DAGs

2025-04-02

Plan

- Identificación
- Introducción a los DAGs
- Caminos causales
- Caminos buenos y malos
- Caminos Front Door y Backdoor
- Caminos abiertos y cerrados
- Selección muestral

Identificación

Consideremos un problema estándar en economía laboral: estimar los retornos de la educación.

Nuestra tarea es aislar el impacto causal de ir a la universidad.

Motivación

Imaginemos un modelo de acumulación de capital humano.

Los salarios dependen de la educación y las habilidades, mientras que la elección educativa está influenciada por la habilidad.

Tal vez obtengamos algo como:

$$W = f_w(E, A, \epsilon_w) = \alpha_1 E + \alpha_2 A + \epsilon_w$$

$$E = f_e(A, \epsilon_e) = \beta A + \epsilon_e$$

$$A = f_A(\epsilon_a) = \epsilon_a$$

Y todos los errores son independientes.

Motivación

Supongamos que la habilidad no es observable.

Si simplemente comparamos los salarios de personas con distintos niveles de educación, obtendremos un estimador sesgado del retorno de la educación.

Podemos usar el modelo para ver que $\frac{cov(W,E)}{var(E)} = \alpha_1 + \alpha_2\beta\frac{var(\epsilon_A)}{var(E)}$.

Motivación

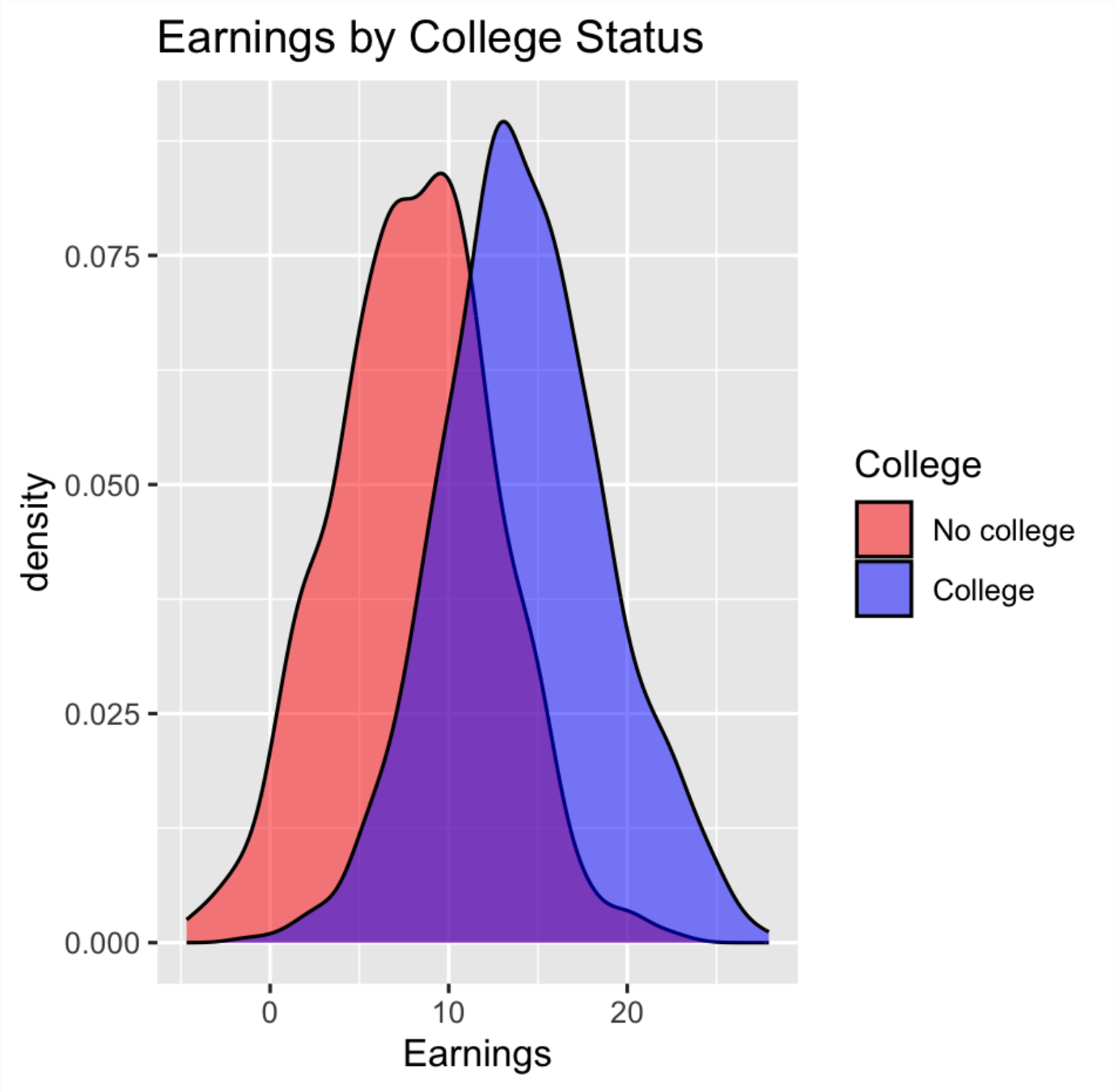
Ahora supongamos que sabemos que la habilidad depende de la genética y del entorno familiar. De hecho, todos los hermanos tienen la misma habilidad.

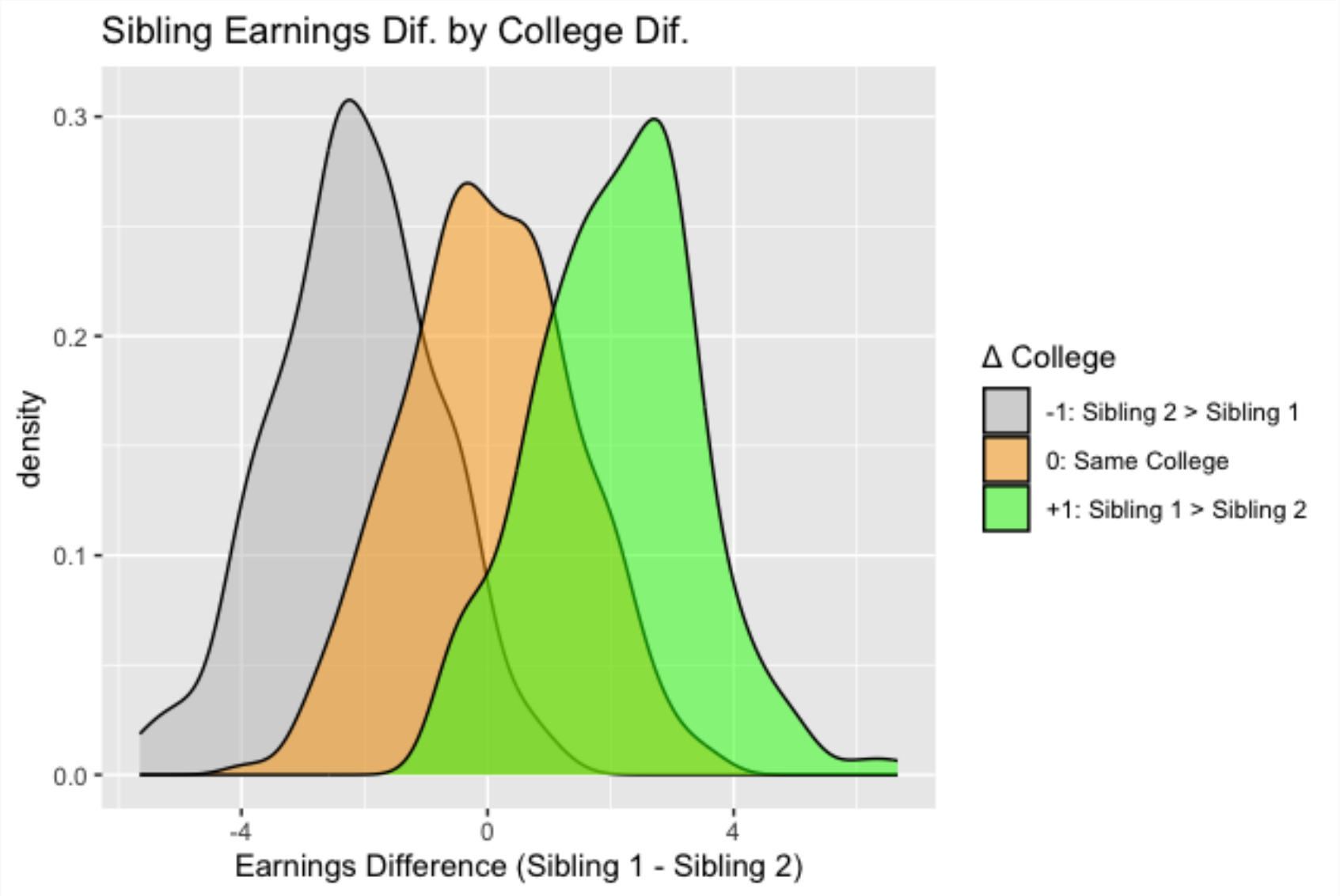
Basándonos en este conocimiento, suponemos que si comparamos dos hermanos con distintos niveles de educación, la habilidad no varía entre ellos. Entonces, cualquier diferencia sistemática entre ellos debe deberse a la universidad.

Simulemos algunos datos para hacerlo más concreto.

Resultados de la simulación

Variable	Ideal	Naive	Difference
college	2.065***	6.161***	
	(0.048)	(0.206)	
ability	4.9745***		
	(0.024)		
delta_college			2.136***
			(0.065)


Earnings by College Status
0.075
0.050
0.025
0.000
density
0
10
20
Earnings
College
No college
College


Sibling Earnings Dif. by College Dif.
density
0.3
0.2
0.1
0.0
-4
0
4
Earnings Difference (Sibling 1 - Sibling 2)
Δ College
-1: Sibling 2 > Sibling 1
0: Same College
+1: Sibling 1 > Sibling 2

Identificación

Usamos nuestro conocimiento del **proceso generador de datos** para elegir una comparasión que corresponde al parámetro teórico que nos interesa.

Identificación no es un proceso *estadístico*, sino que un proceso *conceptual*.

La estratégia de identificación no está en los datos, está en el plano teórico.

Identificación

El conocimiento (parcial) sobre el DGP (proceso generador de datos) puede provenir de diversas fuentes:

- Teoría económica: expectativas racionales, maximización de beneficios
- Otras ciencias: epidemiología, genética, meteorología
- Conocimiento contextual: cómo funcionan las políticas públicas, reglas administrativas
- Contexto histórico: patrones de desarrollo

Identificación

Lo que queremos aprender del **DGP** es esencialmente:

- Qué variables están causalmente relacionadas
- Qué variables no están causalmente relacionadas
- Qué variables están correlacionadas debido a causas comunes
- Qué variables debemos mantener fijas para identificar relaciones causales

Identificación

Podríamos escribir un modelo formal y trabajar a través de las ecuaciones.

Sin embargo, eso puede ser bastante tedioso y complicado.

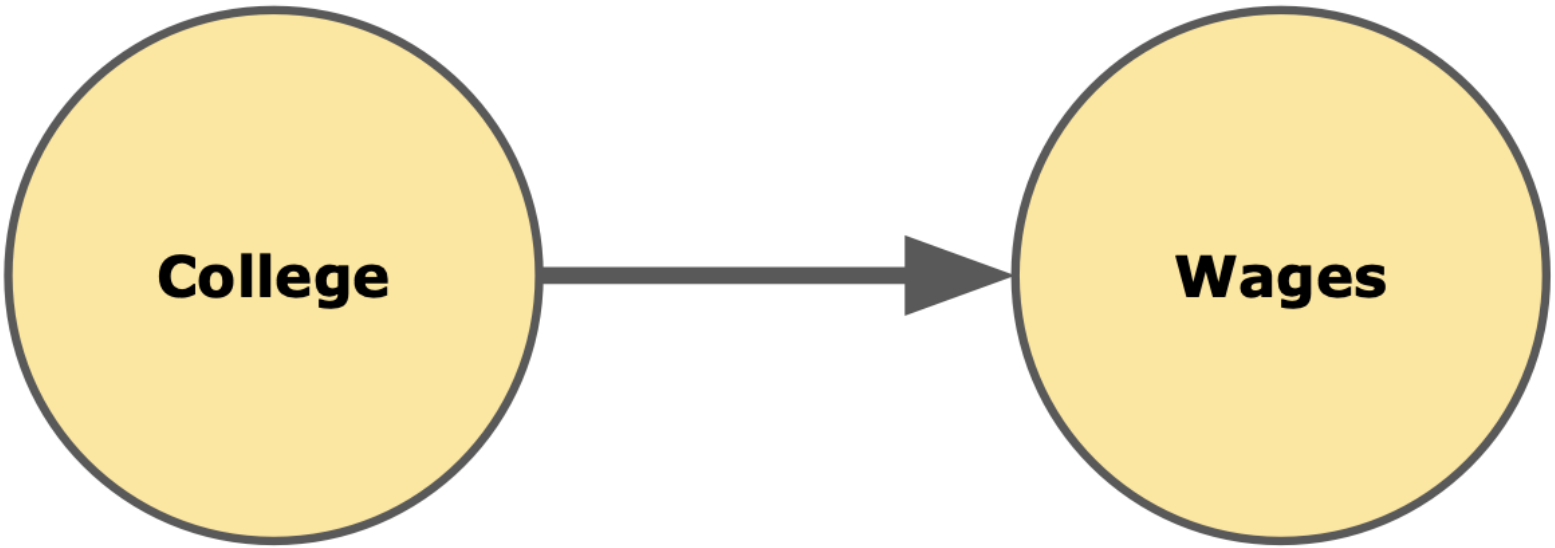
Nos interesa principalmente saber si existen relaciones causales y en qué dirección: podemos pensar en las formas funcionales y otros detalles más adelante.

Por eso, vamos a intentar resumir la información del DGP de una manera que sea más fácil de trabajar.

Introducción a los diagramas causales

Comencemos representando gráficamente una relación causal.

