Condicionando en Observables

2025-04-23

Introducción

A veces no tenemos (o no podemos tener) una evaluación aleatorizada.

- Problemas éticos con la aleatorización
- El programa ya ocurrió
- Queremos aprender a partir de la variación existente

Pero aún así queremos aprender sobre efectos del tratamiento.

Supuesto de Independencia Condicional

Supongamos que tenemos datos sobre un tratamiento D y queremos conocer su efecto causal sobre un resultado Y.

No tenemos aleatorización, pero podemos observar todas las variables X que causan sesgo de selección.

Cuál és la hipótesis apropiada?

a) $(Y^1, Y^0) \perp D$
b) $Y \perp D|X$
c) $(Y^1, Y^0) \perp D|X$
d) $(Y^1, Y^0) \perp X|D$
e) $Y \perp (X, D)$

Supuesto de Independencia Condicional

Si creemos que al condicionar en X se cierran todos los caminos de backdoor, entonces podemos obtener los efectos causales.

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

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

Supuesto de Independencia Condicional

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

Interpretación:

1. No hay asociación entre el tratamiento y los potenciales resultados *más allá de* X
2. *Entre individuos con el mismo valor de* X, el tratamiento es independiente de los resultados potenciales.
3. *Condicionando en* X, el tratamiento es tan bueno como aleatorio.

Condicionar en X

¿Qué significa condicionar en X?

En teoría, significa simplemente que mantenemos X constante al variar D.

Lo que queremos calcular es:

$$\begin{aligned} E[Y^1 - Y^0] =& E[E[Y^1 - Y^0|X]] \\ =& E[E[Y^1|X] - E[Y^0|X]] \\ =& E[E[Y|D=1,X] - E[Y|D=0,X]] \\ =& \sum_x Pr(X=x)(E[Y|D=1,X] - E[Y|D=0,X]) \end{aligned}$$

Sin embargo, en la práctica, hay muchas formas de hacerlo.

Estimación

Hay muchos enfoques distintos para lidiar con selección en observables.

Todos dependen del mismo supuesto (CIA).

Pero tienen ventajas y desventajas.

- Regresión
- Subclasificación
- Emparejamiento
  - Emparejamiento exacto
  - Vecino más cercano
- Métodos basados en el propensity score
  - Regresión
  - Ponderación
  - Emparejamiento

Subclasificación

La subclasificación es un método más antiguo, pero va directo al punto correcto.

- Queremos mantener fija alguna característica, así que dividimos la muestra por valores de X.
- Calculamos el efecto del tratamiento en cada valor de X como $E[Y|D = 1, X = x] - E[Y|D = 0, X = x]$
- Agregamos según la distribución de X.

Subclasificación

Hagamos un ejercicio en R para ver cómo funciona esto.

Nos interesa estudiar las tasas de supervivencia en el Titanic.
- Ser pasajero de primera clase está asociado a una mayor probabilidad de sobrevivir.
- Pero hay más mujeres en primera clase, y las mujeres también tenían mucha más probabilidad de sobrevivir.
- Por lo tanto, para entender el impacto causal de estar en primera clase, debemos controlar por sexo.

El supuesto clave es que sexo y edad son los únicos confundidores relevantes.

Subclasificación

Una vez que tenemos los efectos por grupo, $E[\delta_i|X = x]$, eso ya nos dice toda la información sobre los efectos causales. El paso de agregación es solo una forma de condensar la información.

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

$$ATE = \sum_x P(X = x)E[\delta|X = x]$$

$$ATT = \sum_x P(X = x|D = 1)E[\delta|X = x]$$

$$ATU = \sum_x P(X = x|D = 0)E[\delta|X = x]$$

Subclasificación

Para calcular $E[Y|D=1, X=x] - E[Y|D=0, X=x]$, 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 tuviéramos la edad exacta de cada pasajero. Seguramente habría muchas edades en las que no encontraríamos personas de todos los grupos para comparar.

