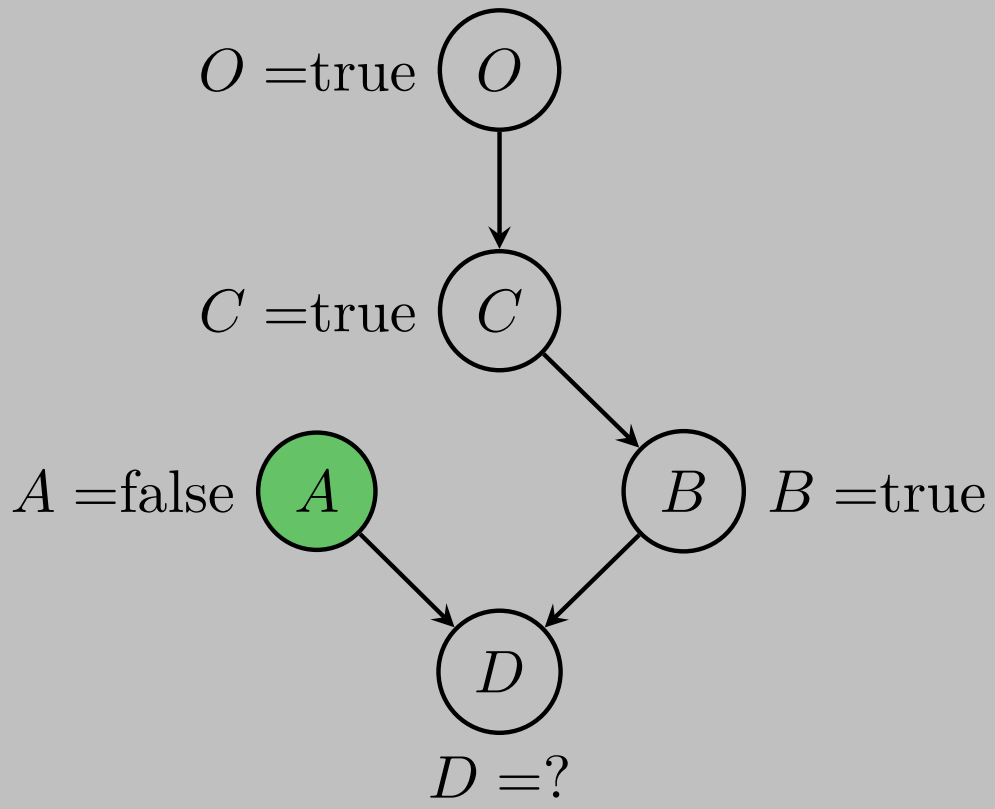
Contrafactual

Suponga que el prisionero está muerto. De esto podemos conlcuir que A disparó, B disparó, que el capitán dio la señal, y que la corte dio la orden. Pero

- ¿Qué hubiera pasado si A hubiera decidido no disparar? ¿El prisionero estaría vivo?



Esta pregunta requiere que comparemos el mundo real con uno ficticio y contradictorio donde A no disparó. En el mundo ficticio eliminamos las flechas que apuntan a A y fijamos su valor a falso, dejando el resto de la historia pasada igual a como sucedió en el mundo real.



Concluimos que el prisionero también estaría muerto en este mundo ficticio, pues el disparo de B lo habría matado. La valentía de A no le habría salvado la vida.

Sin duda, esta es una de las razones de que existan los pelotones de fusilamiento: garantiza que se realice la orden de la corte y quita la posible sensación de culpa de cada uno de los individuos del pelotón, quienes pueden decir que sus acciones no causaron la muerte del prisionera, pues este estaría muerto de cualquier manera.

Si juntamos reportes de distintas "ejecuciones", nuestros datos se verían algo como:

O	C	A	B	D
true	true	true	true	true
true	true	true	true	true
false	false	false	false	false

No hay manera en que estos datos, en ausencia de una relación causasl puedan predecir el resultado de persuadir al soldado A de disparar, no importa cuántos datos tengamos.

O	C	A	B	D
true	true	false	true	?

Relación entre probabilidad y causalidad

La udea de causas y efectos es más fundamental que la idea de probabilidad. Aprendemos sobre causas y efectos antes de que aprendamos sobre matemáticas.

