# Condicionando en Observables

2026-04-20

# Cuando no podemos aleatorizar

# Plan

**Semana 1**

1. ¿Por qué necesitamos métodos observacionales?
2. El Supuesto de Independencia Condicional
3. Subclasificación
4. Regresión como estrategia de control
5. Emparejamiento (Matching)

**Semana 2 (Bloque 1)**

6. Propensity Score
7. Síntesis: un supuesto, muchos métodos

# Objetivos

**Semana 1:**

- Explicar el Supuesto de Independencia Condicional y conectarlo con el criterio de puerta trasera
- Usar subclasificación para estimar efectos de tratamiento a mano
- Describir qué gana y qué pierde regresión como estrategia de control
- Ejecutar e interpretar matching por vecino más cercano

**Semana 2 (Bloque 1):**

- Explicar cómo el propensity score reduce la dimensionalidad del condicionamiento
- Distinguir emparejamiento sobre $\hat{p}(X)$ e IPW como dos formas de usar el PS
- Interpretar un histograma de propensity score y una tabla de balance

## El problema

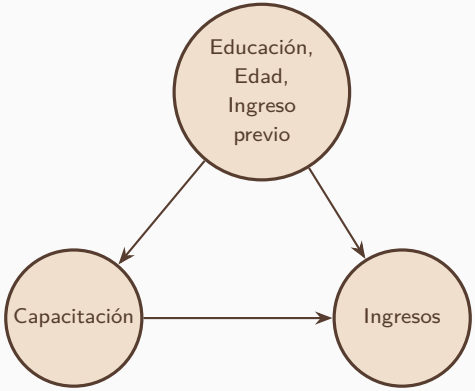
Para la gran mayoría de las preguntas de investigación, no podemos usar experimentos aleatorizados.

Pero aún así queremos aprender sobre efectos de tratamiento.

**Ejemplo:** El gobierno quiere evaluar un programa de capacitación laboral (tipo SENCE). ¿Mejora los ingresos de los participantes?

El problema: la participación es voluntaria. Las personas que se inscriben son distintas de las que no.

# El problema: DAG




Las características $X$ afectan tanto la participación como los ingresos futuros.

El camino de backdoor $D \leftarrow X \rightarrow Y$ está **abierto**.

## ¿Qué sale mal?

La diferencia ingenua $E[Y|D=1] - E[Y|D=0]$ mezcla:

- El efecto causal del programa
- Las diferencias preexistentes entre participantes y no participantes

Si las personas con menor educación se inscriben más, y la menor educación se asocia con menores ingresos, el sesgo de selección puede **subestimar** el efecto real del programa.

## La estrategia: cerrar el camino de backdoor

Si podemos identificar **todas** las variables $X$ que causan sesgo, y condicionar en ellas, cerramos los caminos de backdoor.

Esto es exactamente lo que aprendimos con los DAGs: la idea de hoy es **cómo implementar** ese condicionamiento en la práctica.

# Supuesto de Independencia Condicional

## Supuesto de Independencia Condicional

En lugar de independencia $(Y^1, Y^0) \perp D$, tenemos el **Supuesto de Independencia Condicional** (CIA):

$$(Y^1, Y^0) \perp D|X$$

Interpretación:

1. *Entre individuos con el mismo valor de X*, el tratamiento es independiente de los resultados potenciales.
2. Condicionando en X, no existen caminos de backdoor abiertos.
3. X captura **todas** las razones por las cuales tratados y no tratados difieren (aparte del tratamiento mismo).

Esto es un **supuesto** – no se puede testear. Requiere conocimiento del contexto.

# Subclasificación

## Ejercicio: evaluando un programa

Imagina que te encargan evaluar un programa del gobierno. Solo tienes la siguiente información:

| D | N | $\bar{Y}$ |
|---|---|---|
| 0 | 150 | 1.67 |
| 1 | 250 | 2.40 |

**Discutan en parejas (2 min):** ¿Cuál es tu estimación del efecto de tratamiento? ¿Bajo qué supuesto?

## Subclasificación

Ahora supón que el programa fue administrado en dos tipos de áreas: urbanas y rurales, y las comunas urbanas recibieron más tratamiento:

| Tipo | D | N | $\bar{Y}$ |
|---|---|---|---|
| Urb | 0 | 100 | 1 |
| Urb | 1 | 200 | 2 |
| Rur | 0 | 50 | 3 |
| Rur | 1 | 50 | 4 |

¿Cuál es el efecto de tratamiento en comunas urbanas? ¿Y rurales?

¿Cuál es el efecto promedio de tratamiento?

## Subclasificación

Ahora vamos a cambiar un poco los datos:

| Tipo | D | N | $\bar{Y}$ |
|---|---|---|---|
| Urb | 0 | 100 | 1 |
| Urb | 1 | 200 | 3 |
| Rur | 0 | 50 | 3 |
| Rur | 1 | 50 | 4 |

**Discutan en parejas (2 min):** ¿Cuál es el efecto promedio de tratamiento?

¿Y si queremos aplicar esos resultados a todo Chile, donde el 88% de la población es urbana?

## Los pesos importan

Podemos agregar la información según la distribución de $X$ en cualquier grupo que nos interese:

$$ATE = \sum_x \hat{\tau}_x P(X = x)$$

$$ATT = \sum_x \hat{\tau}_x P(X = x|D = 1)$$

$$ATU = \sum_x \hat{\tau}_x P(X = x|D = 0)$$

Cuáles probabilidades usamos determina sobre qué población estamos hablando.