Subclasificación

La maldición de la dimensionalidad: A medida que crece la dimensión de X, se vuelve más difícil encontrar soporte común.

Imagina que tuviéramos no solo edad y sexo, sino también raza, religión, profesión, región de origen. El número de grupos a comparar crece exponencialmente con la cantidad de características.

Subclasificación: Resumen

- Supuestos: CIA, soporte común
- Se calculan diferencias por grupo y luego se agregan con pesos muestrales
- Se puede calcular ATE o ATT (dependiendo de los pesos)
- Maldición de la dimensionalidad: difícil de aplicar con múltiples controles continuos
- No se usa mucho hoy en día, pero es similar al emparejamiento

Regresión

Regresión

El enfoque más común para controlar por confundidores observables es la regresión.

Muy fácil: simplemente se incluye X como control.

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

Regresión

Recuerda que en regresión múltiple, el τ puede interpretarse como proveniente de una regresión de residuos:

$$\tilde{Y}_i = \tilde{\alpha} + \tau \tilde{D}_i + u_i$$

$$D_i = \gamma_0 + \gamma_1 X_i + \tilde{D}_i$$

Así, cualquier diferencia sistemática relacionada con X_i no se carga sobre D_i.

Regresión

La regresión maneja bien algunos de los problemas de la subclasificación.

- La linealidad de los parámetros permite tratar naturalmente variables continuas.
- La maldición de la dimensionalidad no es tan problemática, porque imponemos separabilidad en los coeficientes.

Sin embargo, las cosas son un poco más complicadas de lo que parecen.

Regresión

¿Recupera la regresión el ATE o el ATT (o algo distinto)?

Regresión

¿Recupera la regresión el ATE o el ATT (o algo distinto)?

Resulta que es más complicado.

Ejemplo: imagina un programa con aleatorización estratificada. Hay solo dos grupos, A y B. Fue más caro tratar al grupo B, así que solo el 20% recibió tratamiento.

X	D	E[Y|X,D]	N
A	0	2	50
A	1	5	50
B	0	5	80
B	1	6	20

- ¿Cuál es el efecto de tratamiento para cada grupo?
- ¿Cuál es el ATT?

X	D	E[Y]	N
A	0	2	50
A	1	5	50
B	0	5	80
B	1	6	20

Es fácil ver que el efecto de tratamiento es 3 para el grupo A y 1 para el grupo B. ATE $= 2$.

```
##
## Call:
## lm(formula = Y ~ D + X, data = df)
##
## Coefficients:
## (Intercept)            D            X
##       2.390        2.220        2.366
```

Resulta que la estimación es $2.2 > 2$. ¿Qué está pasando?

Regresión

Cuando usamos regresión para controlar, el estimador es un promedio ponderado de los efectos del tratamiento, dando más peso a los grupos con mayor varianza en el tratamiento.

En este caso:

$$\beta = \frac{\tau_A w_A + \tau_B w_B}{w_A + w_B}$$

$$w_S = n_S \text{Var}(D|X = S) = n_S p_S (1 - p_S)$$

Los grupos A y B tienen el mismo tamaño, pero A tiene mayor varianza en el tratamiento, así que recibe más peso.

Entonces, la regresión estima un promedio ponderado de efectos del tratamiento, pero los pesos no corresponden a la proporción poblacional.

¿Es esto un problema?

Entonces, la regresión estima un promedio ponderado de efectos del tratamiento, pero los pesos no corresponden a la proporción poblacional.

¿Es esto un problema?

1. La estimación no representa exactamente a ningún grupo de interés.
2. Pero siempre es un "valor intermedio" entre los distintos efectos del tratamiento.
3. En la práctica, puede que no haga mucha diferencia.
4. Por lo general, mientras se pueda obtener un promedio ponderado de efectos del tratamiento, eso es "suficientemente bueno".