Entender el significado de "causa" ha sido el centro de atención de varios filósofos, quienes han tratado de definir causalidad en términos de probabilidad, usando la idea de que X causa Y si X aumenta la probabilidad de Y.

Por ejemplo, decimos "conducir en estado de ebriedad cusa accidentes" sabiendo que la causa hace más probable la consecuencia, no que suceda con total certeza.

Pero este incremento en la probabilidad puede suceder por otras razones, puede ser que X sea una causa de Y o que exista otra variable Z que sea causa común de ambos. ¡Ese es el problema!

Operador *do*

Para rescatar la idea del incrmento en la probabilidad, usamos el operador *do*. Mientras que $\mathbb{P}(Y|X = x)$ denota la distribución observada, que corresponde con mirar. Es decir, $\mathbb{P}(Y|X = x)$ corresponde con la distribución de Y entre aquellos individuos cuyo valor de X es x.

Por otro lado, $\mathbb{P}(Y|\text{do}(X = x))$ corresponde a la distribución intervenida, que corresponde con hacer. Es decir, $\mathbb{P}(Y|\text{do}(X = x))$ corresponde con la distribución de Y si todos en los individuos tuvieran el valor de X igual a x. Por lo tanto, $\mathbb{P}(Y|\text{do}(X = x))$ describe el efecto causal de X en Y.

Factorización truncada

Hemos comentado que realizar la intervención $\text{do}(X_i = x_i)$, en el DAG $\mathcal{G}$ se obtiene al crear un nuevo DAG $\mathcal{G}'$, donde las aristas entre PA_i y X_i son borrados, y fijamos el valor de X_i en x_i. Esto es equivalente a remover el término $p(X_i|\text{PA}_i)$ en la fórmula de factorización:

$$p(X_1, \ldots, X_n) = \prod_{j=1}^{n} p(X_j|\text{PA}_j),$$

y fijar $X_i = x_i$. Entonces,

$$\begin{aligned} &p(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n|\text{do}(X_i = x_i)) \\ &= p^{\mathcal{G}'}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n|X_i = x_i) \\ &= \prod_{j \neq i} p(X_j|\text{PA}_j, X_i = x_i) \end{aligned}$$

Paradoja de Simpson

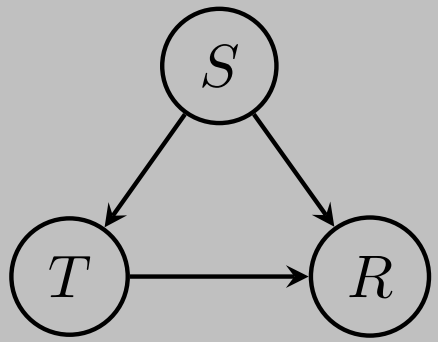
Suponga que observamos 700 pacientes quienes pueden haber elegido someterse a un tratamiento o no hacerlo.

	Sin tratamiento
Hombres	234 recuperados de 270 (87%)
Mujeres	55 recuperados de 80 (69%)
Hombers y mujeres	289 recuperados de 350 (83%)

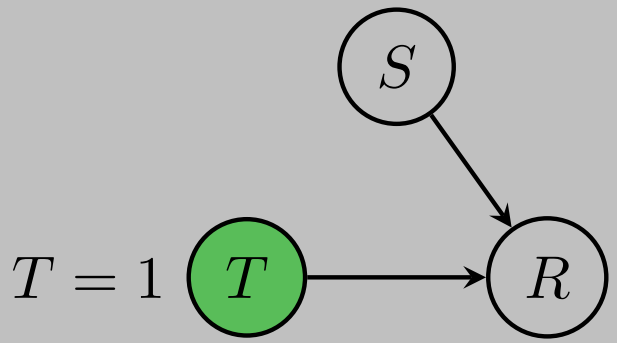
	Con tratamiento
Hombres	81 recuperados de 87 (93%)
Mujeres	192 recuperados de 263 (73%)
Hombres y mujeres	273 recuperados de 350 (78%)

¿Se debería de recomendar el tratamiento o no?