## Soporte común

Para calcular $\hat{\tau}_x$, necesitamos tener unidades tratadas y de control para cada valor de $X$.

- A esto se le llama el **Supuesto de Soporte Común**.

Es difícil tener soporte común con $X$ continua.

- Imagina si queremos efectos de tratamiento para la edad exacta de cada persona. Seguramente habría muchas edades sin personas en ambos grupos.

## La maldición de la dimensionalidad

A medida que crece la dimensión de $X$, se vuelve más difícil encontrar soporte común.

Con una variable binaria (urbano/rural), tenemos 2 grupos. Con edad, sexo, región, profesión... el número de grupos crece exponencialmente.

Necesitamos métodos más inteligentes para condicionar en $X$.

Tres estrategias: **Regresión**, **Emparejamiento**, **Propensity Score**.

Todas dependen del mismo supuesto (CIA). Difieren en cómo implementan el condicionamiento.

# Regresión

## Regresión como estrategia de control

La subclasificación requiere calcular efectos célula por célula. Con $X$ continua o multidimensional, eso es imposible.

La regresión resuelve el problema **imponiendo una forma funcional**:

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

El coeficiente $\hat{\tau}$ es "el efecto del tratamiento manteniendo $X$ constante."

## ¿Qué gana y qué pierde la regresión?

**Gana:**

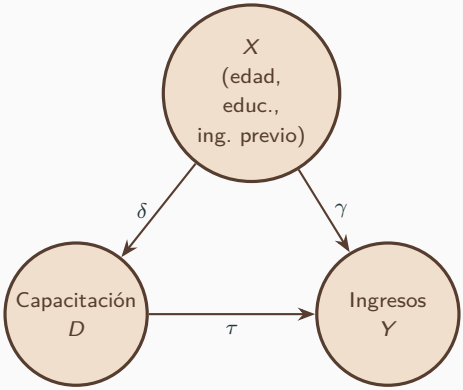
- Maneja variables continuas naturalmente
- La maldición de la dimensionalidad no es tan problemática: imponemos separabilidad

**Pierde:**

- Impone **linealidad**. Si la relación entre $X$ e $Y$ es altamente no lineal, la regresión puede no cerrar adecuadamente los backdoors
- Si se omite un confundidor, el sesgo contamina $\hat{\tau}$ de manera predecible (lo veremos a continuación)

## Cerrando el backdoor con regresión

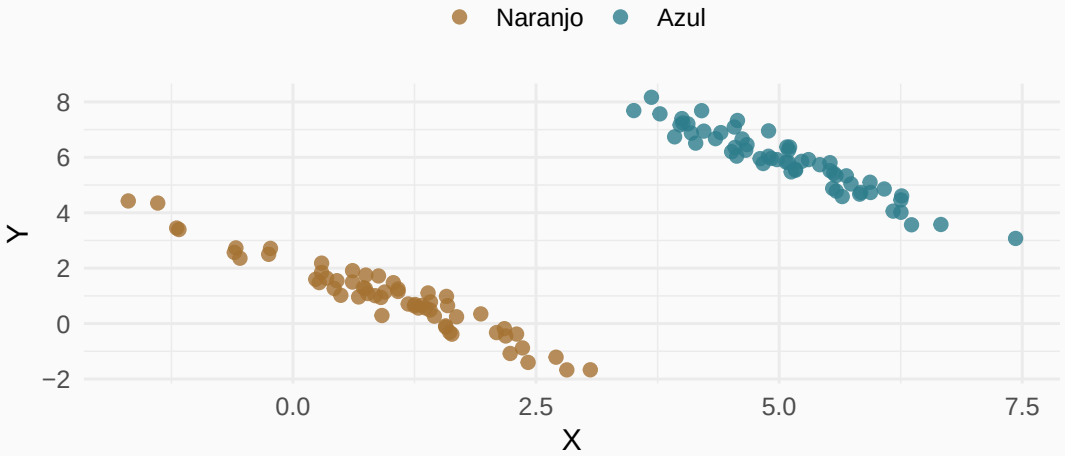
El DAG del programa de capacitación tiene un camino de backdoor abierto:




Al incluir $X$ en la regresión $Y_i = \alpha + \tau D_i + \beta X_i + u_i$, **bloqueamos** el camino $D \leftarrow X \rightarrow Y$.

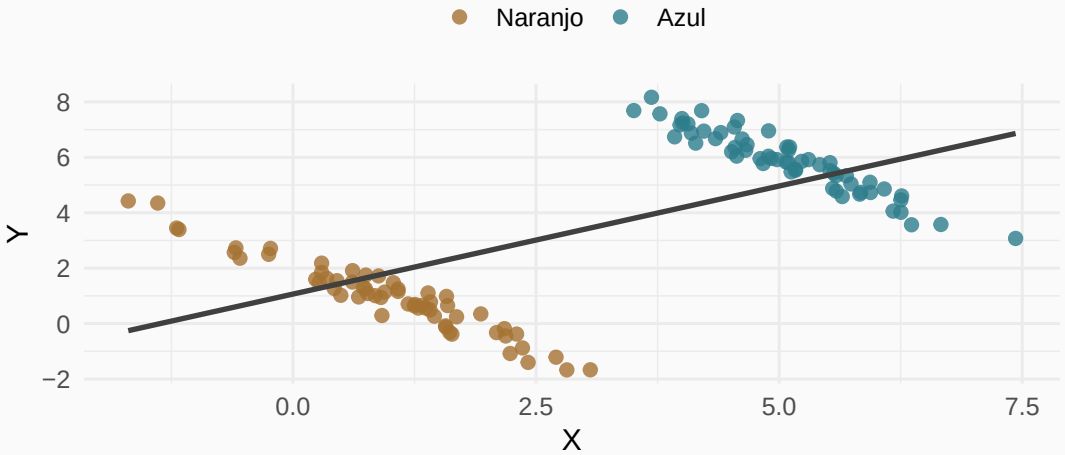
## La intuición: dos grupos, dos patrones

Observamos dos grupos (naranjo y azul). Dentro de cada grupo la relación entre $X$ e $Y$ es **negativa** — pero el grupo azul tiene $X$ e $Y$ más altos.