- Los nodos representan variables
- Las flechas entre ellos representan una relación causal



Introducción a los diagramas causales

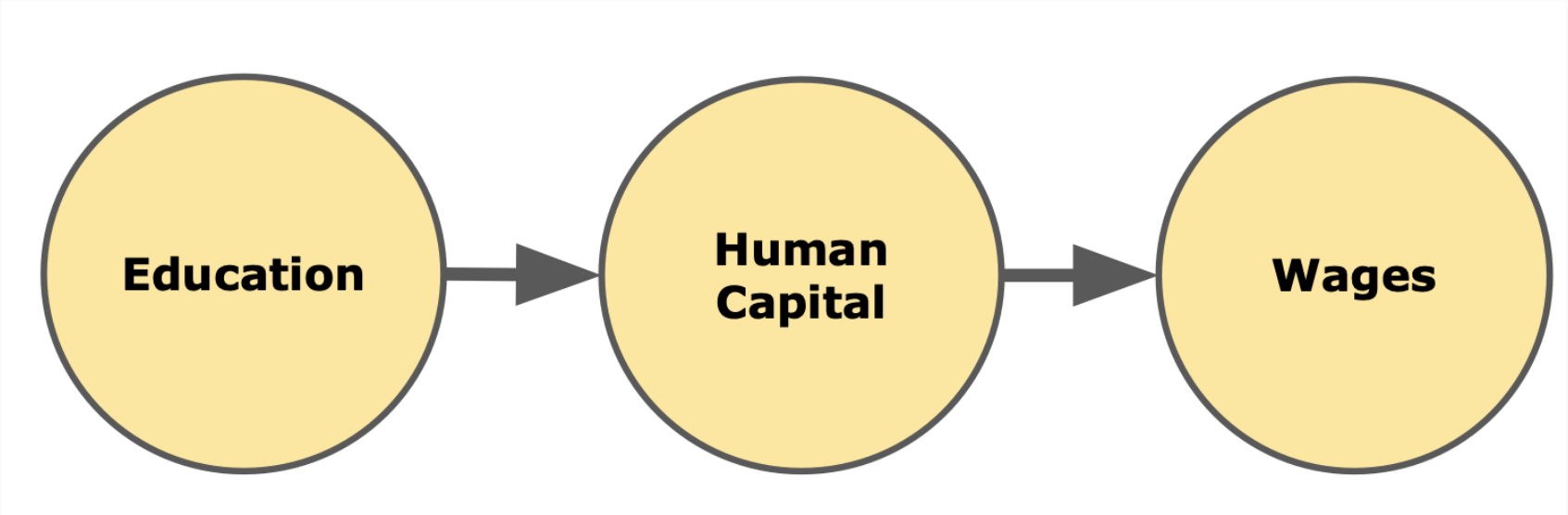
Cuando tenemos un diagrama con $X \rightarrow Y$ abstraemos de:

- La magnitud de la relación
- El signo de la relación
- La forma funcional
- Lo único que importa es que exista alguna relación causal.

Por supuesto, cualquier proceso generador de datos interesante tendrá más de dos variables.

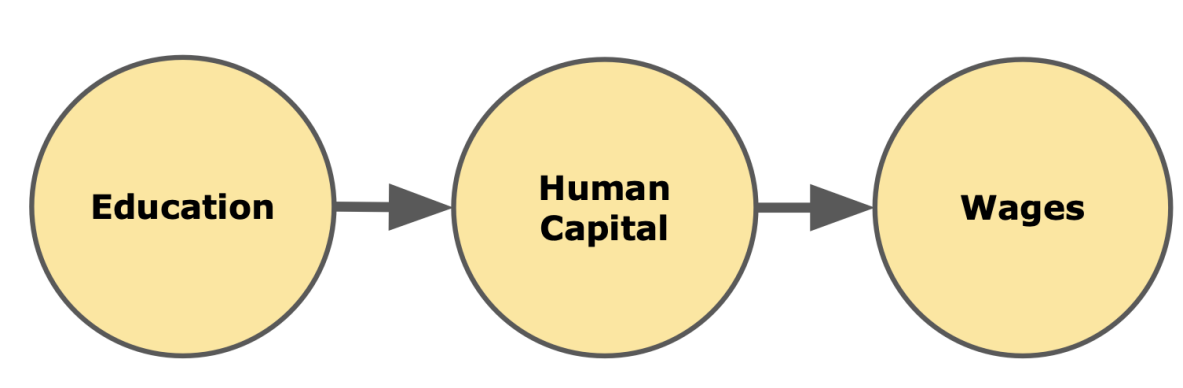
Introducción a los diagramas causales

Veamos este caso:



¿Qué nos dice este diagrama?

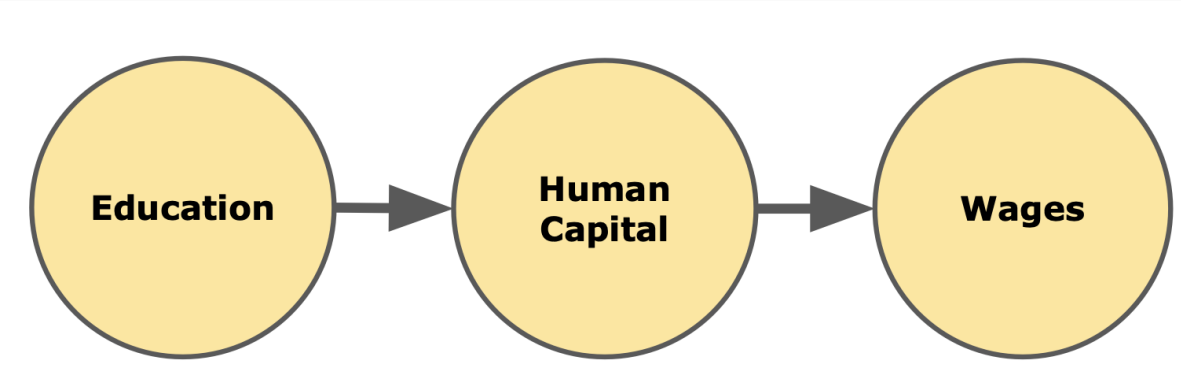
Introducción a los diagramas causales



¿Qué nos dice este diagrama?

- La educación causa capital humano
- El capital humano causa salarios
- La educación causa salarios indirectamente, pero no directamente

Introducción a los diagramas causales

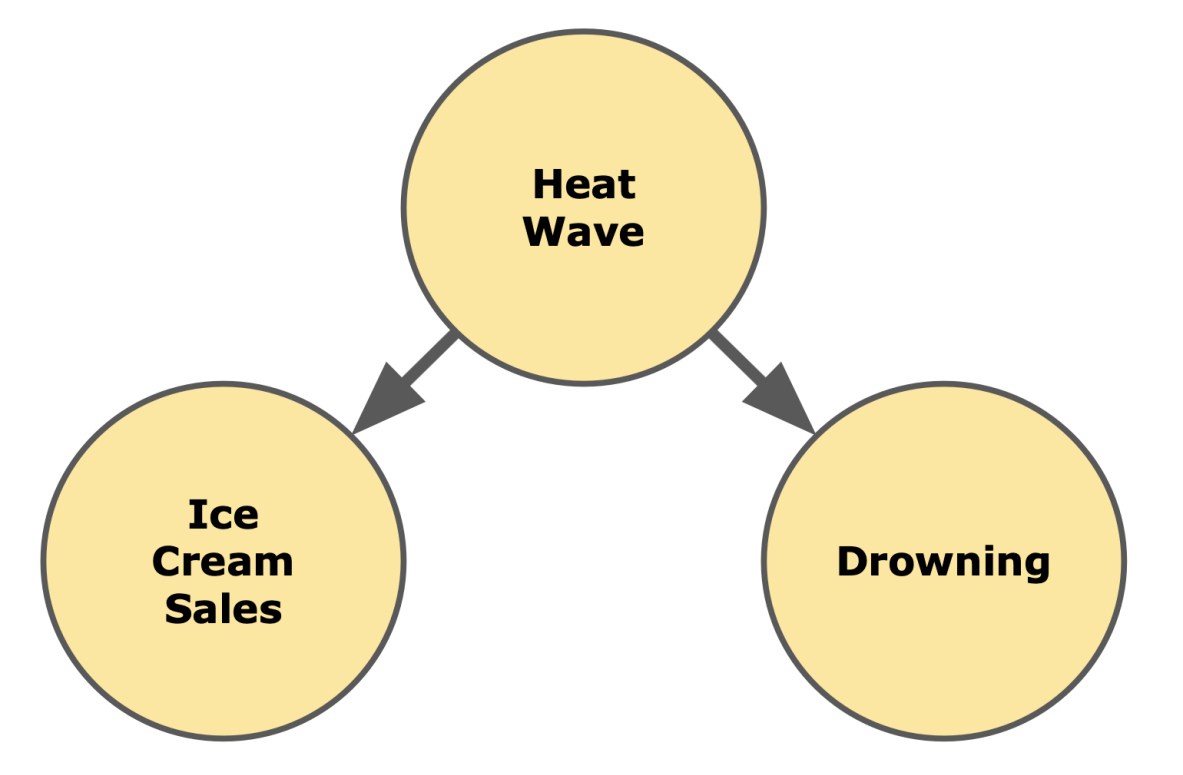


- Si mantenemos constante el capital humano, no debería haber correlación entre educación y salarios

$$Wages \perp Education | HumanCapital$$

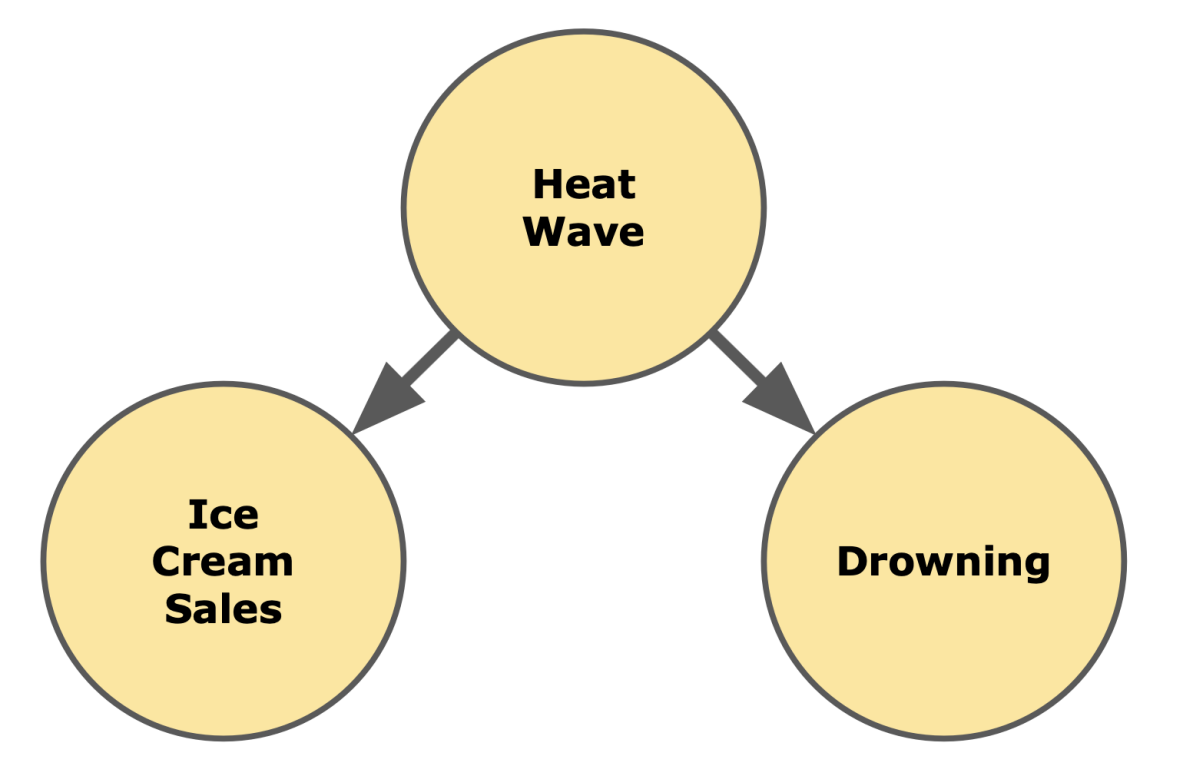
- Mantener constante el capital humano bloquea el camino de dependencia
- Podemos decir que el capital humano es un **mediador**

Introducción a los diagramas causales



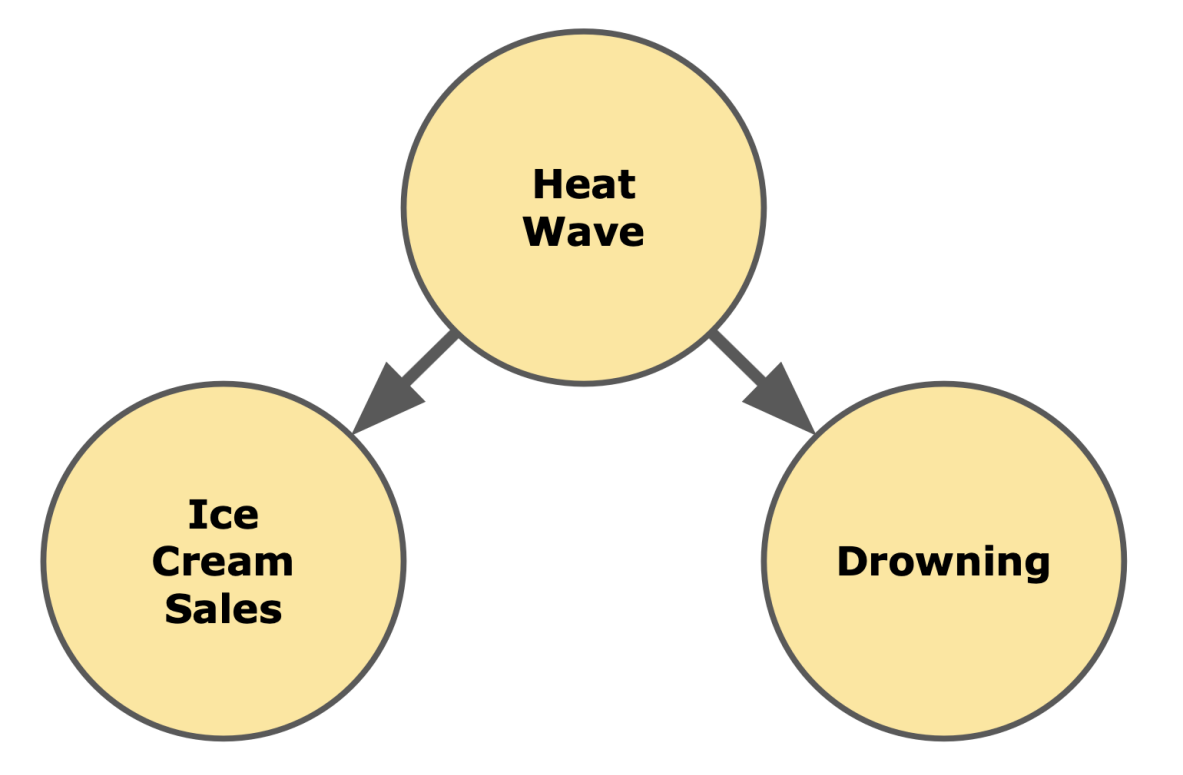
¿Qué tipo de correlaciones podemos esperar en este diagrama?