Para responder la pregunta necesitamos calcular el efecto causal del tratamiento en la recuperación.



DAG: $\mathcal{G}$



DAG: $\mathcal{G}'$

$$
\begin{aligned}
&\mathbb{P}(R=1|\mathrm{do}(T=1)) \\
&= \mathbb{P}^{\mathcal{G}'}(R=1|T=1) \\
&= \mathbb{P}^{\mathcal{G}'}(R=1|S=1,T=1)\mathbb{P}^{\mathcal{G}'}(S=1|T=1) \\
&\quad + \mathbb{P}^{\mathcal{G}'}(R=1|S=0,T=1)\mathbb{P}^{\mathcal{G}'}(S=0|T=1)
\end{aligned}
$$

Efecto promedio del tratamiento (ATE)

$$\begin{aligned}\mathbb{P}(R=1|\text{do}(T=1)) =& \mathbb{P}(R=1|S=1,T=1)\mathbb{P}(S=1) \\ &+ \mathbb{P}(R=1|S=0,T=1)\mathbb{P}(S=0) \\ =& \frac{81}{87}\left(\frac{87+270}{700}\right) + \frac{192}{263}\left(\frac{263+80}{700}\right) \\ \approx& 0.833\end{aligned}$$

$$\begin{aligned}\mathbb{P}(R=1|\text{do}(T=0)) =& \mathbb{P}(R=1|S=1,T=0)\mathbb{P}(S=1) \\ &+ \mathbb{P}(R=1|S=0,T=0)\mathbb{P}(S=0) \\ =& \frac{234}{270}\left(\frac{87+270}{700}\right) + \frac{55}{80}\left(\frac{263+80}{700}\right) \\ \approx& 0.779\end{aligned}$$

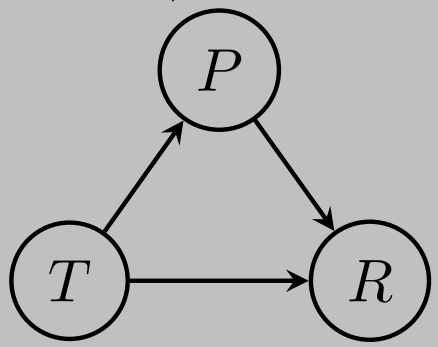
En promedio, se recuperan 5.4% más pacientes cuando se someten al tratamiento.

Suponga ahora, que en vez del sexo del paciente, lo que se registra es su presión sanguínea, la cual puede ser alta o baja. Además, suponga que el tratamiento tiene un efecto directo en la recuperación del paciente, y un efecto indirecto alterando la presión sanguínea

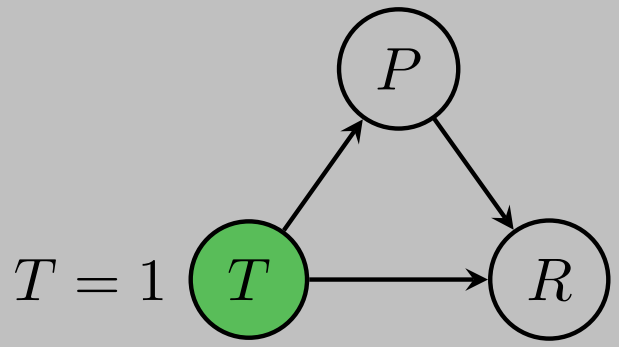
	Con tratamiento
Baja presión	234 recuperados de 270 (87%)
Alta presión	55 recuperados de 80 (69%)
Ambos	289 recuperados de 350 (83%)

	Sin tratamiento
Baja presión	81 recuperados de 87 (93%)
Alta presión	192 recuperados de 263 (73%)
Ambos	273 recuperados de 350 (78%)

En este caso, nuestros gráfico causales serían:



DAG: $\mathcal{G}$



DAG: $\mathcal{G}'$

Efecto causal promedio (ACE)

$$\begin{aligned}\mathbb{P}(R=1|\text{do}(T=1)) =& \mathbb{P}(R=1|P=1,T=1)\mathbb{P}(P=1|T=1) \\ &+ \mathbb{P}(R=1|P=0,T=1)\mathbb{P}(P=0|T=1) \\ =& \frac{234}{270}\left(\frac{270}{350}\right) + \frac{55}{80}\left(\frac{80}{350}\right) \\ \approx& 0.826\end{aligned}$$

$$\begin{aligned}\mathbb{P}(R=1|\text{do}(T=0)) =& \mathbb{P}(R=1|P=1,T=0)\mathbb{P}(P=1|T=0) \\ &+ \mathbb{P}(R=1|P=0,T=0)\mathbb{P}(P=0|T=0) \\ =& \frac{81}{87}\left(\frac{87}{350}\right) + \frac{192}{263}\left(\frac{263}{350}\right) \\ \approx& 0.78\end{aligned}$$

En promedio, se recuperan 4.6% más pacientes cuando se someten al tratamiento.

Criterio de la puerta trasera

Decimos que un conjunto de variables Z satisface el criterio de la puerta trasera relativo a un par ordenado de variables (X, Y) en un DAG $\mathcal{G}$ si:

1. Ninguna de las variables en Z es descendiente de X.
2. Z bloquea todos los caminos entre X y Y que tienen una flecha apuntando a X

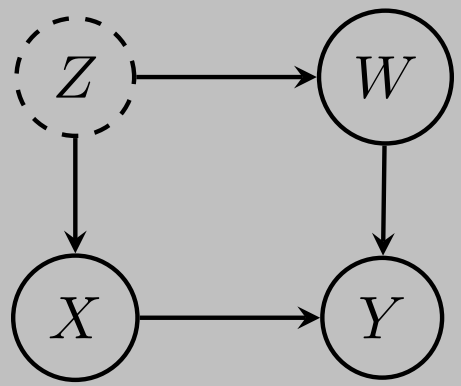
Si Z satisface el criterio de la puerta trasera relativo a (X, Y), entonces el efecto causal de X en Y puede ser calculado mediante la fórmula:

$$p(Y|\text{do}(X = x)) = \sum_z p(Y|x, z)p(z)$$

Nota: Los padres de X siempre satisfacen el criterio de la puerta trasera.

Ejemplo

Considere el siguiente DAG que modela la relación entre un tratamiento X y recuperarse de alguna enfermedad Y. Además suponga que el estado de salud W también afecta la posibilidad de recuperarse, y que el estado socioeconómico afecta al estado de salud y a la posibilidad de tener acceso al tratamiento X.



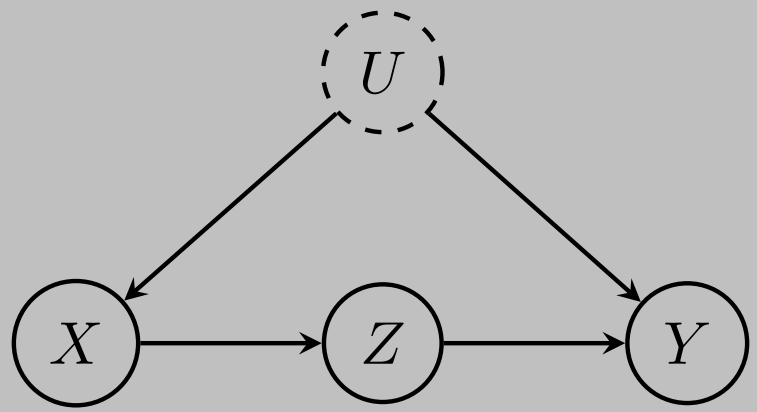
Sin embargo, en el estudio realizado no se recopiló información sobre el estado socioeconómico.

Como además de Z, W también bloquea el camino no causal $X \leftarrow Z \rightarrow W \rightarrow Y$, entonces podemos calcular el efecto causal de X en Y mediante:

$$\mathbb{P}(Y = y|\text{do}(X = x)) = \sum_{w} \mathbb{P}(Y = y|X = x, W = x)\mathbb{P}(W = w)$$