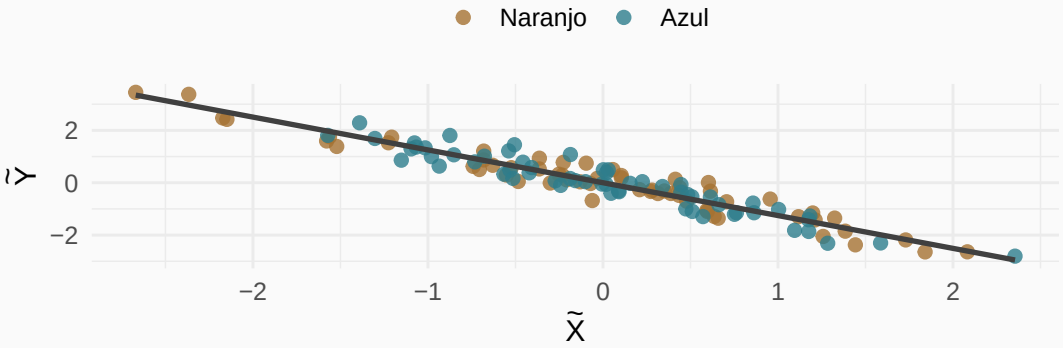
## La intuición: la regresión simple nos engaña

Si ignoramos el color y estimamos $Y = \alpha + \beta X + u$, la pendiente es **positiva** — lo contrario de la relación dentro de cada grupo.



# La intuición: residualizar — teorema FWL

Residualicemos $X$ e $Y$ sobre el color (restando la media de cada grupo). La pendiente es ahora **negativa** — la relación intra-grupo.




Equivale a estimar $Y \sim X +$ color: **regresión múltiple = regresión simple en los residuos.**

## ¿Qué estimand nos da la regresión?

Cuando los efectos son heterogéneos, $\hat{\tau}_{OLS}$ es un promedio de los efectos locales $\tau(x)$, **ponderado** por $\text{Var}(D \mid X = x)$.

- Los estratos donde $D$ varía mucho (muchos tratados y controles) reciben más peso.
- Si $\tau$ es constante: $\hat{\tau}_{OLS} = ATE$. Si no, no coincide en general con ATE ni ATT.

Consecuencia: regresión y matching pueden dar resultados distintos aunque CIA se cumpla — estiman objetos distintos.

## Sesgo por variable omitida (SVO)

Si omitimos $X$ de la regresión, el estimador converge a:

$$\text{plim}(\hat{\tau}_{\text{corta}}) = \tau + \underbrace{\gamma}_{X \to Y} \cdot \underbrace{\delta}_{D \leftarrow X}$$

- $\gamma$: efecto de $X$ sobre $Y$ (flecha $X \to Y$ en el DAG)
- $\delta$: correlación entre $D$ y $X$ (flecha $D \leftarrow X$ en el DAG)

**Ejercicio de signo:** En el programa de capacitación, la educación aumenta los ingresos ($\gamma > 0$) y las personas más educadas participan más ($\delta > 0$). ¿Hacia qué lado sesga ignorar la educación?

$\gamma > 0,\ \delta > 0 \Rightarrow$ sesgo hacia arriba — sobreestimamos el efecto del programa.

## SVO: demostración numérica

Simulamos $Y = \tau D + \gamma X + \varepsilon$ con $\tau = 1$, $\gamma = 1.5$, y $D = 0.5\,X + \eta$:

```
set.seed(1); n_ovb <- 5000
X_ovb <- rnorm(n_ovb); D_ovb <- 0.5 * X_ovb + rnorm(n_ovb)
Y_ovb <- 1 * D_ovb + 1.5 * X_ovb + rnorm(n_ovb)
tau_corta <- coef(lm(Y_ovb ~ D_ovb))["D_ovb"]
tau_larga <- coef(lm(Y_ovb ~ D_ovb + X_ovb))["D_ovb"]
delta_hat <- coef(lm(X_ovb ~ D_ovb))["D_ovb"]
round(c("tau corto"     = tau_corta,
        "tau largo"     = tau_larga,
        "gamma x delta" = 1.5 * delta_hat), 2)
```

```
##     tau corto.D_ovb     tau largo.D_ovb gamma x delta.D_ovb
##                1.64                1.00                0.63
```

## OVB en la práctica: retornos a la educación

Card (1993) — NLS Young Men. ¿Cuánto aumenta el salario por un año extra de educación?