Introducción a los diagramas causales



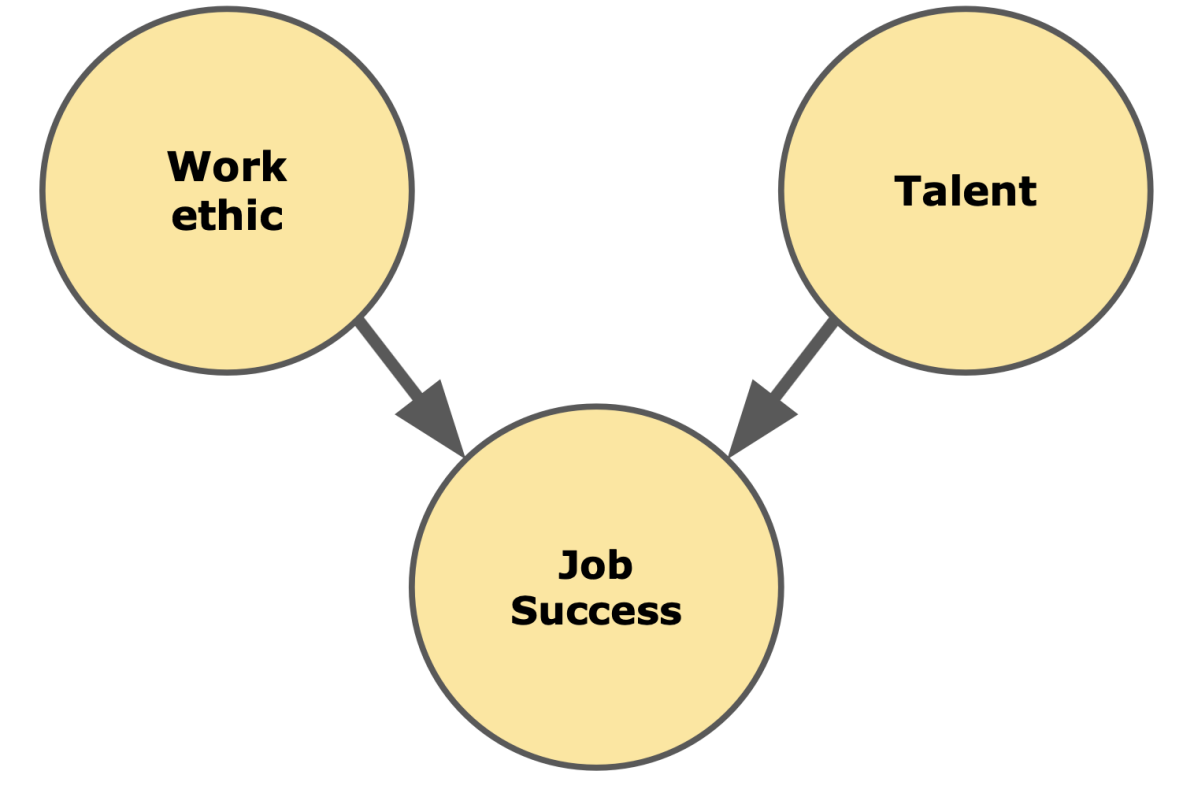
- Las ventas de helado están correlacionadas con los ahogamientos, porque tienen una causa común.
- Pero no están correlacionadas si condicionamos en la temperatura.

Intro to Causal Diagrams



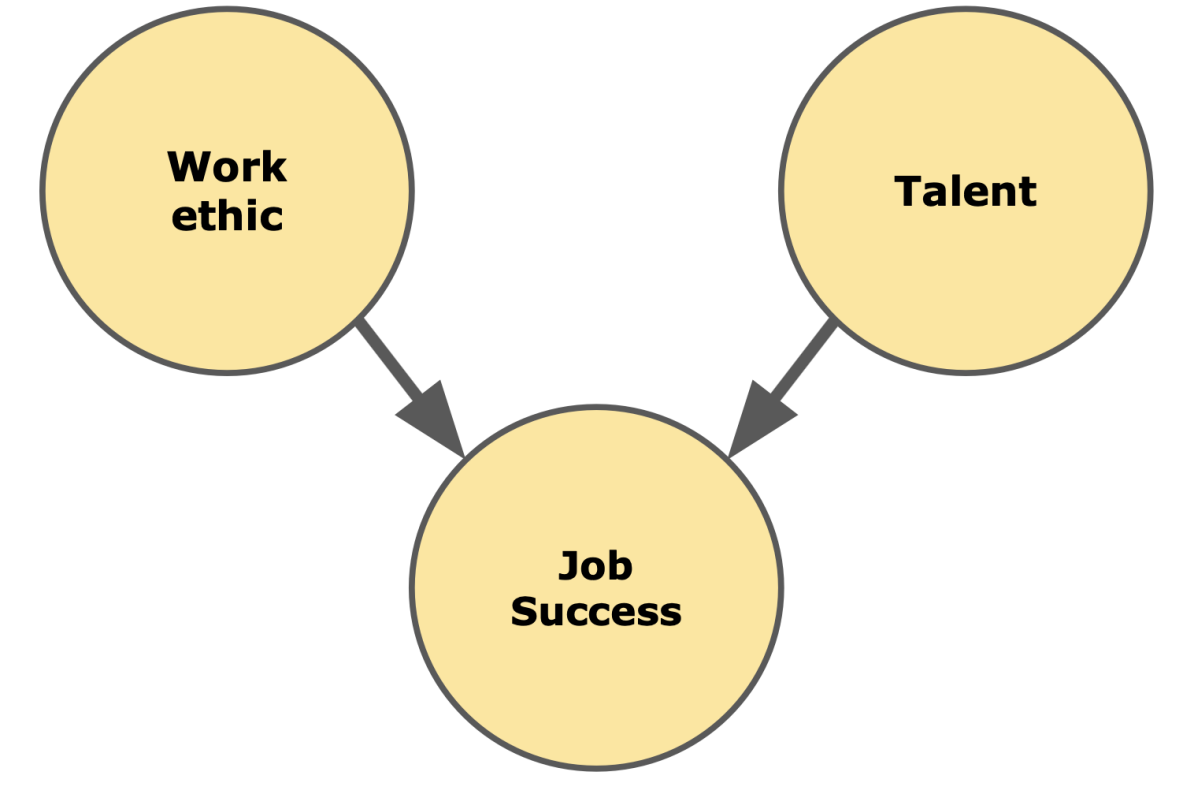
- La estructura de correlación es la misma que en el caso del mediador.
- Pero esta relación no es causal.
- Llamamos a esto una correlación espuria, o un camino de backdoor (Backdoor Path).

Intro to Causal Diagrams



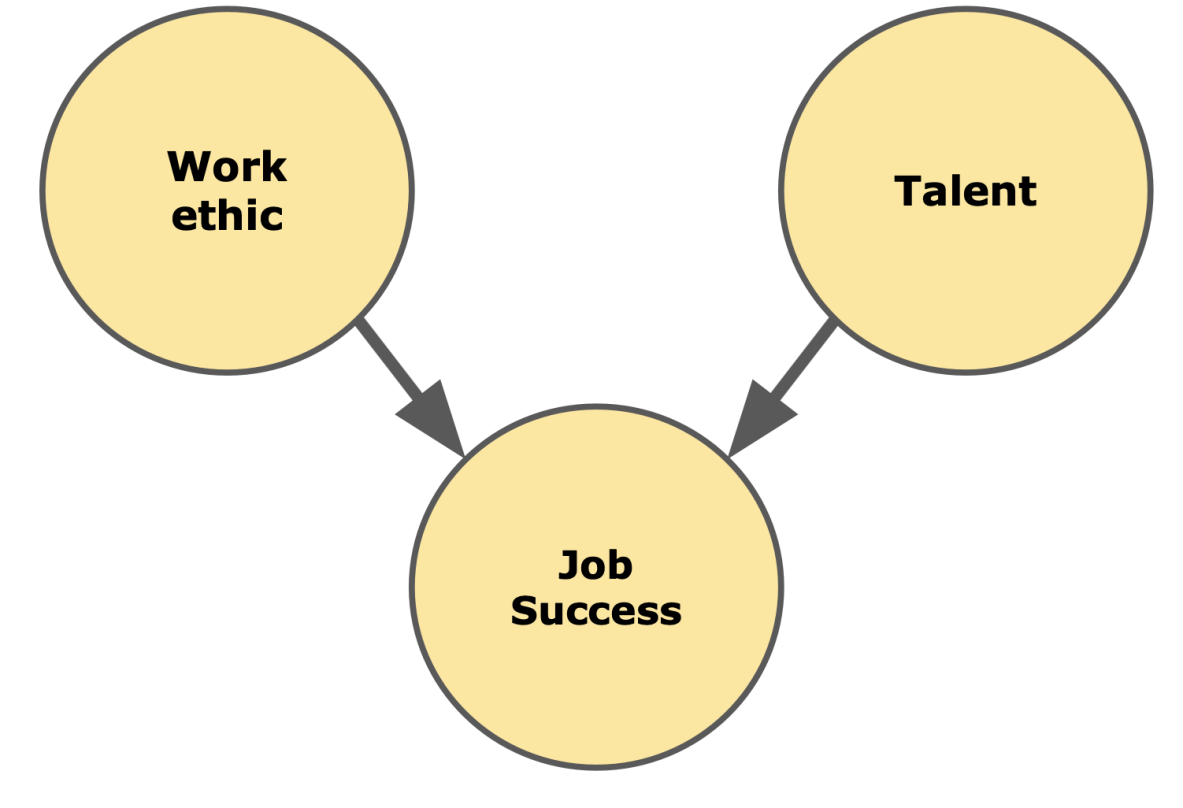
¿Qué tipo de correlaciones podemos esperar en este caso?

Intro to Causal Diagrams



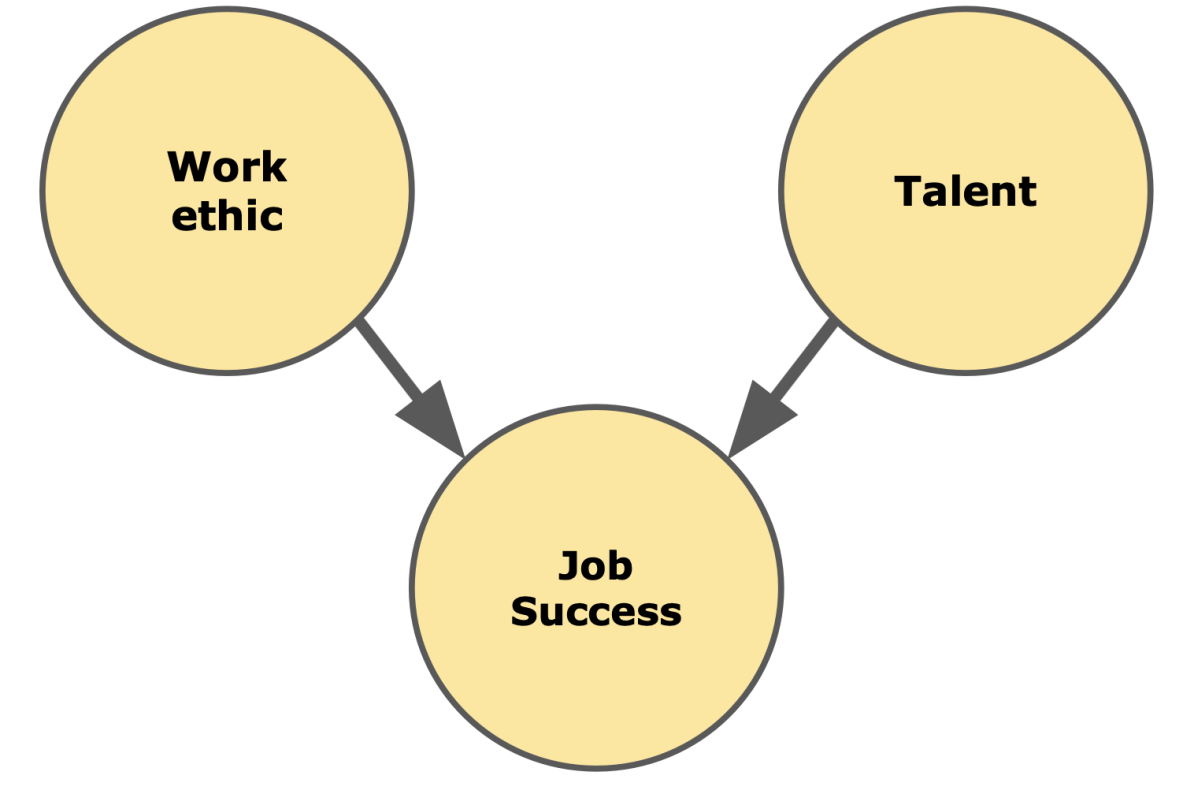
- El éxito laboral está correlacionado tanto con la ética de trabajo como con el talento.
- Pero la ética de trabajo y el talento son independientes.