Criterio de la puerta delantera

Considere el siguiente DAG:



donde la(s) variable(s) U no es/son observada(s).
La distribución conjunta de X, Y, Z, U puede ser escrita como

$$p(X, Y, Z, U) = p(U)p(X|U)p(Z|X)p(Y|Z, U).$$

Usando la fórmula de la factorización truncada,

$$p(Y, Z, U|\text{do}(X = x)) = p(U)p(Z|x)p(Y|Z, U),$$

marginalizando sobre Z y U, obtenemos

$$p(Y|\text{do}(X = x)) = \sum_z p(z|x) \sum_u p(u)p(Y|z, u).$$

Por otro lado, como $Z \perp\!\!\!\perp U|X$, entonces $p(U|Z,X) = p(U|X)$. Y como $Y \perp\!\!\!\perp X|Z,U$, entonces $p(Y|X,Z,U) = p(Y|Z,U)$. De estas dos expresiones obtenemos

$$\begin{aligned}\sum_u p(Y|Z,u)p(u) &= \sum_x \sum_u p(Y|Z,u)p(u|x)p(x) \\ &= \sum_x \sum_u p(Y|Z,x,u)p(u|,Z,x)p(x) \\ &= \sum_x p(Y|Z,x)p(x).\end{aligned}$$

Ahora, recuerde que

$$p(Y|\text{do}(X=x)) = \sum_z p(z|x) \sum_u p(u)p(Y|z,u).$$

Por lo tanto,

$$p(Y|\text{do}(X=x)) = \sum_z p(z|x) \sum_{x'} p(Y|z,x')p(x').$$

Decimos que un conjunto de variables Z satisface el criterio de la puerta delantera relativo a un par ordenado de variables (X, Y) en un DAG $\mathcal{G}$ si:

1. Z intercepta todos los caminos causales de X a Y.
2. No existe ningún camino no causal entre X y Z.
3. Todos los caminos no causales entre Z y Y están bloqueados por X.

Si Z satisface el criterio de la puerta delantera relativo a (X, Y), entonces el efecto causal de X en Y puede ser calculado mediante la fórmula:

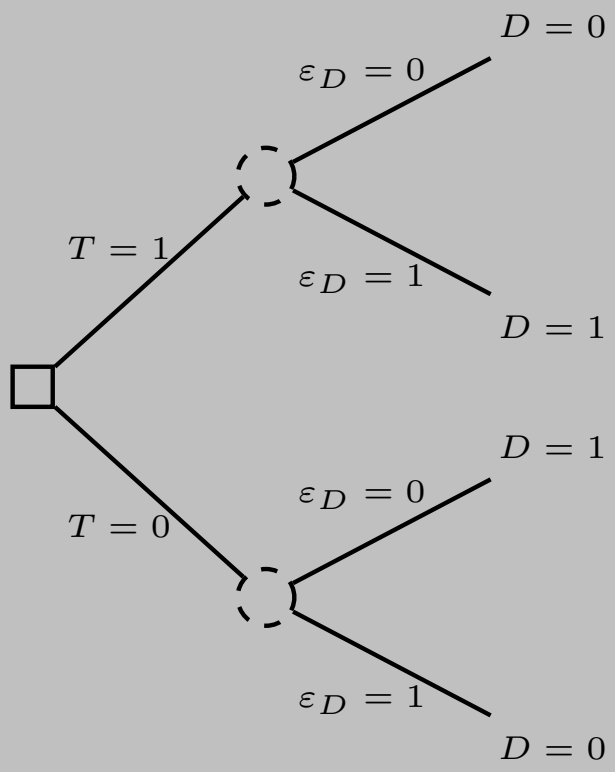
$$p(Y|\text{do}(X = x)) = \sum_z p(z|x) \sum_{x'} p(Y|z, x')p(x').$$

Modelos causales estructurales

Modelos causales estructurales (SCM)

Suponga que existe un tratamiento para una enfermadad mortal. Para el 99% de la pobalación el tratamiento ($T = 1$) funciona y se recuperan ($D = 0$). El 1% restante tiene una extraña condición ($\varepsilon_D = 1$) que la hace inmune a la enfermedad, pero el tratamiento se vuelve fatal.

Esta rara condición es desconocidad para un médico, cuya decisión de administrar el tratamiento es entonces independiente de ε_D. Denotamos esta variable como ε_T.