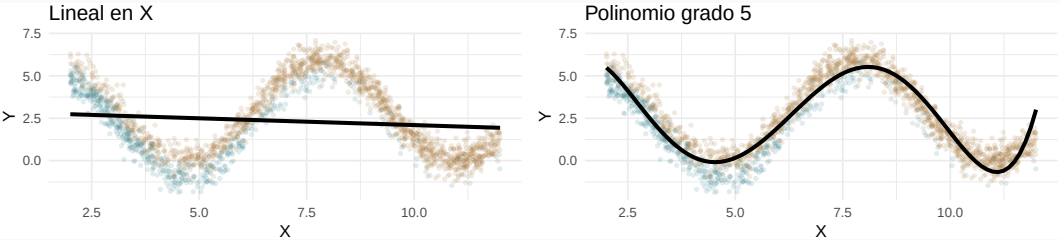
Omitir la **habilidad** es el sesgo clásico: personas más hábiles estudian más Y ganan más, independiente de la educación.

| Especificación | $\hat{\beta}_{\text{educ}}$ | SE | N |
|---|---|---|---|
| (1) Solo educ | 0.0521 | 0.0029 | 3010 |
| (2) + exper, demografía | 0.0740 | 0.0035 | 3010 |
| (3) + IQ (habilidad) | 0.0693 | 0.0049 | 2061 |

El coeficiente de educ cae al añadir IQ: $\gamma_{\text{IQ}\to\text{wage}} > 0$ y $\delta_{\text{IQ}\to\text{educ}} > 0$, así que $\gamma \cdot \delta > 0$ — el DAG predice la dirección del sesgo.

# Demostración: la forma funcional importa

Datos simulados: $Y = 2 + \sin(X) \cdot 3 + D + \varepsilon$. El efecto verdadero de $D$ es **1**.




```
##   Lineal.D Flexible.D
##      1.282      1.004
```

La especificación lineal sesga $\hat{\tau}$: el modelo no captura la curva, así que la variación de $X$ no explicada queda contaminada. El polinomio recupera el efecto verdadero.

Problema: ¿cuánta flexibilidad es suficiente? Si no queremos apostar por una forma

## Regresión como imputación

Igual que matching, la regresión se puede leer como **imputación de resultados potenciales**:

- **Matching** imputa $\hat{Y}_i^0 = Y_{m(i)}$: el $Y$ del vecino más similar en $X$.
- **Regresión** imputa $\hat{Y}_i^0 = \hat{\alpha} + \hat{\beta} X_i$: la predicción del modelo al fijar $D = 0$.

En ambos casos, $\hat{\tau}$ es el promedio de las diferencias entre el $Y$ observado y el $\hat{Y}$ contrafactual imputado.

Matching imputa con datos; regresión imputa con el modelo.

## Regresión: Resumen

- Supuesto clave: CIA + forma funcional correcta
- La intuición FWL: $\hat{\tau}$ compara variación en $D$ no explicada por $X$
- Si se omite un confundidor, el SVO es $\gamma \cdot \delta$ — predecible a partir del DAG
- Simple y familiar, pero depende de la especificación

# Pausa

## Descanso

A continuación: **Emparejamiento** — otra forma de implementar el mismo supuesto CIA, sin imponer linealidad.

# Emparejamiento

## La idea del emparejamiento

Para calcular efectos del tratamiento, queremos estimar el resultado contrafactual para cada unidad.

¿Qué tal si simplemente elegimos una unidad "similar" con la asignación opuesta?

El **emparejamiento** consiste en asignar una o más unidades de control a cada unidad tratada, basándonos en la similitud en $X$.

## Un ejemplo sencillo

3 tratados y **6 posibles controles**. La única covariable es la edad.

**Tratados ($D = 1$)**

| ID | Edad | $Y$ |
|---|---|---|
| T1 | 25 | 12 |
| T2 | 30 | 15 |
| T3 | 35 | 18 |

**Controles ($D = 0$)**

| ID | Edad | $Y$ |
|---|---|---|
| C1 | 22 | 9 |
| C2 | 25 | 10 |
| C3 | 28 | 13 |
| C4 | 30 | 12 |
| C5 | 35 | 14 |
| C6 | 40 | 16 |

¿Qué control corresponde a cada tratado? ¿Cuáles quedan sin pareja?

## Empareja a cada tratado

Emparejamos por edad idéntica:

- T1 (edad 25) $\rightarrow$ C2: diferencia $= 12 - 10 = 2$
- T2 (edad 30) $\rightarrow$ C4: diferencia $= 15 - 12 = 3$
- T3 (edad 35) $\rightarrow$ C5: diferencia $= 18 - 14 = 4$

$$\widehat{ATT} = \frac{2 + 3 + 4}{3} = 3$$

C1, C3 y C6 quedan sin pareja — sus edades no coinciden con ningún tratado.

Acabamos de hacer matching exacto a mano. Formalicémoslo.

## Emparejamiento: Estimador

Llamemos a la unidad emparejada con $i$ como $m(i)$. El estimador más simple es:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{m(i)})$$

Nota: el matching estima naturalmente el **ATT**, porque partimos de cada unidad tratada.

## Matching como imputación

Para cada unidad tratada $i$, observamos $Y_i^1$ pero **no** $Y_i^0$.

El matching **imputa** ese resultado faltante: $\hat{Y}_i^0 = Y_{m(i)}$, el resultado de la unidad de control más similar.

El ATT estimado es entonces simplemente el promedio de las diferencias imputadas:

$$\hat{\tau}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i^1 - \hat{Y}_i^0) = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{m(i)})$$

Esta es la misma ecuación de antes, pero vista como **imputación de contrafactuales** — la misma lógica del marco de resultados potenciales.

## Emparejamiento exacto

El caso más simple: cada unidad tratada se empareja con una unidad de control que tiene **exactamente** los mismos valores de $X$.

## Un caso más grande: los datos

6 tratados y 8 controles. Queremos emparejar por edad.