Intro to Causal Diagrams



¿Qué pasa si condicionamos en el éxito laboral?

Intro to Causal Diagrams



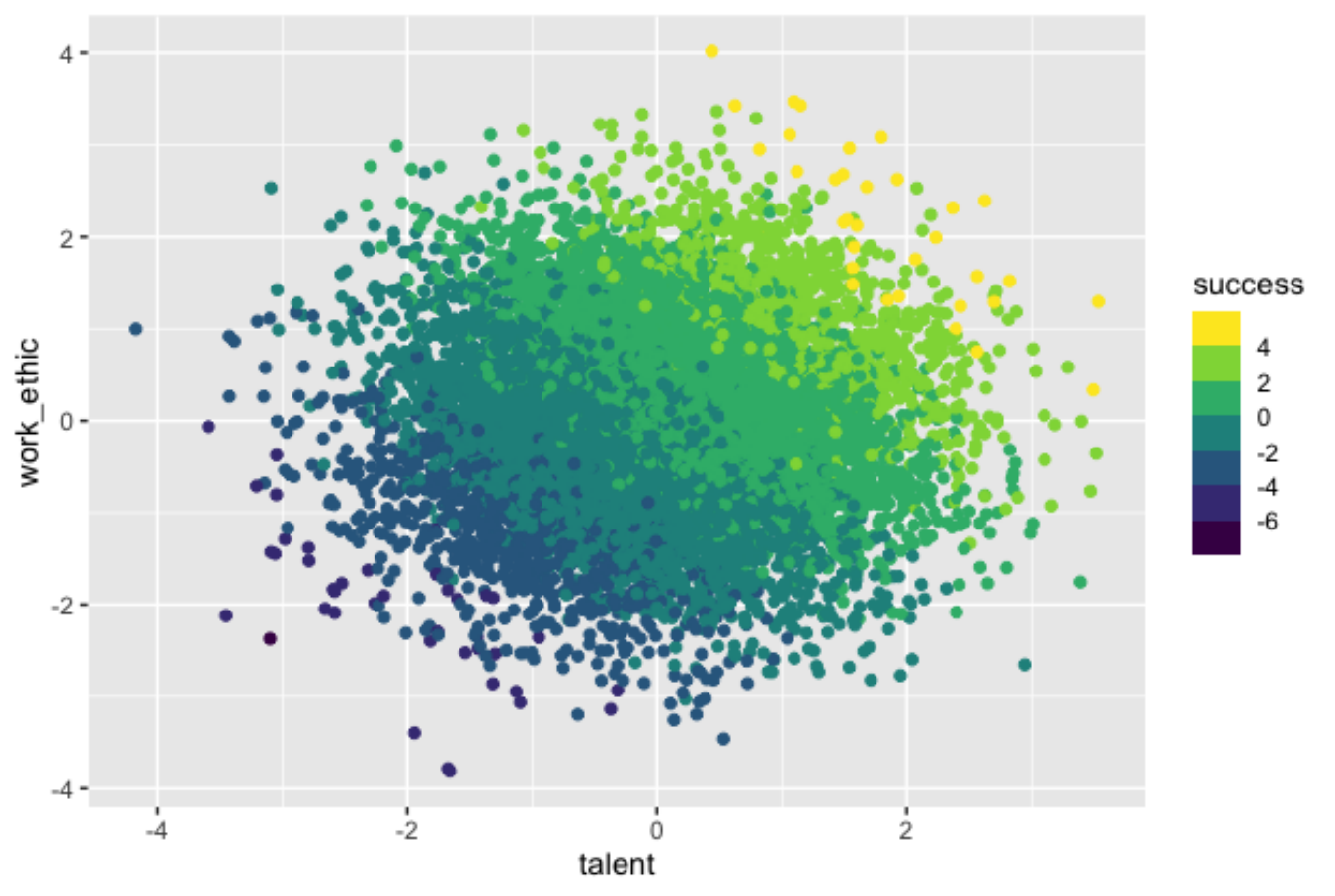
¿Qué pasa si condicionamos en el éxito laboral?

¡Si condicionamos, **creamos una correlación!**

Collider

Este diagrama causal se llama un colisionador (collider).

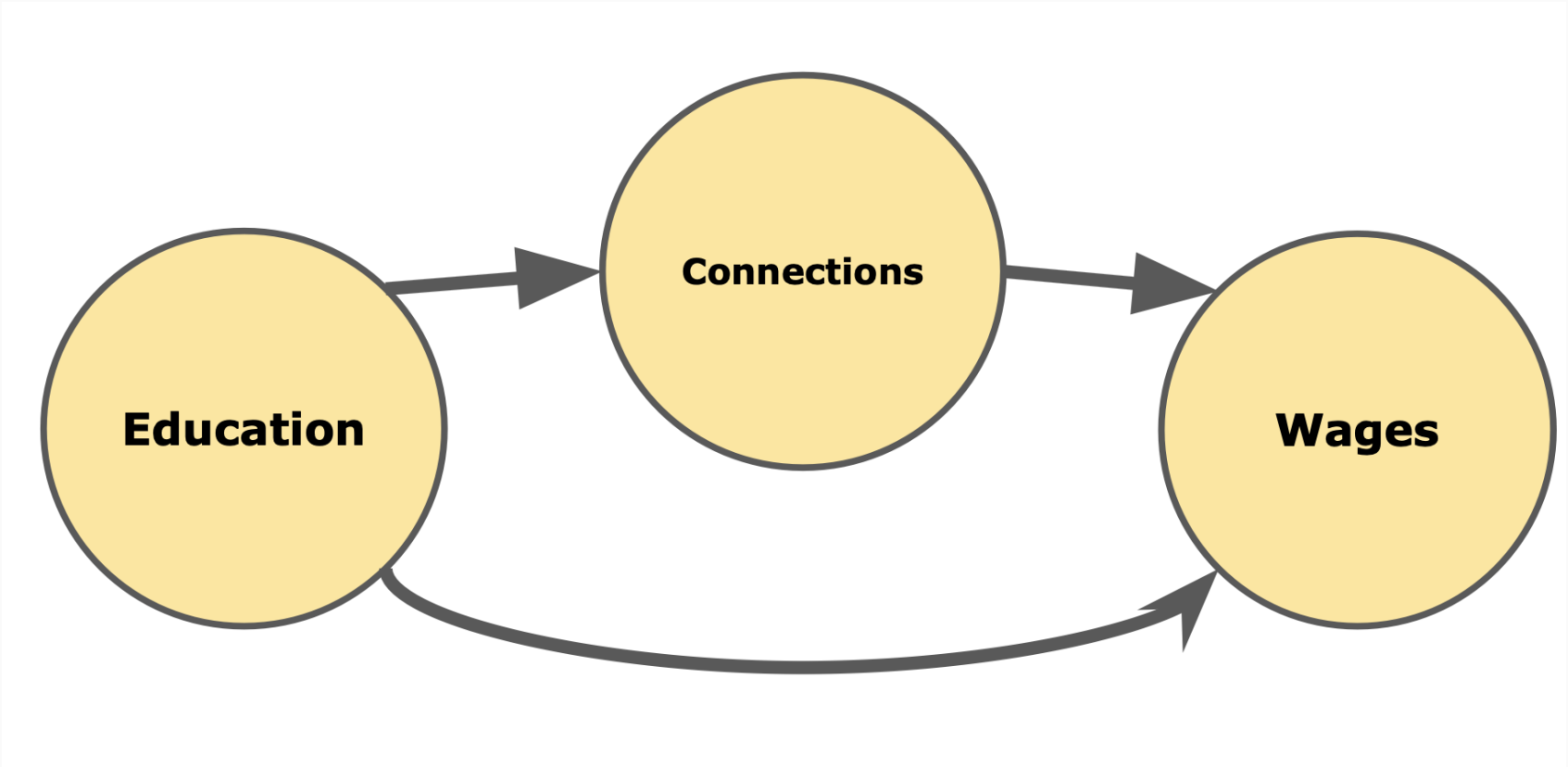
Condicionar en el colisionador genera lo que se llama sesgo del colisionador (*collider bias*): una forma de correlación espuria.



Tomemos cada una de estas tres estructuras y agreguemos otro vínculo causal.

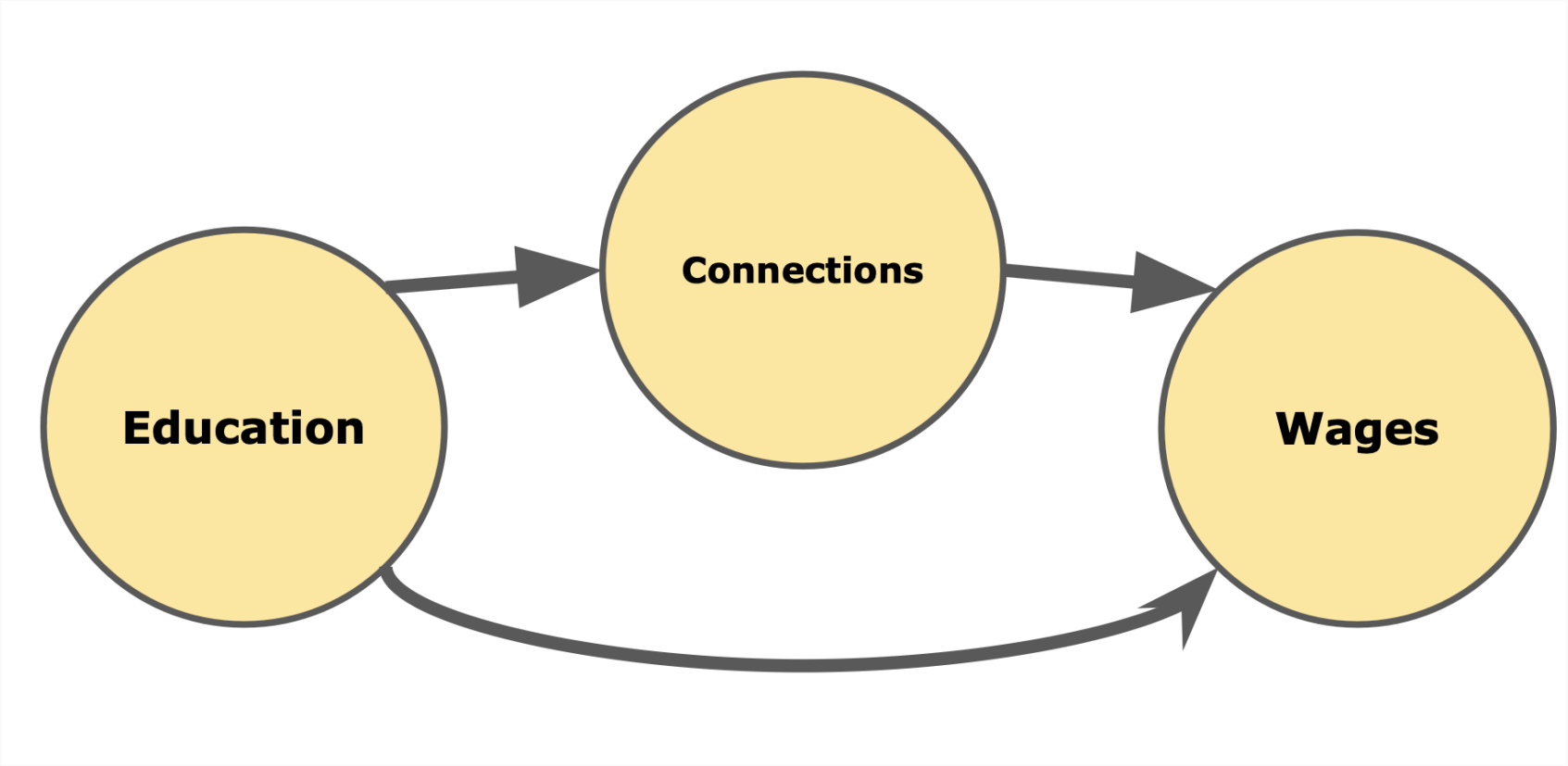
Nuestro objetivo es entender qué relaciones son causales, y cómo se correlacionan las variables, con y sin condicionamiento.

Introducción a los diagramas causales



¿Cuál es el efecto **causal** de la educación sobre los salarios?