En este caso tenemos el grafo causal: $T \to D \leftarrow \varepsilon_D$ y las *ecuaciones causales estructurales (SCEs)*:

$$\mathfrak{C} : \begin{cases} T := \varepsilon_T, \\ D := T\varepsilon_D + (1-T)(1-\varepsilon_D), \end{cases}$$

ambos determinan el SCM $\mathfrak{C}$.

Pregunta contrafactural

Un paciente muere ($D = 1$) después de que el médico administrara el tratamiento ($T = 1$). Podemos realizar la pregunta contrafactual *"¿Qué hubiera sucedido si el médico no hubiera administrado el tratamiento?"*

Las personas que mueren bajo el tratamiento es porque cuentan con la extraña condición, que corresponde con $\varepsilon_D = 1$, lo que significa que se recuperarían si y sólo si no se les trata.

El proceso para concluir que una persona fallecida bajo tratamiento ($T = 1, D = 1$) se habría recuperado si no hubiera sido tratada requiere tres pasos.

Abducción y acción

Primero, aplicamos la evidencia con la que contamos, $e : \{T = 1, D = 1\}$, en el modelo y conluímos que e es compatible únicamente con una valor de ε_T y ε_D, $\{\varepsilon_T = 1, \varepsilon_D = 1\}$.

$$\mathfrak{C}|T = 1, D = 1 : \begin{cases} 1 = \varepsilon_T \\ 1 = 1\varepsilon_D + (1 - 1)(1 - \varepsilon_D). \end{cases}$$

$$\Rightarrow \{\varepsilon_T = 1, \varepsilon_D = 1\}$$

Segundo, para calcular el efecto de $\text{do}(T = t)$, sustituimos $T = t$ en la segunda ecuación de $\mathfrak{C}$ e ignoramos la primera ecuación $T := \varepsilon_T$.

$$\mathfrak{C}|\text{do}(T = t) : \Big\{ D := t\varepsilon_D + (1 - t)(1 - \varepsilon_D).$$

Predicción

Tercero, simulamos la condición hipotética "la persona no fue tratada", sustituyendo $t = 0$ en este modelo modificado $\mathfrak{C}$, ignorando la primera ecuación.

$$\mathfrak{C}|T = 1, D = 1, \text{do}(T = 0) : \begin{cases} D := 0(1) + (1 - 0)(1 - 1) \\ \quad = 0. \end{cases}$$

Con lo que conluimos que la personas estaría viva si no hubiera recibido el tratamiento.

Intervención

Note que la acción $\text{do}(T = t)$ corresponde con el modelo intervenido:

$$\mathfrak{C}|\text{do}(T = t) : \Big\{ D := t\varepsilon_D + (1 - t)(1 - \varepsilon_D).$$

$$\mathbb{P}^{\mathfrak{C}|\text{do}(T=1)}(D = 0) = \mathbb{P}(\varepsilon_D = 0) = 0.99,$$

and

$$\mathbb{P}^{\mathfrak{C}|\text{do}(T=0)}(D = 0) = \mathbb{P}(1 - \varepsilon_D = 0) = 0.01.$$

Por lo tanto, podemos argumentar que el médido no actuó de manera negligente (de acuerdo a la información disponible).

Pasos contrafactuales

Estos tres pasos pueden ser generalizados a cualquier modelo causal $\mathfrak{C}$ como sigue.

Dada la evidencia e, para calcular la probabilidad de Y bajo la condición hipotética $X = x$ (donde X es un subconjunto de variables), usando los siguientes tres pasos a $\mathfrak{C}$.

1. **Abducción:** Calculamos la distribución $p(\varepsilon|e)$.
2. **Acción:** Reemplazamos las ecuaciones correspondientes a las variables en el conjunto X por las ecuaciones $X = x$.
3. **Predicción:** Usamos el modelo modificado para calcular la probabilidad de $p(Y|e, \text{do}(X = x))$.

Resultados potenciales

Resultados potenciales

Suponga que recolectamos datos sobre el salario de distintas personas, siendo X los años de experiencia, D el nivel educativo, y S el salario.

Por simplicidad, supongamos que existen tres niveles educativos: $0 =$ media superior, $1 =$ universidad, $2 =$ posgrado.