**Tratados ($D = 1$)**

| ID | Edad | $Y$ |
|---|---|---|
| T1 | 25 | 12 |
| T2 | 28 | 13 |
| T3 | 30 | 15 |
| T4 | 32 | 15 |
| T5 | 35 | 18 |
| T6 | 40 | 20 |

**Controles ($D = 0$)**

| ID | Edad | $Y$ |
|---|---|---|
| C1 | 22 | 9 |
| C2 | 25 | 11 |
| C3 | 28 | 12 |
| C4 | 30 | 13 |
| C5 | 33 | 14 |
| C6 | 35 | 15 |
| C7 | 40 | 17 |
| C8 | 45 | 19 |

¿Qué tratado queda sin pareja?

## Las parejas y el ATT

| Tratado | Edad | $Y_T$ | Control | $Y_C$ | $Y_T - Y_C$ |
|---|---|---|---|---|---|
| T1 | 25 | 12 | C2 | 11 | 1 |
| T2 | 28 | 13 | C3 | 12 | 1 |
| T3 | 30 | 15 | C4 | 13 | 2 |
| T4 | 32 | 15 | — | — | *sin match* |
| T5 | 35 | 18 | C6 | 15 | 3 |
| T6 | 40 | 20 | C7 | 17 | 3 |

$$\widehat{ATT} = \frac{1+1+2+3+3}{5} = 2 \quad \text{(sobre los 5 tratados emparejables)}$$

Controles no utilizados: C1 (22), C5 (33), C8 (45). T4 queda fuera del análisis.

## Pregunta

Si no encontramos pareja para algunas unidades, ¿cuándo obtenemos una estimación no sesgada del ATT?

- a) Siempre que no eliminemos ninguna unidad del grupo de control
- b) Siempre que no eliminemos ninguna unidad del grupo tratado
- c) Nuestra estimación es sesgada si eliminamos cualquier unidad
- d) Eliminar unidades no genera sesgo

## Discusión: ¿cuándo es insesgado el ATT?

Respuesta: **(b)**

**Eliminar controles** que no tienen pareja no genera sesgo en el ATT — sólo perdemos eficiencia, siempre que CIA siga cumpliéndose en las unidades restantes.

**Eliminar tratados** fuera del soporte común *sí* cambia el estimand: ya no estamos estimando el ATT sobre todos los tratados, sino el ATT sobre el subconjunto emparejable.

Perder datos $\neq$ perder identificación. Pero sí cambia la pregunta si excluimos tratados.

## El límite del matching exacto

3 tratados, 9 controles, **3 covariables**: edad, educación (años), experiencia (años).

**Tratados**

| ID | Ed | Edu | Exp |
|---|---|---|---|
| T1 | 25 | 12 | 3 |
| T2 | 30 | 16 | 8 |
| T3 | 35 | 14 | 12 |

**Controles**

| ID | Ed | Edu | Exp |
|---|---|---|---|
| C1 | 25 | 12 | 3 |
| C2 | 30 | 16 | 6 |
| C3 | 30 | 14 | 8 |
| C4 | 28 | 16 | 8 |
| C5 | 35 | 12 | 12 |
| C6 | 32 | 14 | 12 |
| C7 | 35 | 14 | 10 |
| C8 | 28 | 15 | 7 |
| C9 | 33 | 13 | 11 |

¿Puedes encontrar un match exacto (en las 3 covariables) para cada tratado?

## El límite del matching exacto: solución

- **T1** $\rightarrow$ C1: edad 25, educación 12, experiencia 3. **Match perfecto.**

- **T2**: necesita edad 30, educación 16, experiencia 8.
  - C2: coincide en edad y educación, pero exp $= 6 \neq 8$
  - C3: coincide en edad y experiencia, pero edu $= 14 \neq 16$
  - C4: coincide en educación y experiencia, pero edad $= 28 \neq 30$
  - No existe un control con las tres covariables iguales.

- **T3**: necesita edad 35, educación 14, experiencia 12. C5, C6, C7 coinciden en a lo sumo dos de las tres. Tampoco hay match exacto.

Con más covariables, los matches exactos se vuelven imposibles. Necesitamos una noción de **cercanía**.

## Del exacto al aproximado: ¿qué significa “similar”?

Con $X$ continua no hay match exacto. Necesitamos definir **distancia**.

El problema: si edad está en años e ingresos en miles de pesos, la distancia euclidiana sin estandarizar está dominada por los ingresos. Por eso estandarizamos:

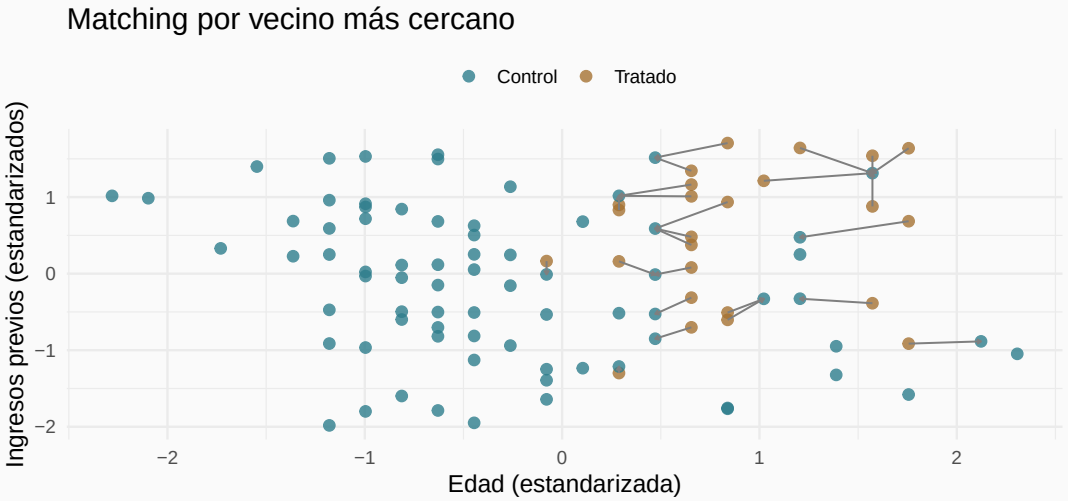
$$d(i,j) = \sqrt{\sum_k \left( \frac{X_{ik} - X_{jk}}{\text{sd}(X_k)} \right)^2}$$