Si nos preocupan los efectos heterogéneos, podemos aliviar este problema usando formas funcionales más flexibles:

```
##
## Call:
## lm(formula = Y ~ D + X + X * D, data = df)
##
## Coefficients:
## (Intercept)            D            X          D:X
##           2            3            3           -2
```

La regresión no sufre tanto por la dimensionalidad porque **impone linealidad**.

Si especificamos mal la forma funcional del control, puede que no cerremos adecuadamente los backdoors.

Si hay relaciones altamente no lineales, la regresión puede no ser el método más adecuado.

Este problema puede aliviarse usando especificaciones más flexibles, pero en la práctica a menudo es mejor usar otros métodos en lugar de correr regresiones muy complicadas.

Regresión: Resumen

- Supuesto clave: CIA, forma funcional
- Estima un efecto promedio ponderado por varianza (VWATE)
- Ventajas: simplicidad, flexibilidad, puede manejar confundidores de alta dimensión
- Desventaja: depende de la forma funcional, el VWATE no tiene una interpretación clara

Emparejamiento

Emparejamiento

Para calcular efectos del tratamiento, queremos una forma de 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) unidad(es) de control a cada unidad tratada, basándonos en la "similitud".

Emparejamiento

En general, llamemos a la unidad emparejada con i como $m(i)$. Entonces, $Y_{m(i)}$ es el resultado de la unidad emparejada. El estimador de emparejamiento más simple es:

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

Hay muchas formas de elegir una pareja. Discutiremos algunas de ellas:

- Emparejamiento exacto
- Emparejamiento por vecino más cercano
- Emparejamiento con caliper

Emparejamiento: Emparejamiento Exacto

El caso más simple de emparejamiento es el emparejamiento exacto.

Esto significa que cada unidad tratada se empareja con una unidad de control que tiene exactamente los mismos valores de X.

Trainees					
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075		31.95	$11,101.25

Figure 1: Treated (Left) and Untreated (Right)

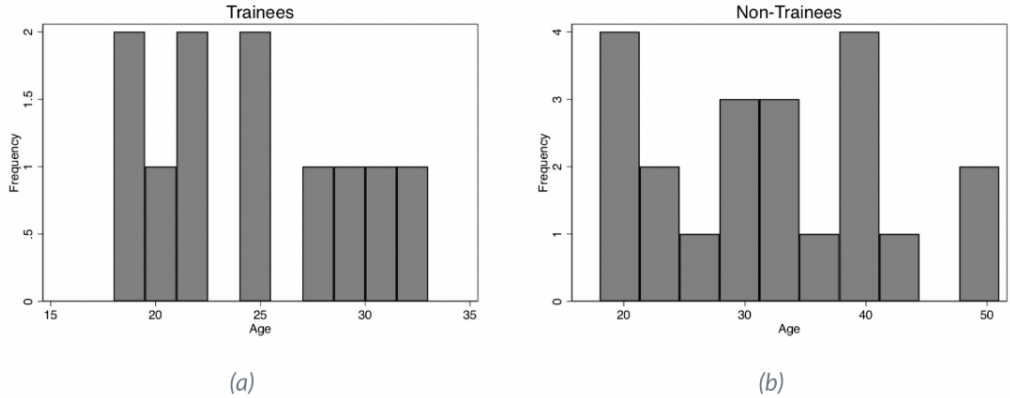
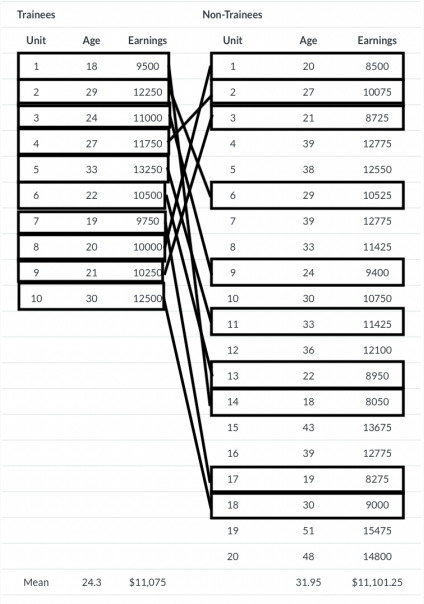


Figure 5.1: Covariate distribution by job trainings and control.