Así, $S_0(u)$ representa el salario del individuo u si u cuenta con eduación media superior pero no con título universitario, y $S_1(u)$ representa el salario de u si u tuviera título universitario.

u	X	D	$S_0(u)$	$S_1(u)$	$S_2(u)$
Alice	6	0	81,000	?	?
Bert	9	1	?	92,500	?
Caroline	9	2	?	?	97,000
David	8	1	?	91,000	?
Ernest	12	1	?	100,000	?
Frances	13	0	97,500	?	?
⋮	⋮	⋮	⋮	⋮	⋮

Una típica pregunta contrafactual es "¿cuánto sería el salario de Alice si tuviera título universitario?" En otras palabras, cuánto vale $S_1(\text{Alice})$?

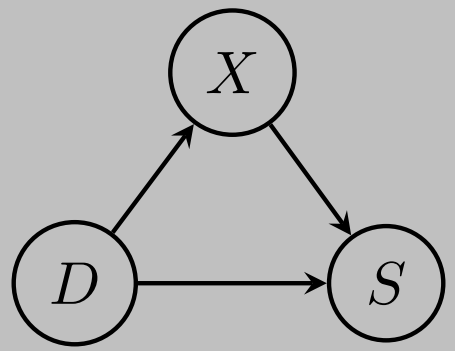
Suponga que decidimos modelar $S = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon_S$, y estimamos β_0, β_1 y β_2. El modelo queda de la siguiente manera:

$$\mathbb{E}[S|X, D] = 65,000 + 2,500X + 5,000D.$$

Haciendo un análisis de este modelo, podemos conlcuir que el salario de Alice, si tuviera título universitario, sería de

$$65,000 + 2,500 \times 6 + 5,000 \times 1 = 85,000.$$

Sin embargo, si Alice hubiera asistido a la universidad, no podría haber usado ese tiempo para ganar experiencia. Esta relación está representada por el siguiente DAG:



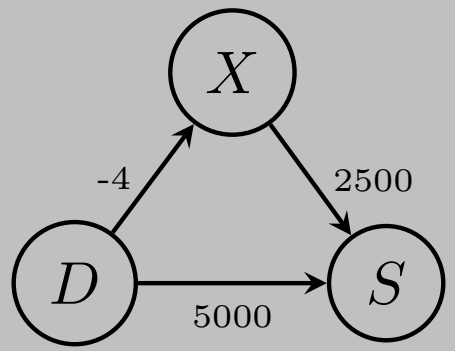
Podemos usar los mismos métodos estadísticos para hallar el mejor modelo lineal. El resultado sería como el anterior, excepto por una diferencia

$$S = 65,000 + 2,500X + 5,000D + \varepsilon_S.$$

También contamos con una ecuación estructural para X, que podría verse como:

$$X = 10 - 4D + \varepsilon_X.$$

Es decir, contamos con el siguiente SCM:



$$\mathfrak{C}: \begin{cases} D := \varepsilon_D, \\ X := 10 - 4D + \varepsilon_X, \\ S := 65000 + 2500X + 5000D + \varepsilon_S. \end{cases}$$

Abducción: Datos de Alice: $e : \{D = 0, X = 6, S = 81,000\}$.

$$\mathfrak{C}|e : \begin{cases} 0 = \varepsilon_D, \\ 6 = 10 - 4 \times 0 + \varepsilon_X, \\ 81,000 = 65000 + 2500 \times 6 + 5000 \times 0 + \varepsilon_S. \end{cases}$$

$$\Rightarrow \{\varepsilon_D = 0, \varepsilon_X = -4, \varepsilon_S = 1000\}$$

Acción:

$$\mathfrak{C}|\mathrm{do}(D = d) : \begin{cases} X := 10 - 4 \times d + \varepsilon_X, \\ S := 65000 + 2500 \times X + 5000 \times d + \varepsilon_S. \end{cases}$$

Predicción:

$$\mathfrak{C}|e, \mathrm{do}(D = 1) : \begin{cases} X := 10 - 4 \times 1 - 4 \\ \qquad = 2, \\ S := 65000 + 2500 \times 2 + 5000 \times 1 + 1000 \\ \qquad = 76000 \end{cases}$$