## Decisiones de implementación: sesgo vs. varianza

Tres trade-offs clave:

- $k$ **vecinos:** más vecinos $\Rightarrow$ menos varianza, pero matches más lejanos $\Rightarrow$ más sesgo.
- **Con/sin reemplazo:** con reemplazo permite mejores matches pero concentra peso en pocos controles $\Rightarrow$ más varianza del estimador.
- **Caliper:** descarta matches con distancia $> c$ $\Rightarrow$ mejor soporte común, pero puede excluir tratados sin pareja $\Rightarrow$ cambia el estimand.

# Demostración: Matching en acción



## Balance antes y después del matching

| group | Edad media | Ingreso medio | Tasa ascenso |
|---|---|---|---|
| Control (pre-match) | 38.8 | 77607 | 0.293 |
| Control emparejado | 44.8 | 81696 | 0.400 |
| Tratados | 45.2 | 81857 | 0.760 |

SMD edad: 1.52 (pre-match) $\rightarrow$ 0.13 (post-match). Regla de referencia: SMD $< 0.1$ indica buen balance.

**ATT estimado:** 0.36

## Matching vs. Regresión

| Dimensión | Regresión | Matching |
|---|---|---|
| Forma funcional | Lineal (típicamente) | No paramétrica |
| Soporte común | Extrapola con todos los datos | Usa sólo unidades comparables |
| Estimand natural | Promedio ponderado | ATT |
| Muchas covariables | Escala bien | Difícil (maldición dimensionalidad) |
| Transparencia | Coeficiente $\hat{\tau}$ | Promedio de diferencias observadas |

Las dos estrategias son **complementarias**. La elección depende de qué sea más creíble en el contexto: la linealidad o el soporte común.

## Heterogeneidad: tres caminos

**Pregunta:** ¿El programa funciona más para jóvenes que para adultos?

- **Regresión:** añadir $D \times 1\{\text{joven}\}$ al modelo — leer la interacción.
- **Matching:** correr matching separadamente en cada grupo de edad.
- **Restricción muestral:** filtrar a jóvenes, estimar con cualquier método.

Ojo: los dos primeros responden "¿difiere el efecto?"; el tercero responde "¿cuál es el efecto en este subgrupo?" — preguntas distintas.

## Emparejamiento: Resumen

- Supuesto clave: CIA + soporte común
- Estima naturalmente el ATT (matching como imputación de $Y^0$)
- No depende de linealidad (ventaja sobre regresión); requiere soporte común
- Dificultades con muchas variables de control (maldición de la dimensionalidad)
- Muchas decisiones de implementación (métrica, caliper, número de vecinos)

# Propensity Score

## El problema de dimensionalidad, de nuevo

El matching funciona bien con 1-2 variables.

Pero si necesitamos controlar por edad, educación, ingreso, empleo previo, región, estado civil...

Encontrar un match en 10 dimensiones es casi imposible.

## La idea clave: el propensity score

El **propensity score** resume el vector $X$ en un solo número:

$$p(X) = \Pr(D = 1 \mid X)$$

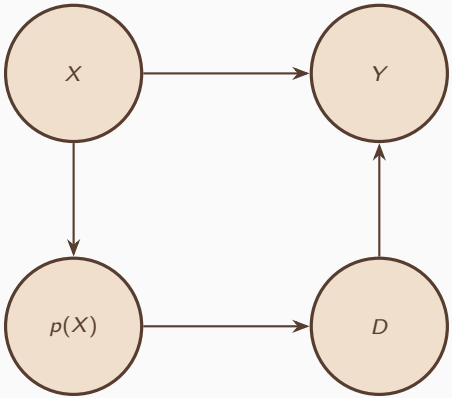
**Teorema (Rosenbaum y Rubin, 1983):** Si CIA vale condicionando en $X$, entonces vale condicionando solo en $p(X)$:

$$(Y^0, Y^1) \perp D \mid p(X)$$

# La idea clave: el propensity score

$p(X)$ es un *balancing score*: dentro de grupos con el mismo $\hat{p}(X)$, tratados y controles son comparables — la asignación es como-si aleatoria.

Implicación: basta condicionar en el escalar $\hat{p}(X)$ en lugar del vector completo $X$.



## La intuición del balancing score

¿Por qué basta condicionar en $\hat{p}(X)$? Porque dentro de cada valor del PS, $X$ se balancea entre grupos.

| Estrato | Grupo | Edad | Educ. | Ingreso previo |
|---|---|---|---|---|
| $\hat{p} \approx 0.30$ | Tratado | 31 | 10 | \$2,100 |
| $\hat{p} \approx 0.30$ | Control | 32 | 11 | \$2,300 |
| $\hat{p} \approx 0.70$ | Tratado | 29 | 13 | \$8,400 |
| $\hat{p} \approx 0.70$ | Control | 28 | 12 | \$7,900 |

Dentro de cada estrato del PS, tratados y controles lucen similares en $X$, aunque el matching solo usó el escalar $\hat{p}(X)$.

