Propensity Score Matching

Version Compacta

Dr. Sergio Bejar Lopez

CIDE

Enero 2026

El Problema: Auto-Seleccion

Ejemplo: Jovenes Construyendo el Futuro
Programa de capacitacion laboral - participacion voluntaria

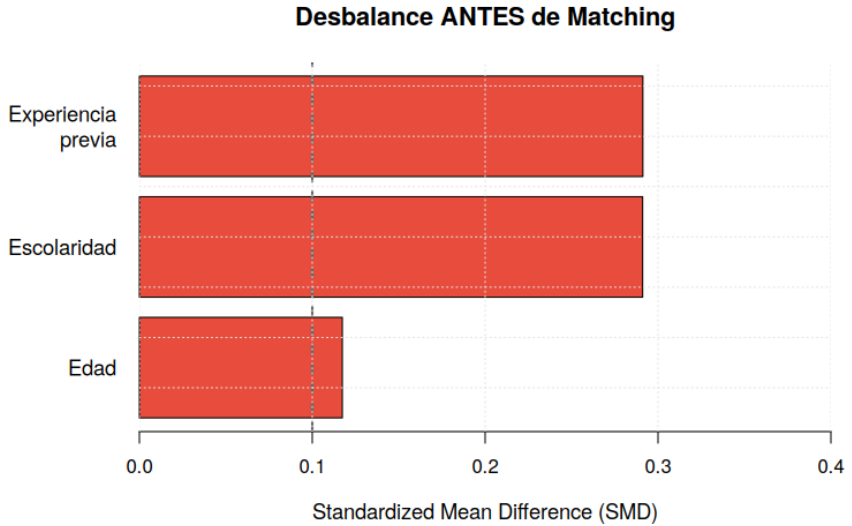
Comparacion ingenua:

Grupo	Empleo Formal
Participantes	44%
No participantes	38%
Diferencia	6 pp

Problema

¿Efecto del programa o diferencias pre-existentes?

Los Grupos NO Son Comparables



La Solucion: Matching

Idea central:

Matching

Comparar cada participante con no-participantes **similares** en caracteristicas observables

Analogia:

- Encontrar "gemelos estadisticos"
- Uno participo, otro no
- Comparar estos pares

Resultado:

- Grupos mas comparables
- Estimacion con menos sesgo
- Controla por observables

Propensity Score: El Resumen

Problema: Muchas variables (edad, educacion, experiencia...)

Solucion: Propensity Score

Definicion

$$e(X_i) = P(\text{Tratamiento} = 1|X_i)$$

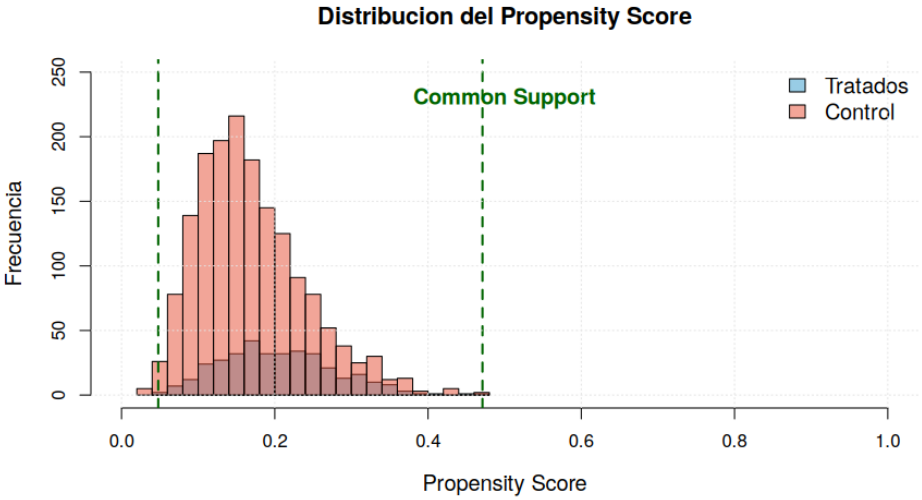
Probabilidad de participar dadas tus caracteristicas

Ventaja:

- Resume TODA la info en 1 numero (entre 0 y 1)
- Matching en PS = matching en todas las variables
- Facil de visualizar

Calculo: Regresion logistica

Supuesto 1: Common Support



Supuesto 2: CIA

Conditional Independence Assumption (CIA)

$$\{Y_i(1), Y_i(0)\} \perp T_i | X_i$$

Condicional en X, el tratamiento es como aleatorio

En palabras:

- Controlando por observables (edad, educacion...)
- NO hay confusores no observados importantes
- Motivacion, talento, redes no medidas pueden violar CIA

CRITICO

CIA NO es testeable - requiere conocimiento del contexto
Si hay seleccion fuerte en no-observables: matching NO funciona

Matching en R: Paso a Paso

```r
library(MatchIt)
library(cobalt)

# Paso 1: Hacer matching
match <- matchit(
  tratamiento ~ edad + mujer + escolaridad + urbano +
                experiencia_previa + ingreso_familiar,
  data = datos,
  method = "nearest",
  distance = "glm"
)

# Paso 2: Verificar balance
bal.tab(match, thresholds = c(m = 0.1))

# Paso 3: Extraer datos matched
datos_matched <- match.data(match)

# Paso 4: Estimar ATT
att <- mean(datos_matched$empleo_formal[datos_matched$tratamiento==1]) -
       mean(datos_matched$empleo_formal[datos_matched$tratamiento==0])
```

Balance: Antes vs Despues

ANTES de matching (desbalanceado):

- Edad: SMD $= 0.31$
- Escolaridad: SMD $= 0.28$
- Experiencia: SMD $= 0.19$

DESPUES de matching (balanceado):

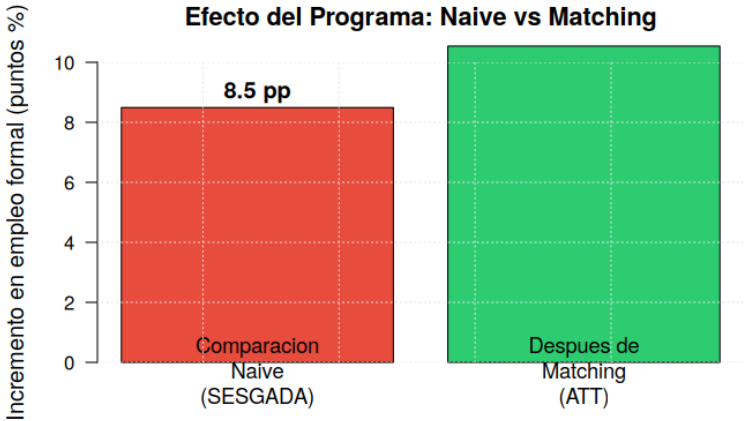
- Edad: SMD $= 0.02$ ✓
- Escolaridad: SMD $= 0.03$ ✓
- Experiencia: SMD $= 0.01$ ✓

Exito!

SMD < 0.1 para todas las variables

Los grupos ahora son comparables

Resultados: Naive vs Matching



Limitaciones de Matching

Ventajas:

- Intuitivo
- Transparente (podemos verificar balance)
- No requiere aleatorizacion

Limitaciones IMPORTANTES:

1. **CIA es fuerte** - si hay confusores no observados, falla
2. **Solo controla observables** - lo no medido sesga
3. **Requiere overlap** - sin common support no funciona
4. **Pierde datos** - controles no matched se descartan

Pregunta clave

¿Tengo medidas de TODAS las variables confusoras?
Si NO, considera DiD, RDD, o IV

Matching vs DiD

	Matching	**DiD**
Datos	Cross-section	Panel
Supuesto	CIA	Tendencias paralelas
Controla	Observables	Fijos + tendencias
Ventaja	No necesita panel	Controla no-observables

Combinables

DiD + Matching = robusto a ambos problemas

Resumen: 4 Puntos Clave

1. **Matching** = comparar similares en observables
2. **Propensity Score** = resume info en 1 numero
3. **Supuestos:** CIA + Common Support
4. **Verificar balance:** SMD < 0.1

Mensaje final

Matching es poderoso PERO solo si CIA es plausible

Codigo Completo

```r
library(MatchIt)
library(cobalt)

# Matching
m <- matchit(tratamiento ~ edad + mujer + escolaridad +
               urbano + experiencia_previa,
             data = datos,
             method = "nearest")

# Balance
bal.tab(m, thresholds = c(m = 0.1))

# Datos matched
datos_m <- match.data(m)

# ATT
att <- mean(datos_m$outcome[datos_m$tratamiento==1]) -
       mean(datos_m$outcome[datos_m$tratamiento==0])

# Regresion (mas robusto)
modelo <- lm(outcome ~ tratamiento,
             data = datos_m,
             weights = weights)
summary(modelo)
```

Matching

Comparar Similares

Dr. Sergio Bejar Lopez
CIDE