Figure 2: Age distribution is unbalanced



Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075		31.95	$11,101.25

Figure 3: Exact Matching

Unit	Age	Earnings
1	18	9500
2	29	12250
3	24	11000
4	27	11750
5	33	13250
6	22	10500
7	19	9750
8	20	10000
9	21	10250
10	30	12500
Mean	24.3	$11,075

Unit	Age	Earnings
14	18	8050
6	29	10525
9	24	9400
8	27	10075
11	33	11425
13	22	8950
17	19	8275
1	20	8500
3	21	8725
10,18	30	9875
	24.3	$9,380

Figure 4: Treated (Left) and Matched Controls (Right)

En este caso, estamos estimando el ATT, ya que estamos emparejando cada unidad *tratada* con una unidad de control, pero eliminando otros no tratados.

Si queremos el ATE, podemos emparejar cada tratado con un control, y cada no tratado con un tratado:

$$\hat{\delta}_{ATE} = \frac{1}{N}\left(\sum_{D_i=1}(Y_i - Y_{m(i)}) + \sum_{D_i=0}(Y_{m(i)} - Y_i)\right)$$

Emparejamiento Exacto

En algunos casos, puede que no encontremos una pareja para cada unidad. ¿En qué caso obtenemos una estimación no sesgada del ATT?

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

Emparejamiento Exacto

A veces también podemos tener muchas unidades con valores idénticos de X, que podrían ser posibles parejas.

¿Qué deberíamos hacer en ese caso?

Emparejamiento Exacto

A veces también podemos tener muchas unidades con valores idénticos de X, que podrían ser posibles parejas.

¿Qué deberíamos hacer en ese caso?

Algunas posibilidades:

- Tomar el promedio del resultado de todos los matches para crear una "unidad virtual emparejada".
- Incluir todos los matches, pero ponderar cada unidad de control por $\frac{1}{N}$ donde N es el número de emparejamientos.
- Elegir aleatoriamente una pareja del grupo de matches posibles.

Emparejamiento Exacto

El emparejamiento exacto funciona mejor cuando X es discreta y de baja dimensión.

- Puede haber rápidamente problemas de soporte común.

Si estamos condicionando en muchas variables, algunas unidades tratadas quedarán sin pareja. El efecto promedio del tratamiento entre las unidades emparejadas puede diferir del ATT.

Para covariables continuas, podemos discretizar el espacio y hacer emparejamiento exacto, pero usualmente es mejor pasar al siguiente enfoque.

Emparejamiento por Vecino Más Cercano

Supongamos que tenemos una variable verdaderamente continua para condicionar, como ingresos antes de la intervención.

- Es muy difícil encontrar emparejamientos exactos, al nivel del peso.

Pero simplemente podemos buscar la unidad con el valor más cercano de ingresos.

Emparejamiento por Vecino Más Cercano

Primero, necesitamos entender qué significa “cercano”.

Si solo hay una variable que debemos controlar, podemos tomar la unidad con menor diferencia.

Pero si hay más variables, ¿qué hacemos?

Emparejamiento por Vecino Más Cercano

Primero estandarizamos cada variable.

Esto hace que cada variable tenga la misma “importancia”.

Pregunta importante: ¿deberíamos estandarizar cada grupo por separado o todos juntos?

Ahora calculamos la distancia de cada unidad tratada a cada unidad de control, usando la Distancia Euclidiana:

$$d_E(X_i, X_j) = \sqrt{(X_i - X_j)\hat{V}^{-1}(X_i - X_j)}$$

Y escogemos la unidad de control con menor distancia.