*Verificaremos esto empíricamente más adelante con la tabla de balance.*

## Cómo usamos el PS

**Paso 1:** Estimar $\hat{p}(X)$ con un logit o probit de $D$ sobre $X$.

**Paso 2 (dos alternativas):**

- **Emparejamiento** sobre $\hat{p}(X)$: cada tratado encuentra su control más cercano en el PS. Estimamos el ATT usando $\hat{p}(X)$ como distancia en lugar del vector $X$.
- **Ponderación** (IPW): ver la siguiente lámina.

*Nota: la subclasificación por deciles del PS también es posible pero no la cubriremos.*

## Inverse Probability Weighting (IPW)

En lugar de descartar o duplicar unidades, **reponderamos todas** usando el PS.

| Estimand | Peso para tratados | Peso para controles |
|---|---|---|
| ATT | $1$ | $\frac{\hat{p}(X)}{1-\hat{p}(X)}$ |
| ATE | $\frac{1}{\hat{p}(X)}$ | $\frac{1}{1-\hat{p}(X)}$ |

**Intuición:** subponderamos controles muy distintos de los tratados (PS bajo) y sobreponderamos los comparables. El resultado es como-si la muestra observacional fuera un experimento más balanceado.

Conexión: Lo veremos de nuevo en DID moderno (Callaway-Sant'Anna, 2021), que usa pesos IPW para estimar efectos con tratamiento escalonado.

## ¿Cómo sabemos si funciona?

Si el PS está bien estimado, **condicional en** $p(X)$, $X$ debe estar balanceado entre tratados y controles.

Lo verificamos con una **tabla de balance**: comparar medias de $X$ entre grupos dentro de estratos del PS.

Si hay desbalance: volver al paso 1 e intentar otra especificación (más variables, interacciones, forma funcional diferente).

## Soporte común en el PS

Si las distribuciones del PS para tratados y controles no se superponen, no podemos estimar efectos.

El **histograma del PS** es la herramienta diagnóstica clave:

- Si tratados y controles se concentran en rangos similares $\rightarrow$ buen soporte
- Si están en extremos opuestos $\rightarrow$ hay un problema de selección severo

## Demostración: NSW y controles del CPS

El **National Supported Work (NSW)** fue un experimento aleatorizado en EE.UU. El efecto experimental sobre ingresos fue $\approx$ \$1,800.

¿Podemos replicar ese resultado usando datos observacionales? Reemplazamos el grupo de control experimental con datos del **Current Population Survey (CPS)** – una muestra muy diferente.

## Regresión en datos NSW

Antes de usar el PS, veamos cuánto logra la regresión:

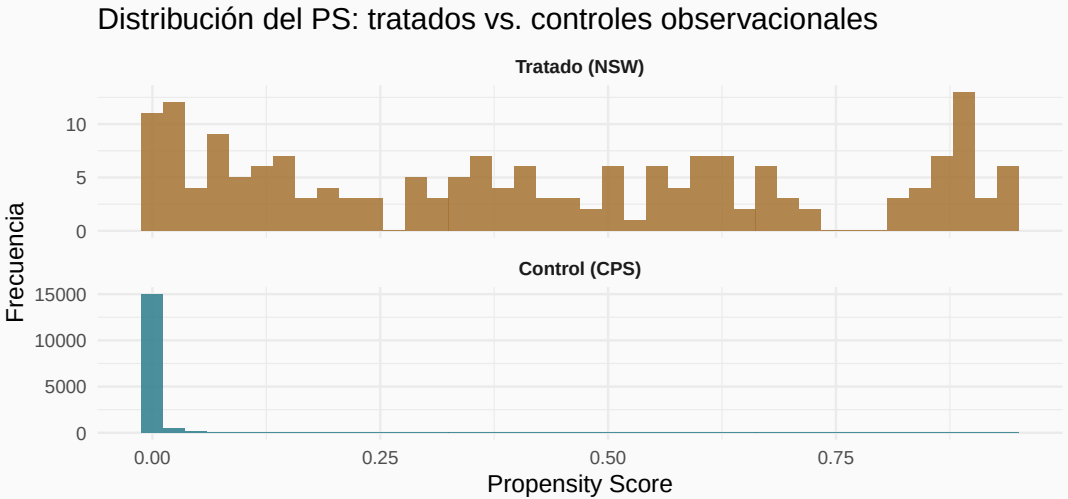
```
# Sin controles
ols_naive    <- lm(re78 ~ treat, data = data_ps)
# Con controles
ols_controls <- lm(re78 ~ treat + age + educ + black + hisp +
                     marr + nodegree + re74 + re75, data = data_ps)

round(c("OLS sin controles"   = coef(ols_naive)["treat"],
        "OLS con controles"   = coef(ols_controls)["treat"],
        "Benchmark exp."      = 1794), 0)
```

```
## OLS sin controles.treat OLS con controles.treat          Benchmark exp.
##                   -8498                     699                    1794
```

## Regresión NSW: ¿cuánto nos acercamos?

La regresión ingenua sobreestima el efecto (SVO positivo: los controles del CPS son más jóvenes y menos educados). Agregar controles reduce el sesgo, pero no lo elimina: el soporte común entre NSW y CPS es limitado. Los métodos de PS a continuación atacan exactamente ese problema.