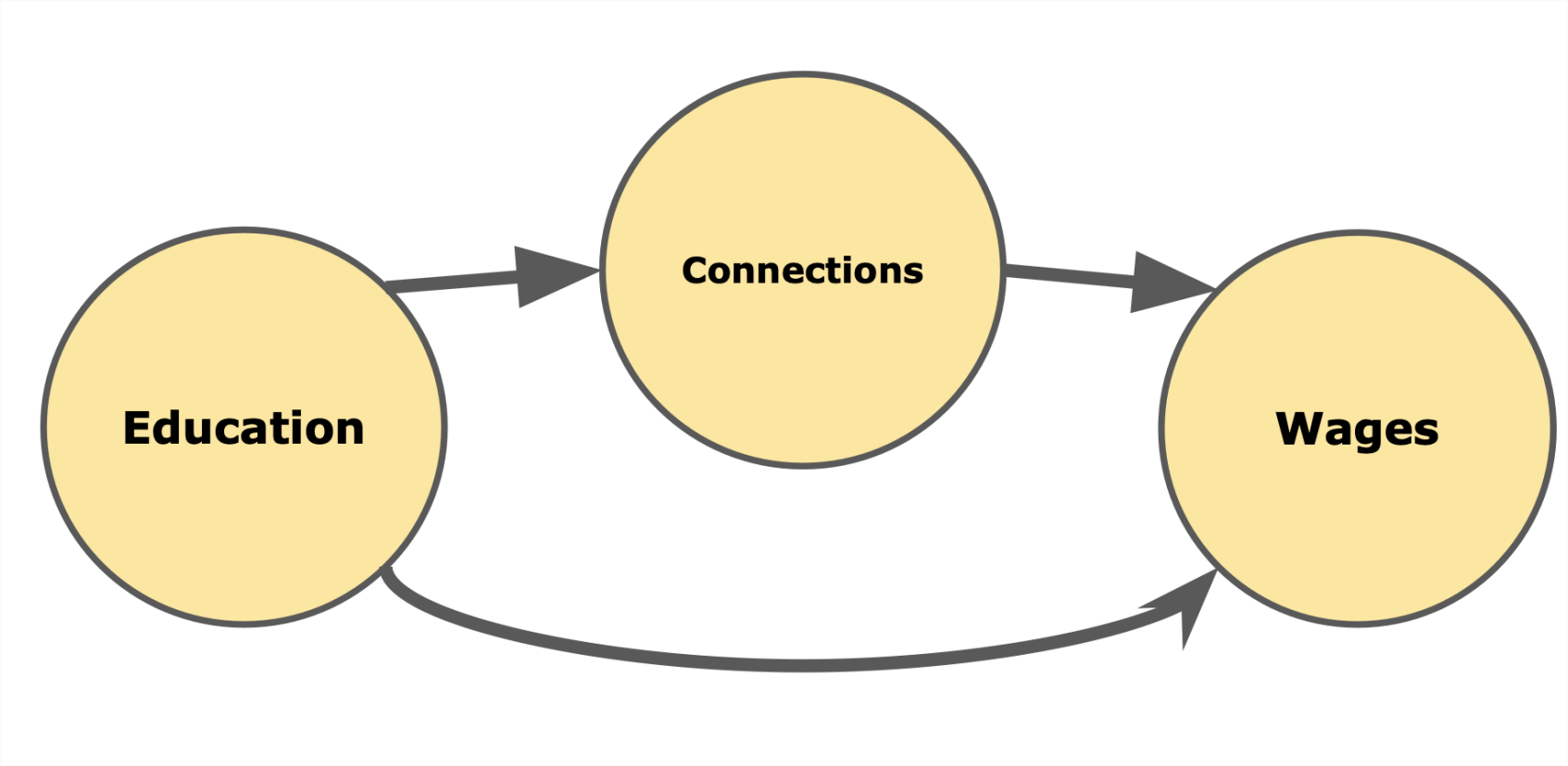
Introducción a los diagramas causales



El efecto causal consiste tanto en el efecto directo como en el efecto indirecto a través de las conexiones.

Podemos estimar el efecto causal simplemente observando la relación entre las dos variables.

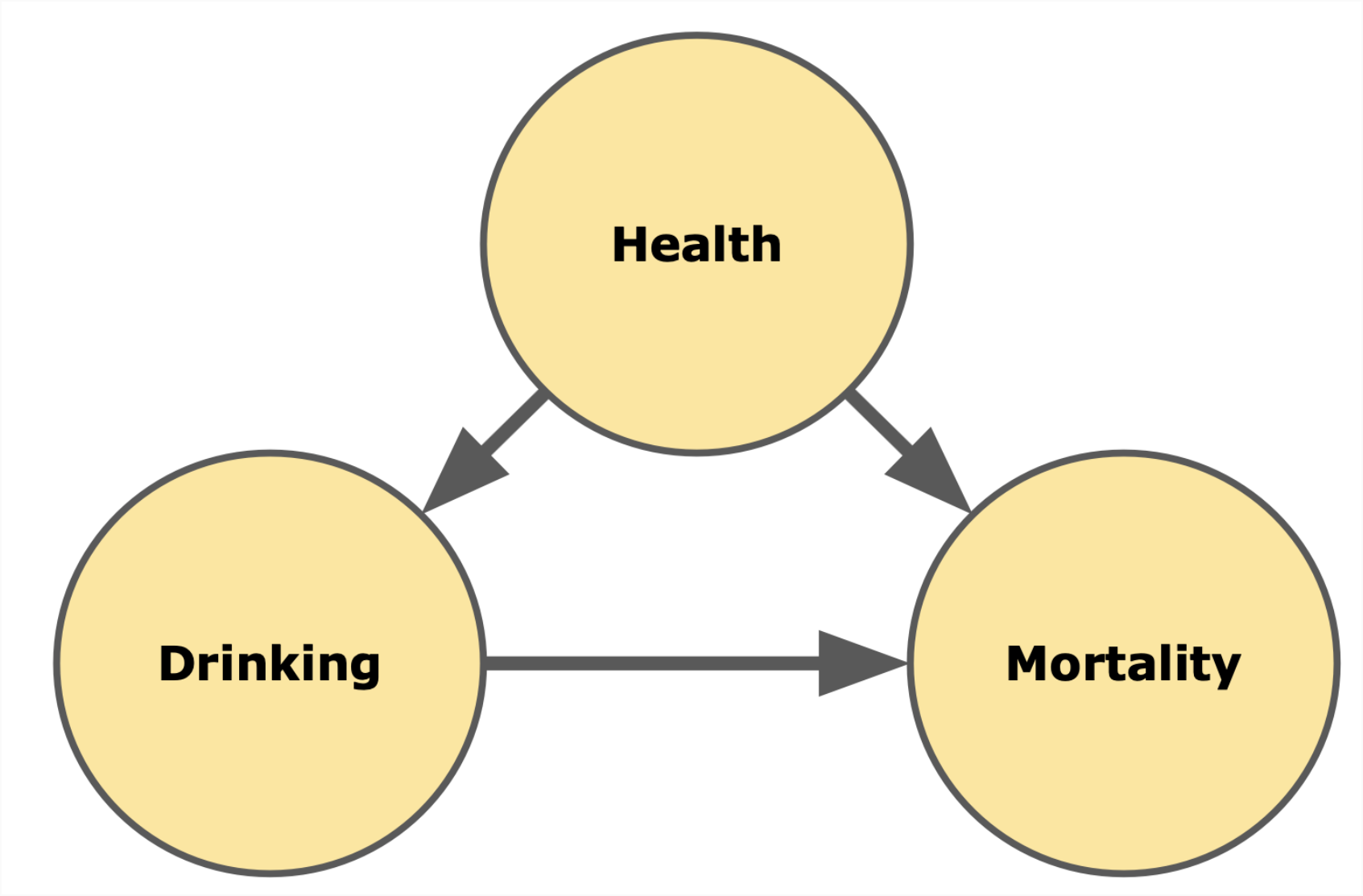
Introducción a los diagramas causales



Sin embargo, puede que nos interese estudiar un mecanismo específico de forma aislada.

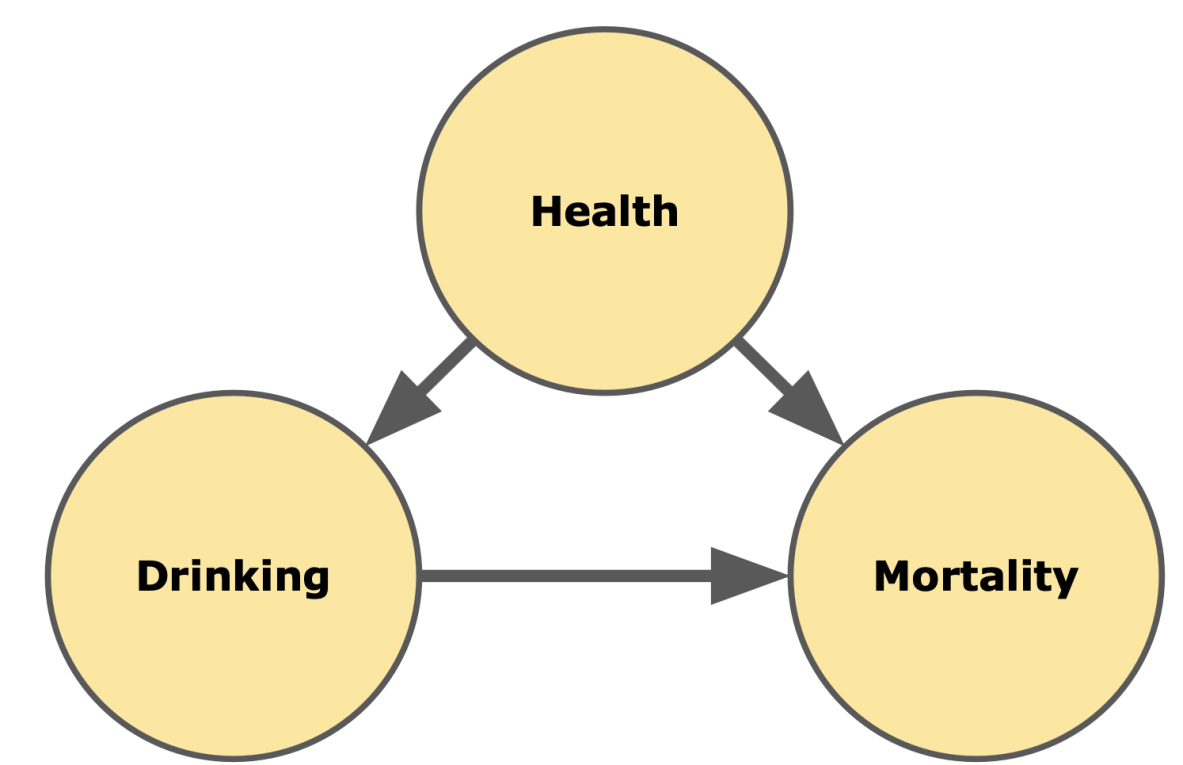
Si condicionamos en las conexiones, podemos obtener el **efecto directo puro**.

Introducción a los diagramas causales



¿Qué deberíamos esperar aquí?

Introducción a los diagramas causales

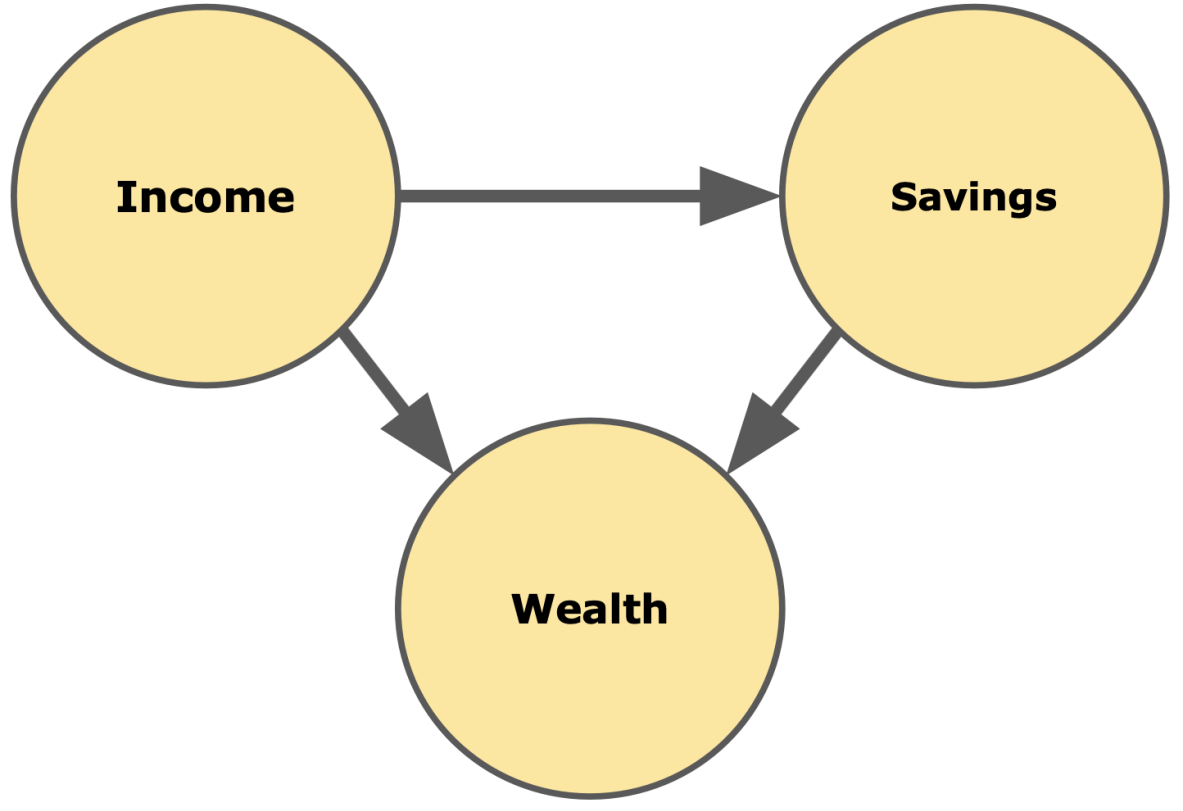


Este es el clásico **confusor (confounder)**.

- Salud crea una correlación espuria entre el alcohol y mortalidad.
- Si no condicionamos en la salud, obtendremos una mezcla del efecto directo y de la correlación espuria.