Hay muchas variaciones de esto.

Primero, en lugar de la distancia euclidiana, la métrica más común es la distancia de Mahalanobis. La misma idea, pero considerando varianzas y covarianzas.

Euclidiana:

$$d_E(X_i, X_j) = \sqrt{(X_i - X_j)\hat{V}^{-1}(X_i - X_j)}$$

Distancia de Mahalanobis:

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)\hat{\Sigma}^{-1}(X_i - X_j)}$$

Donde $\hat{V}$ es una matriz diagonal con varianzas, y $\hat{\Sigma}$ es la matriz de varianza-covarianza.

Emparejamiento

A veces, el control más cercano no está realmente tan cerca. Tal vez las diferencias son demasiado grandes.

- Podemos aceptar que algunas unidades no tengan buenos controles: parecido a no tener soporte común.
- A veces el promedio de los 3 controles más cercanos puede ser una mejor comparación que el vecino más cercano.

Emparejamiento con Caliper

Enfrentamos una disyuntiva entre sesgo y varianza:

- Incluir más unidades de comparación nos permite mantener más observaciones $\rightarrow$ más precisión
- Pero los peores emparejamientos introducen sesgo

Una forma de manejar esto: incluir todos los emparejamientos hasta una distancia máxima D, llamada *caliper*.

Así, conservamos pocos emparejamientos si no hay buenos, y muchos si los hay.

Otra forma de lidiar con la calidad variable de los emparejamientos es ponderar.

Idea básica: ponderar las unidades de control según qué tan cercanas estén. Los buenos emparejamientos reciben más peso; los no tan buenos reciben menos, pero aún cuentan.

Esto se llama *kernel weighting*.

Kernel Weighting

Supongamos que d es la distancia entre una unidad de control potencial y una unidad tratada. Bajo ponderación con kernel, esta unidad recibe peso:

$$K(d) = \max\left(\frac{3}{4}(1 - d^2), 0\right)$$

Así, las unidades con distancia cero reciben el mayor peso, y las unidades con más de 1 unidad de distancia reciben peso 0. Nota: esto se aplica a distancias estandarizadas.

Entonces:

$$Y_{m(i)} = \frac{\sum_{D_j=0} K(d_{ji}) Y_j}{\sum_{D_j=0} K(d_{ji})}$$

Emparejamiento

¿Qué pasa si hay múltiples emparejamientos exactos?

- Podemos usar todos y asignar peso $1/M$ a cada uno (o usar el promedio de Y)
- En la práctica, a veces está bien escoger uno aleatoriamente del conjunto

¿Deberíamos hacer el emparejamiento con o sin reemplazo?

- Está bien permitir reemplazo, para que una unidad sirva como control de más de un tratado.
- Puede ser problemático si mucho peso cae sobre unas pocas unidades: podrían ser *outliers*.

Emparejamiento

En algunos contextos, el emparejamiento puede ser computacionalmente intensivo.

Hay que calcular la distancia entre cada unidad tratada y cada unidad de control. Para bases de datos muy grandes, esto puede volverse un problema.

Resumen

Regresión vs Emparejamiento:

Similitudes:

- Ambos dependen de la Independencia Condicional para la identificación.
- Dos formas de cerrar caminos de backdoor.
- Si hay confundidores no observados, ambos tienen problemas.

Diferencias:

- La regresión es más simple, implica menos decisiones arbitrarias.
- La regresión tiene menos problemas con la dimensionalidad.
- El emparejamiento no depende de la linealidad como la regresión.
- La regresión estima un parámetro ponderado por la varianza.