# Distribución del Propensity Score




Los controles del CPS tienen PS concentrado cerca de 0. El soporte común es limitado.

## Soporte común: implicaciones

**Piensa y discute (3 minutos)**

El histograma muestra que los controles del CPS se concentran cerca de $\hat{p}(X) = 0$, mientras los tratados del NSW se distribuyen más ampliamente.

1. ¿Qué implica esto para estimar el **ATT**?
2. ¿Y para estimar el **ATE**?

## Estimación con PS — NN Matching

```r
# NN Matching en el PS
treated_ps <- data_ps |> filter(treat == 1)
control_ps <- data_ps |> filter(treat == 0)
nn_match <- treated_ps |> rowwise() |>
  mutate(match_y = control_ps$re78[
    which.min(abs(control_ps$pscore - pscore))]) |>
  ungroup()
att_nn <- mean(nn_match$re78 - nn_match$match_y)
```

## Estimación con PS — IPW

```
# Pesos ATT: tratados = 1, controles = p/(1-p)
w_att <- if_else(data_ps$treat == 1, 1,
                 data_ps$pscore / (1 - data_ps$pscore))
att_ipw <- with(data_ps,
  weighted.mean(re78[treat == 1], w_att[treat == 1]) -
  weighted.mean(re78[treat == 0], w_att[treat == 0]))
round(att_ipw, 0)
```

```
## [1] 1782
```

Cada control recibe un peso proporcional a cuán "parecido" es a los tratados según el PS: controles con $\hat{p} \approx 0$ apenas contribuyen.

## Balance post-emparejamiento en NSW

| Grupo | Edad | Educ. | Ing. 1974 | Ing. 1975 |
|---|---|---|---|---|
| Control CPS (post) | 25.7 | 10.8 | 1839 | 1475 |
| Control CPS (pre) | 33.2 | 12.0 | 14017 | 13651 |
| Tratados (NSW) | 25.8 | 10.3 | 2096 | 1532 |

SMD edad: -0.8 $\rightarrow$ 0.02. SMD ingreso 1974: -1.57 $\rightarrow$ 0.05. Emparejar sobre $\hat{p}(X)$ balancea las covariables aunque el emparejamiento fue en un solo escalar.

## Resultados: todos los métodos

| Método | Estimación ($) |
|---|---|
| OLS sin controles | -8,498 |
| OLS con controles | 699 |
| NN Matching en p(X) | 952 |
| IPW (ATT) | 1,782 |
| Benchmark experimental | 1,794 |

Los métodos de PS se acercan más al benchmark experimental que la regresión sola. El soporte común y la correcta comparación importan.

## Limitaciones del PS

- **Especificación**: el PS es un modelo; si está mal especificado, $\hat{p}(X)$ no balancea correctamente.
- **Pesos extremos en matching**: cuando $\hat{p}(X) \approx 0$ o $\approx 1$, pocas unidades comparables — el método pierde potencia.
- **Pesos extremos en IPW**: cuando $\hat{p}(X) \to 0$ o $\to 1$, algunos controles reciben pesos muy grandes y dominan el estimador. Se recomienda *trim* de pesos o restringir al soporte común.
- **Paradoja de King & Nielsen**: en algunos casos, emparejar sobre PS puede *aumentar* el desbalance respecto a emparejar sobre $X$ directamente. No es una bala de plata.

El PS no reemplaza CIA: sigue siendo un supuesto no testeable. Solo ataca la dimensionalidad.

## Propensity Score: Resumen

- **Reduce la dimensionalidad**: controla por un solo escalar $p(X)$ en lugar del vector $X$
- **Dos formas de usarlo**: emparejamiento sobre $\hat{p}(X)$ (ATT) o ponderación IPW (ATT o ATE)
- **Verificar balance**: si $\hat{p}(X)$ está bien especificado, $X$ debe quedar balanceado dentro de estratos del PS
- **Soporte común es crucial**: sin superposición entre tratados y controles, no hay comparación válida
- Lo veremos de nuevo en **DID moderno**: los estimadores de tratamiento escalonado usan pesos IPW

# Síntesis

## Un supuesto, muchos métodos

Todos los métodos de esta clase dependen del mismo supuesto: **CIA**.

Difieren en cómo implementan el condicionamiento:

| Método | Cómo condiciona | Estima | Riesgo |
|---|---|---|---|
| Subclasificación | Estratifica por $X$ | ATE/ATT | Dimensionalidad |
| Regresión | Control lineal | Prom. ponderado | Forma funcional + SVO |
| Matching | Busca unidades similares | ATT | Soporte común |
| Propensity Score | Reduce $X$ a $p(X)$ | ATT | PS mal especificado |

## La amenaza no observable

Todos estos métodos **fallan** si CIA no se cumple.

Si existe un confundidor que no observamos ni podemos medir, ninguno de estos métodos puede salvarnos.

Esto motiva el resto del curso:

- **Variables instrumentales**: usan variación exógena que no depende de $X$
- **Regresión discontinua**: explota un umbral arbitrario
- **Diferencias en diferencias**: usa variación temporal

Cada uno resuelve el problema de los confundidores no observados de una manera diferente.

## Próxima clase

- **Tema:** Cumplimiento Imperfecto y el LATE
- **Conexión:** ¿Qué pasa cuando las personas no cumplen con su asignación de tratamiento? El incumplimiento es otra forma de selección — y nos llevará naturalmente a las variables instrumentales.