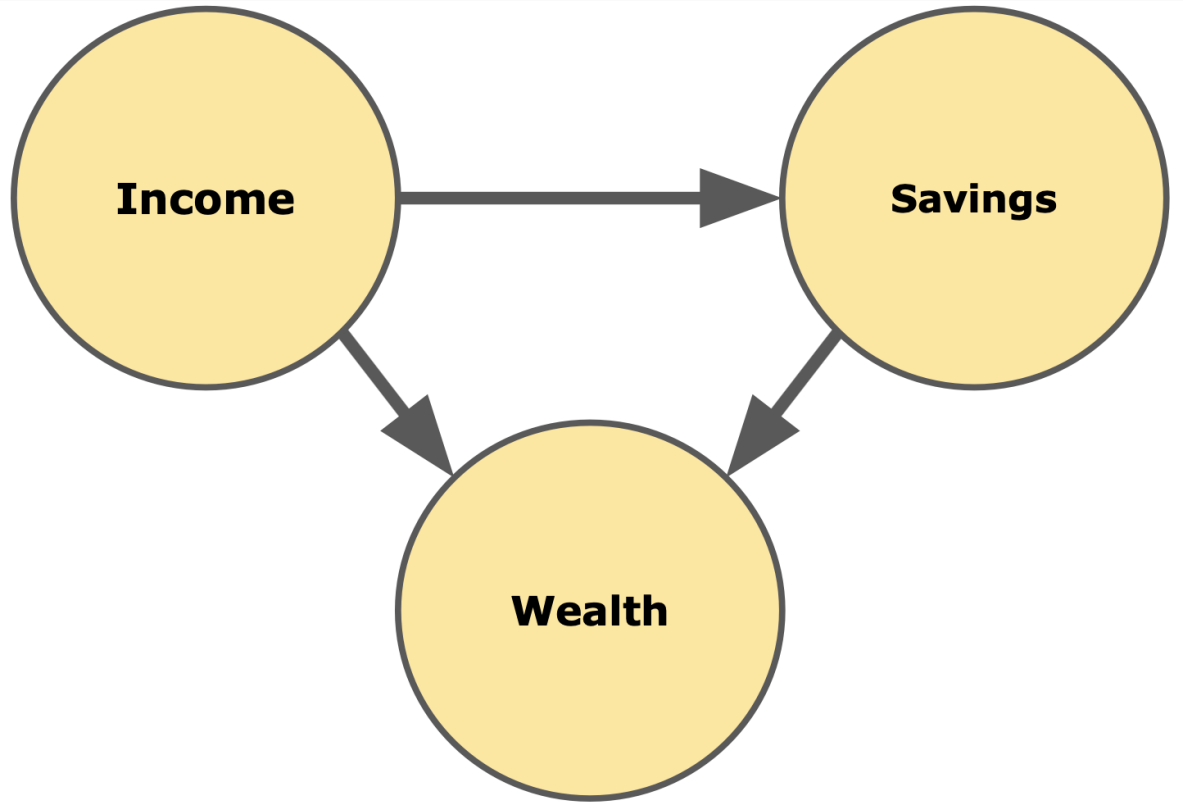
Introducción a los diagramas causales

Queremos estudiar si un ingreso más alto lleva a un mayores ahorros, pero nos preocupa el papel del patrimonio total.



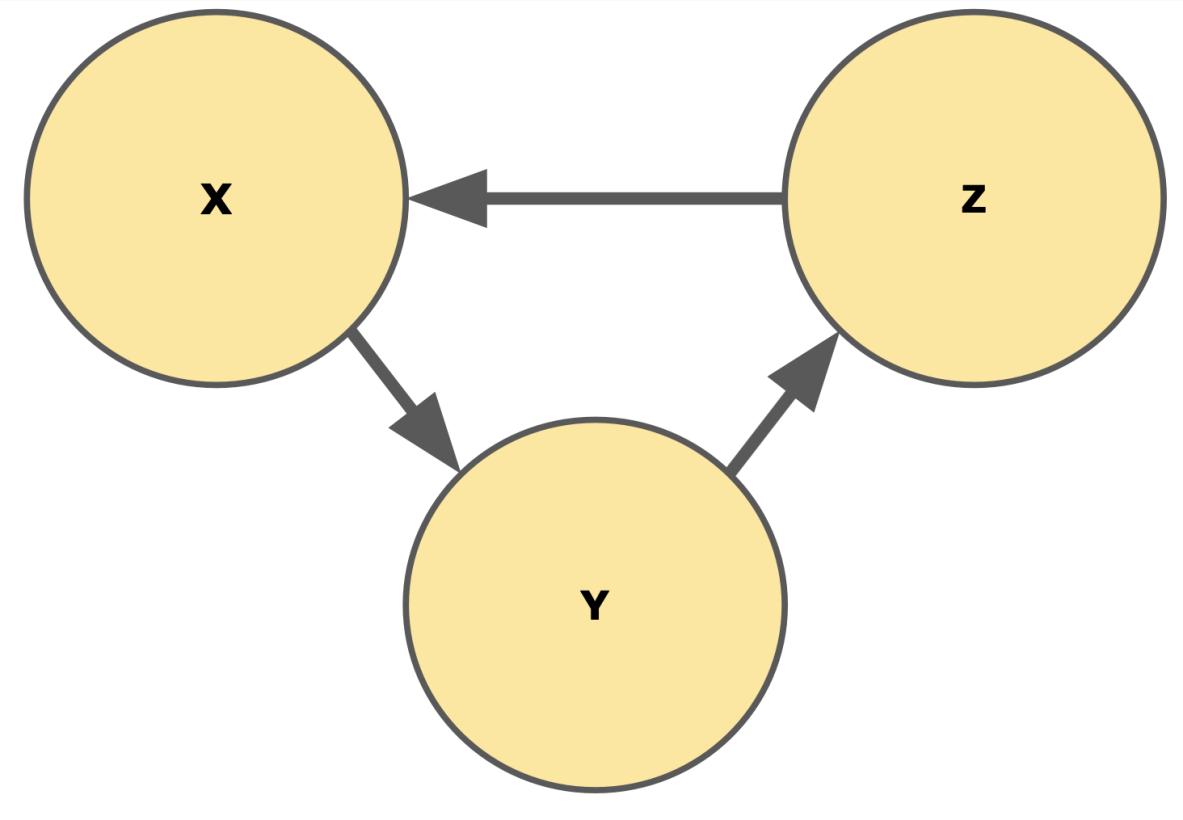
¿Qué sugiere este diagrama que deberíamos hacer?

Introducción a los diagramas causales



- Podemos estimar la relación directamente.
- El patrimonio no es un confusor, sino un colisionador.
- Si condicionamos en el patrimonio, creamos un sesgo.

Introducción a los diagramas causales



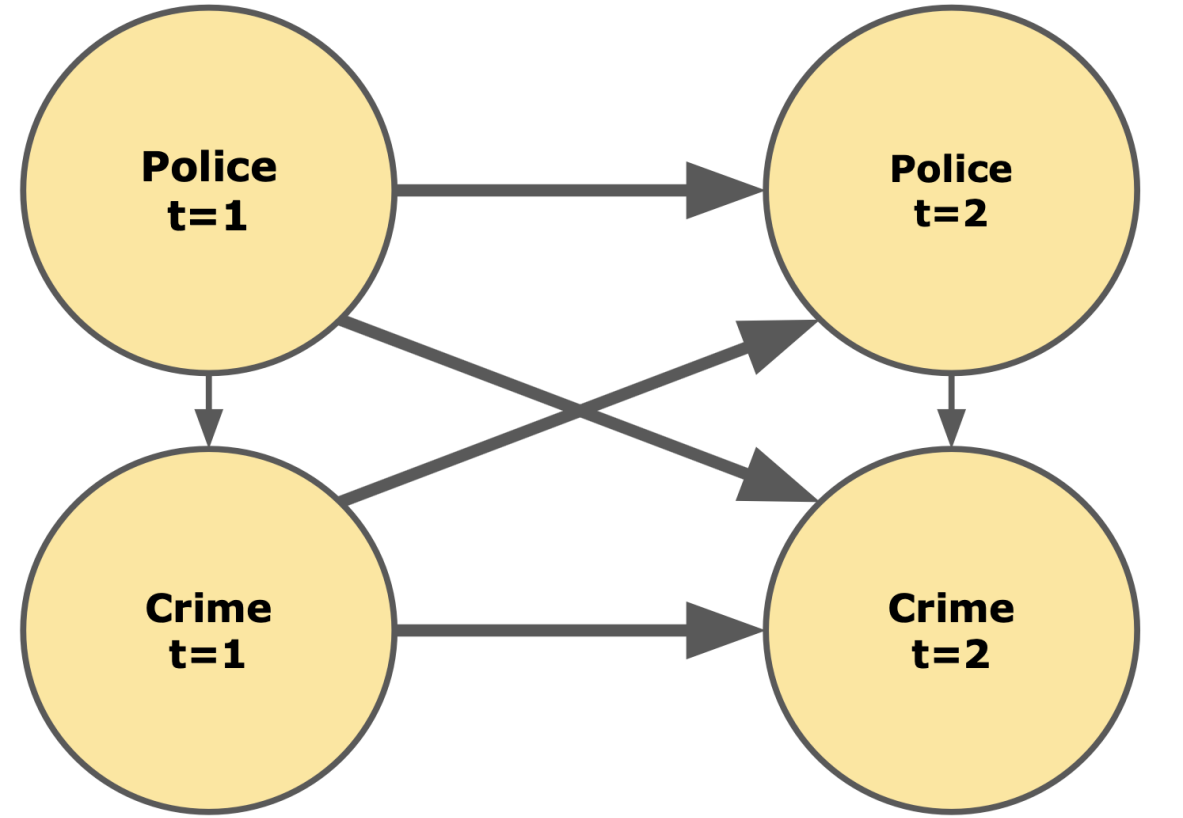
La última estructura posible con 3 nodos y 3 flechas es el ciclo.

Para trabajar con diagramas causales, **NO PODEMOS** tener ciclos.

Intro do DAGs

¿Cómo podemos tratar con modelos que incluyen retroalimentaciones (feedbacks)?

Una opción es incorporar explícitamente el tiempo:



Introducción a los diagramas causales

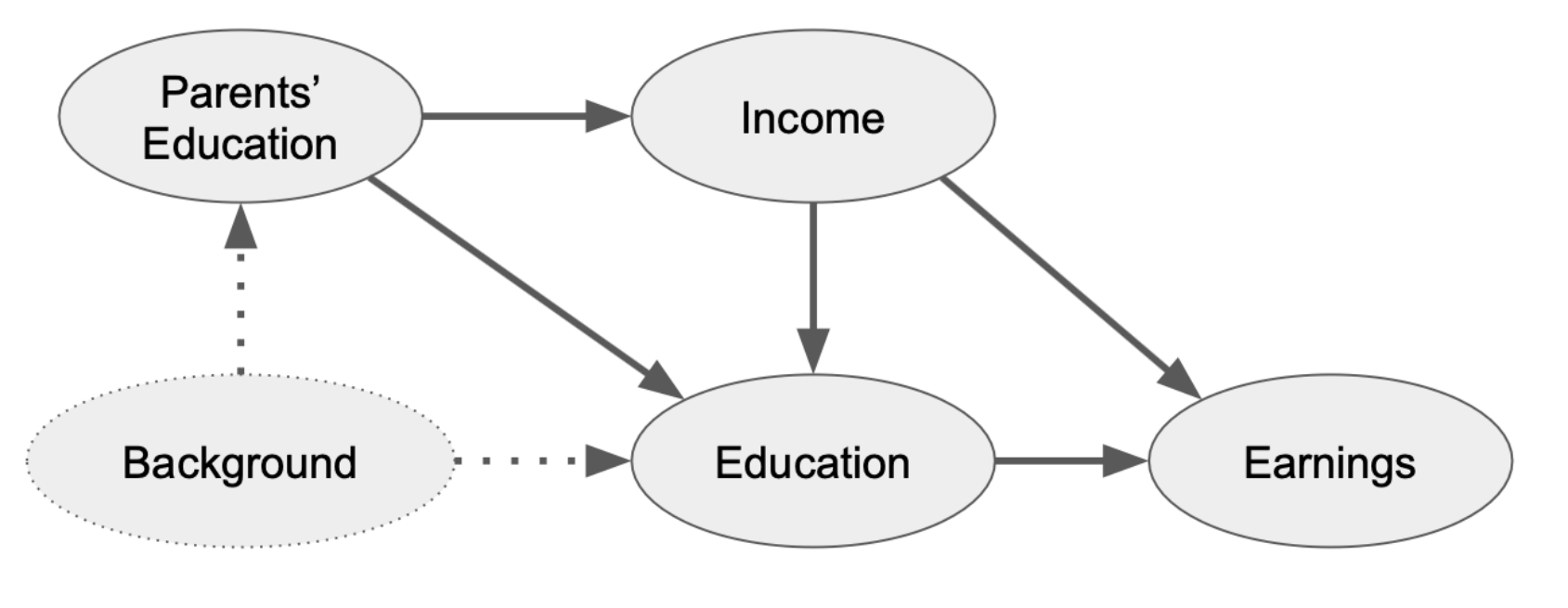
Vamos a usar esta herramienta para representar el DGP.

- Es importante incluir todas las variables relevantes.
- Igualmente importante es pensar en qué variables no están conectadas.
- La independencia es una suposición fuerte.

Vamos a denotar los factores no observables con líneas punteadas.

Introducción a los diagramas causales

¿Qué historia nos cuenta esta figura?



Introducción a los diagramas causales

Todos caminos posibles desde Educación hasta Salários

- Educación $\rightarrow$ Salarios
- Educación $\leftarrow$ Ingresos $\rightarrow$ Salarios
- Educación $\leftarrow$ Educ. Parental $\rightarrow$ Ingresos $\rightarrow$ Salarios
- Educación $\leftarrow$ Antecedentes $\rightarrow$ Educ. Parental $\rightarrow$ Ingresos $\rightarrow$ Salarios

Good and Bad Paths

Toda investigación comienza con una pregunta de interés.

Los caminos causales "buenos" son los que corresponden a tu pregunta:

- Pueden ser todos los caminos causales de una variable a otra
- O puede ser que te interese solo uno en particular

Los caminos "malos" son aquellos que generan correlaciones espurias, no relacionadas con tu pregunta.

Caminos Front-Door y Backdoor

- Un camino Front-Door solo tiene flechas desde la variable de interés hacia el resultado
  - Representa una asociación verdaderamente causal
- Un camino Backdoor incluye flechas en dirección contraria
- Educación $\leftarrow$ Ingresos Familiares $\rightarrow$ Salarios es un camino de backdoor
  - Los caminos de backdoor crean correlaciones espurias si están abiertos
- Si cerramos todos los caminos de backdoor, hemos identificado el efecto causal.

Caminos abiertos y cerrados

- ¿Cómo tratamos con los caminos de backdoor?
  - ¡Los cerramos!
- Esto significa que intentamos detener la cadena causal
- Hay dos formas principales:
  - Control estadístico
  - Un camino también se cierra si tiene un **colisionador**

Caminos abiertos y cerrados

La forma principal de cerrar un camino es controlar al menos una de las variables en ese camino

- Esto significa usar estadística para "mantenerla constante"
- Podemos usar controles en regresiones u otros métodos similares.

Caminos abiertos y cerrados

Un colisionador es una variable que recibe flechas de ambos lados:

- El camino $X \to Z \to Y$ es directo
- El camino $X \to Z \leftarrow Y$ está bloqueado

Advertencia: Controlar un colisionador abre este camino bloqueado

Example 1

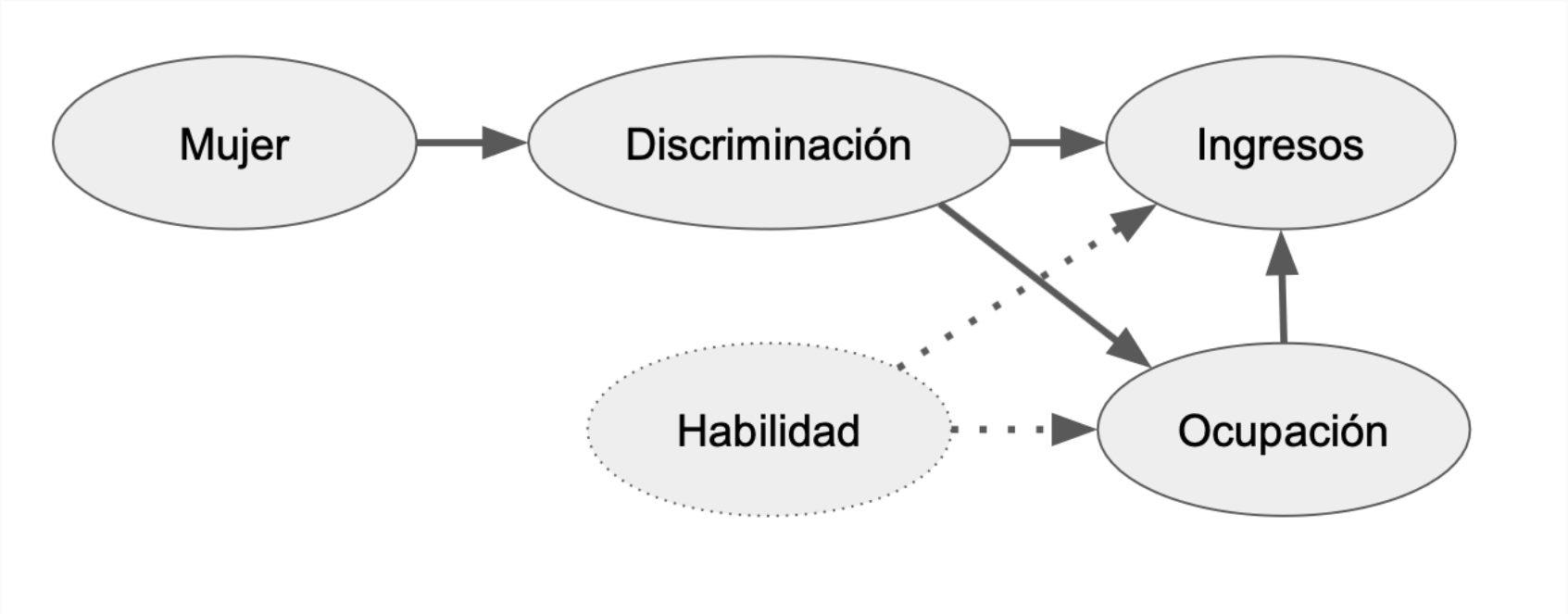
Existe un gran debate sobre la discriminación de género en el mercado laboral, principalmente en el setor de tecnonología.

Google, por ejemplo, fue acusado de ofrecer *menores salários* para sus trabajadoras mujeres.

Google dice que, cuando se compara hombres y mujeres con el mismo nivel de puesto, desempeño, antiguedad, y ubicación, sus salarios son basicamente idénticos.

Example 1

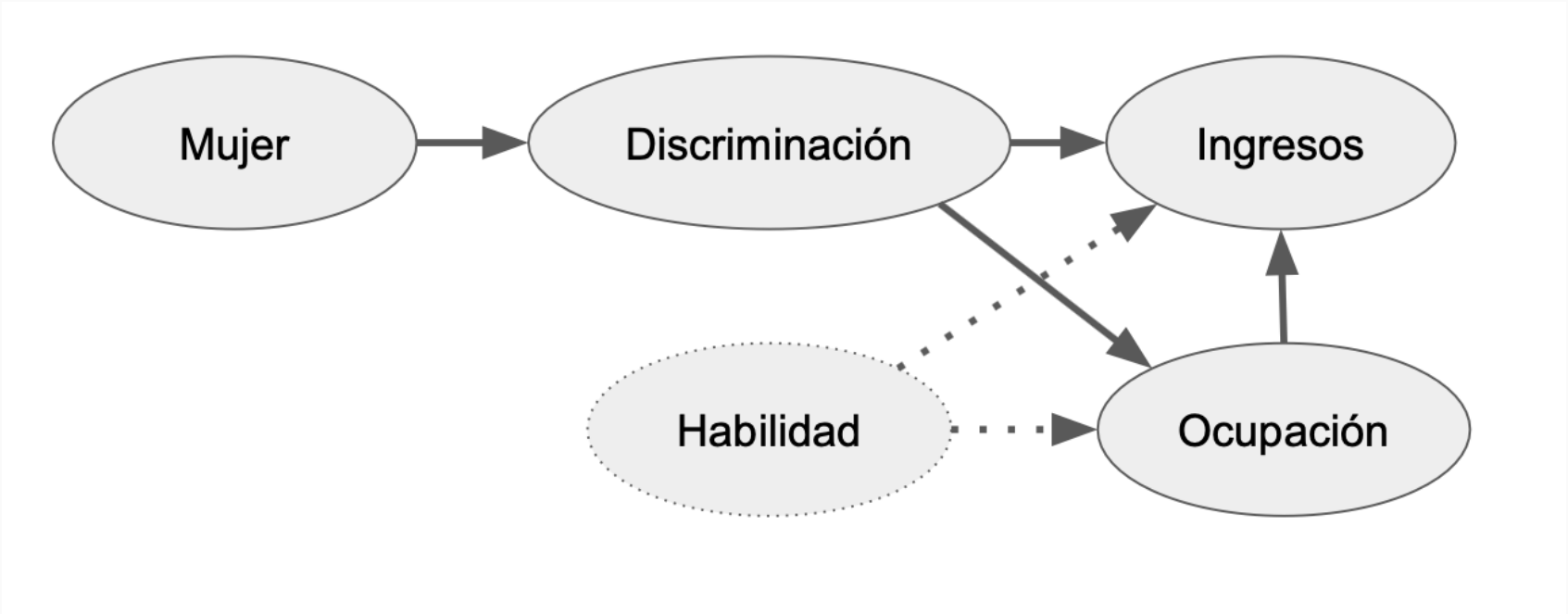
Vamos trabajar con ese modelo.



Cuales son los caminos causales desde “Mujer” hacia “Ingresos”?

Example 1

Vamos trabajar con ese modelo.



¿Que va pasar cuando controlamos por Ocupación?

Ejemplo 1

- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\leftarrow$ Habilidad $\rightarrow$ Ingresos

Ejemplo 1

- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\leftarrow$ Habilidad $\rightarrow$ Ingresos

Los dos primeros son caminos causales "buenos". Representan dos formas diferentes en que la discriminación afeta los ingresos.

El tercero camino es un Backdoor Path, pero está cerado porque Ocupación es un collider.

Ejemplo 1

- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\rightarrow$ Ingresos
- Mujer $\rightarrow$ Discriminación $\rightarrow$ Ocupación $\leftarrow$ Habilidad $\rightarrow$ Ingresos

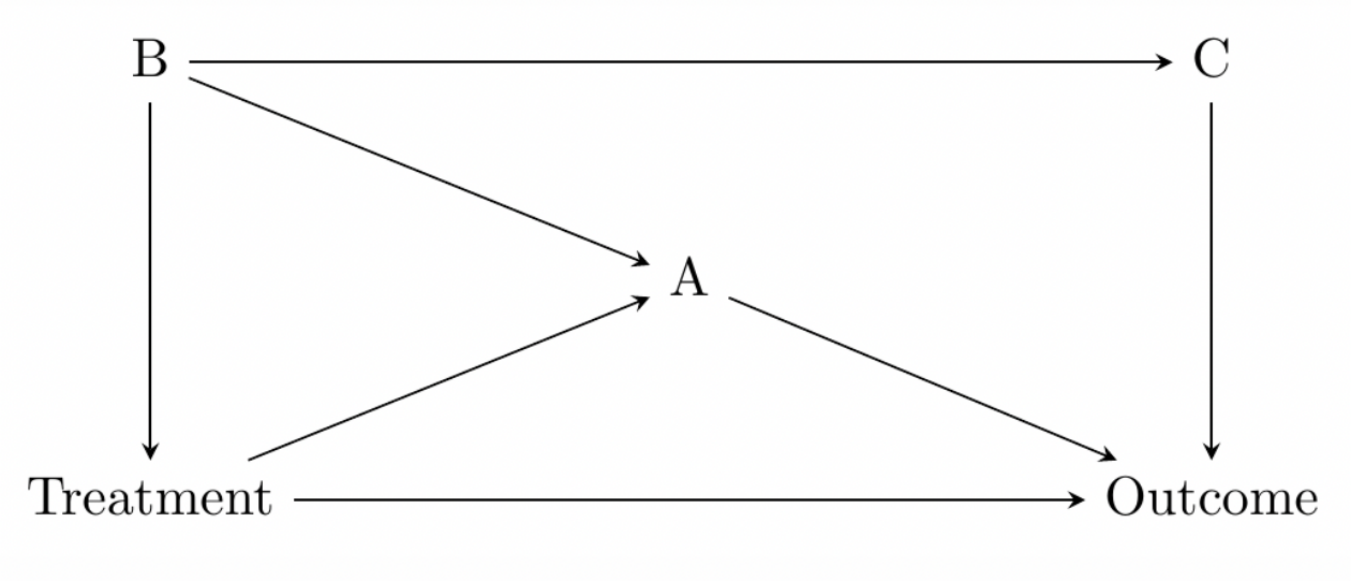
Si condicionamos a ocupación, ceramos el segundo camino, pero abrimos el tercero.

Ejemplo 1

Vamos hacer una simulación para pruebar los resultados:

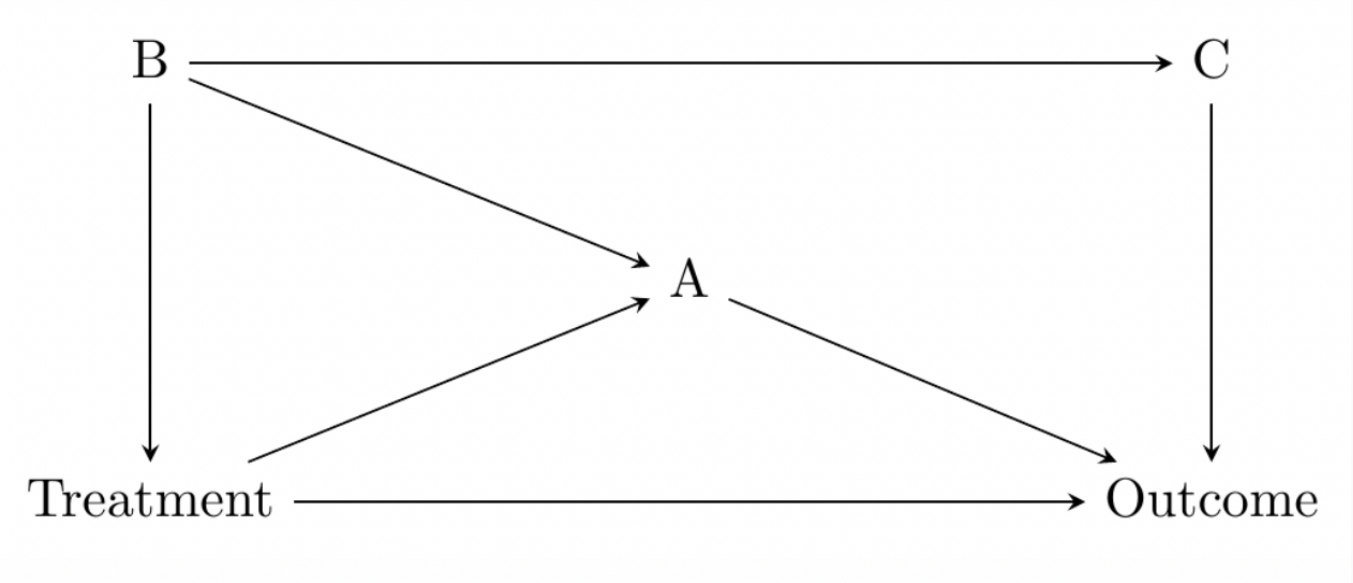
	Incondicional	Condicional (1)	Condicional (2)
Mujer	-3.001***	0.634***	-0.936***
	(0.084)	(0.030)	(0.029)
Ocupación		1.801***	1.011***
		(0.006)	(0.010)
Habilidad			1.969***
			(0.022)
R2	0.112	0.909	0.949

Ejemplo 2



- 1. Escriba todos los caminos entre Treatment y Outcome.

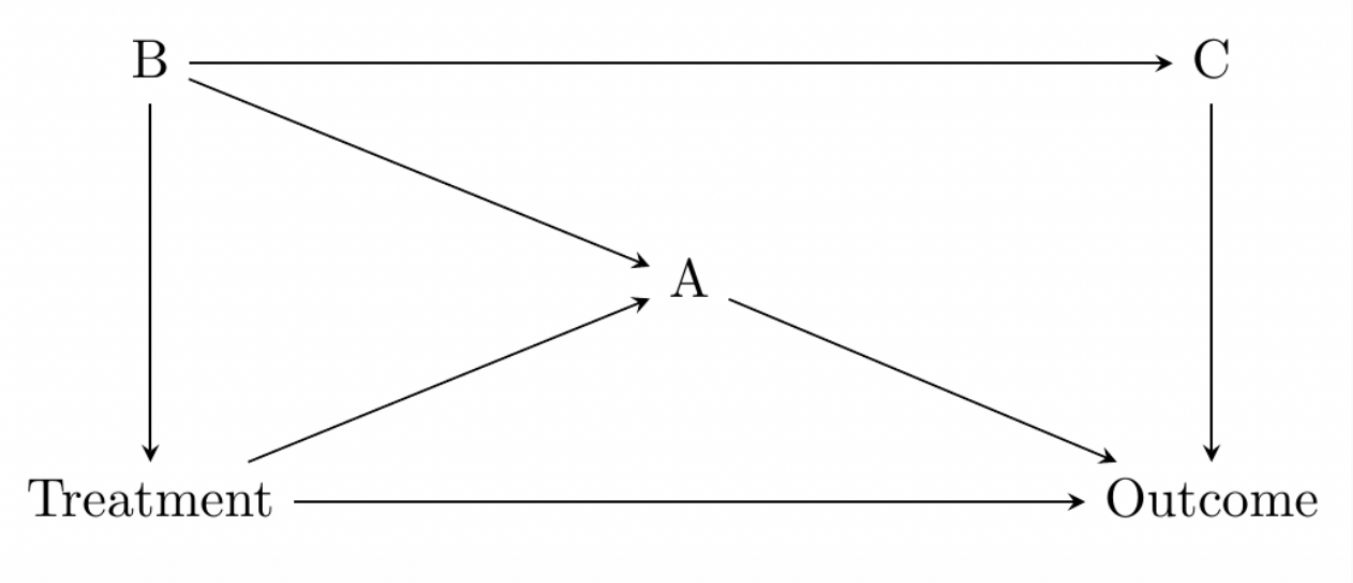
Ejemplo 2



Queremos identificar todos los efectos causales de T a O. ¿Qué estrategias funcionarían?

- a) Condicionar solo en B
- b) Condicionar solo en C
- c) Condicionar en A y C
- d) Condicionar en B y C
- e) Condicionar en A, B y C

Ejemplo 2



3. Ahora nos interesa solamente el efecto directo de T a O. ¿Qué estrategias funcionarían?

- a) Condicionar solo en B
- b) Condicionar solo en C
- c) Condicionar en A y C
- d) Condicionar en B y C
- e) Condicionar en A, B y C

Seleción Muestral

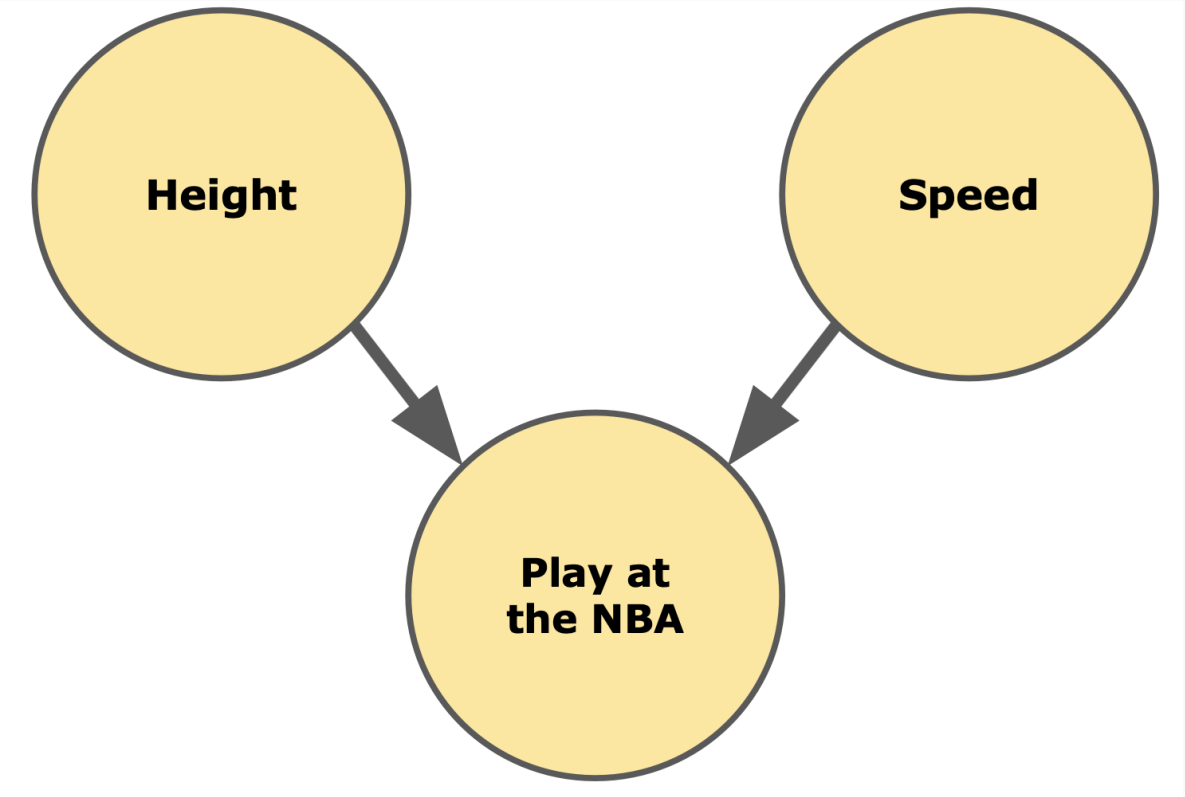
Las principales formas de cerrar un camino causal son condicionando estadisticamente, o si tenemos un collider.

Pero existe una otra situación comun en que un camino se ciera.

Un camino está cerado si la muestra fue selecionada con base en una característica.

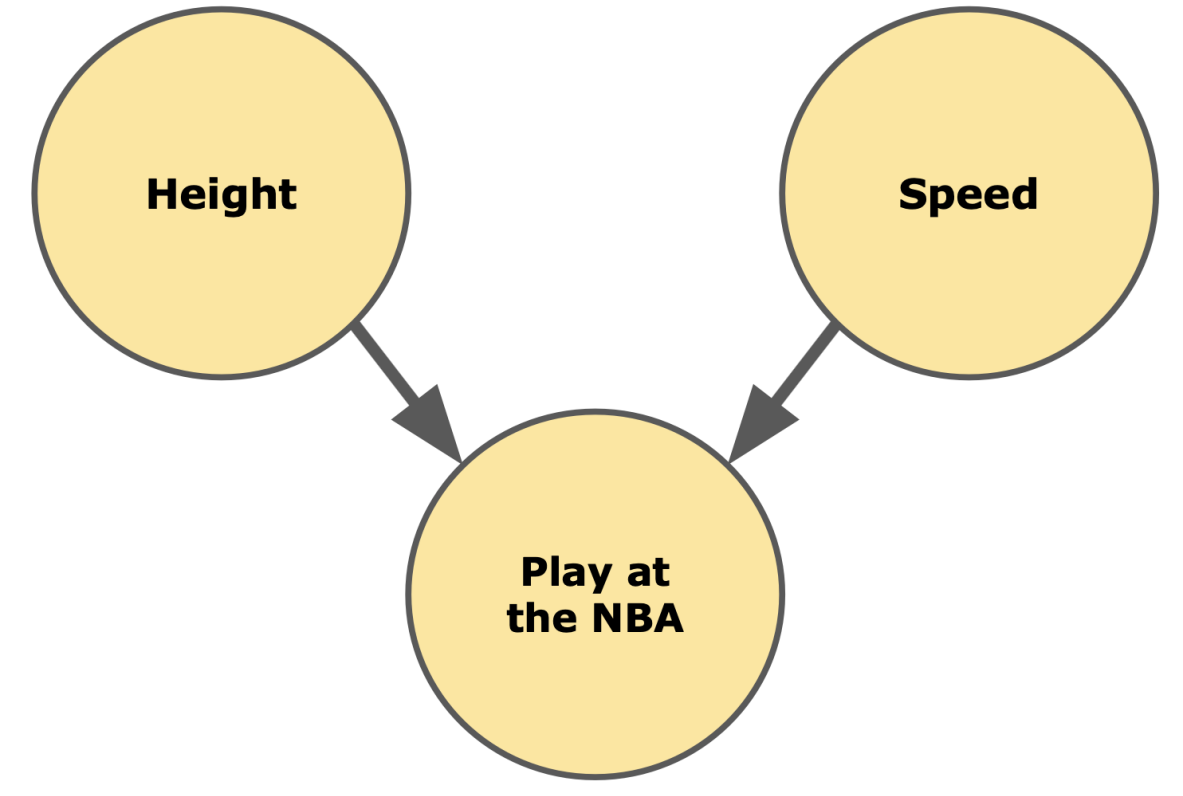
Eso es como una forma no-intencional de controlar por la variable.

Seleción Muestral - Jugadores de Baloncesto



- Supone que no existe una relación entre altura y velocidad.
- Pero sabemos que jugar en la NBA es un collider.
- Si controlamos por NBA, encontramos una relación negativa.

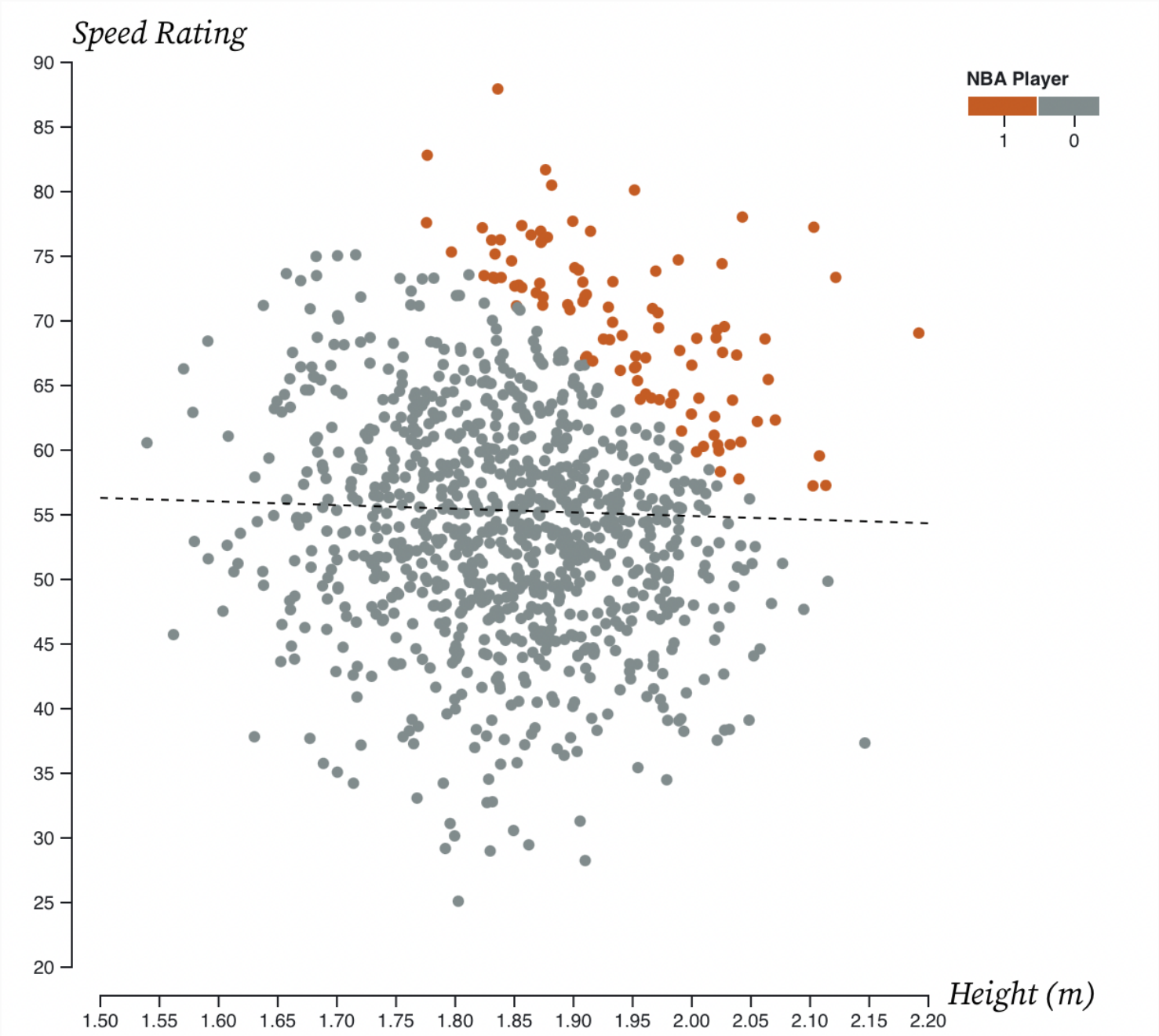
Seleción Muestral - Jugadores de Baloncesto



Si la muestra es solamente de jugadores de la NBA, estamos basicamente controlando por esa variable

Portanto, dentre jugadores de la NBA, podemos ver una relación negativa, mismo si esa relación no existe en la población en general.

Seleción Muestral - Jugadores de Baloncesto



Pruebando el Diagrama

- Una ultima cosa que se puede hacer es utilizar el diagrama para formular hipotesis.
- Se puede analisar la relación dentre qualquieres dos variables.
- Si el diagrama nos dice que dos variables deben ser independientes, pero están correlacionadas en la realidade, puede ser que nos falta algun elemento.
- De la misma forma, podemos analisar relaciones condicionales o no-condicionales.
- Si todas las relaciones están de acuerdo con el diagrama, es un señal de confianza en el diagrama.