Propensity Score

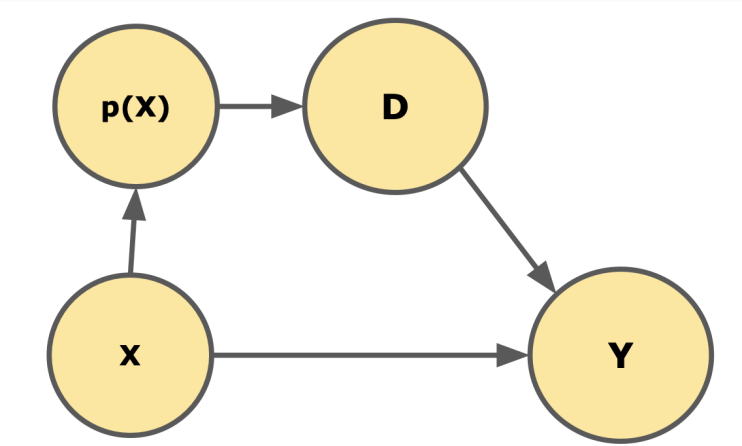
Propensity Score

Los métodos basados en propensity score ayudan a resolver algunos de los problemas del emparejamiento con covariables.

La idea principal es que, en lugar de controlar por un vector grande de características X_i, podemos resumirlo y controlar solo por un escalar $p(X)$.

Esto reduce la dimensionalidad del control y ayuda con el problema de la maldición de la dimensionalidad.

La idea clave detrás del método es que necesitamos controlar por X porque afecta la probabilidad de tratamiento. Pero podemos **estimar esta probabilidad** y controlar *directamente* por ella.



Propensity Score

La estimación procede así:

1. Estimar la probabilidad de tratamiento dado X, haciendo una regresión de D sobre X usando un Probit o Logit. Obtener las probabilidades predichas.
2. Estimar el efecto de D sobre Y controlando por p(X).

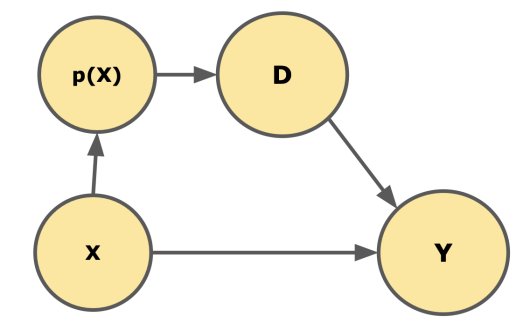
Hay varias formas de hacer el paso 2.

Propensity Score

Todos los métodos de propensity score dependen de estimar correctamente la función de propensity score.

- Necesitamos obtener correctamente las probabilidades predichas de tratamiento condicional en X.

Considerando el diagrama causal, ¿cómo podemos probar si p(x) está bien estimado?



Propensity Score

El propensity score debe mediar el camino de X a D.

Por lo tanto, condicional en p(X), no debe haber relación entre X y D.

Necesitamos probar si la muestra está **balanceada**.

Usualmente se prueba el balance de forma parecida a un experimento: con una tabla de balance.

- Podemos probar el balance **condicional al PS** observando dentro de grupos con valores similares del PS.

¿Qué pasa si hay desbalance?

Entonces necesitamos volver atrás e intentar otra especificación en la estimación del PS.

- Más variables, diferente forma funcional, interacciones, etc.

Otra cosa que debemos revisar es el Soporte Común en el PS.

Si la mayoría de los tratados tienen valores muy altos de PS, y la mayoría de los controles tienen valores muy bajos, con poco solapamiento, es probable que el grupo de control no sea adecuado.

A menudo tiene sentido limitar la estimación al grupo con solapamiento en los valores.

Propensity Score + Regresión

Ahora, ¿cómo estimamos los efectos?

Una posibilidad es simplemente hacer una regresión de Y sobre D y p(X).

No se usa mucho en la práctica porque queremos controlar de forma flexible. Imponer linealidad en p(X) suele ser demasiado restrictivo y no suficiente para cerrar los backdoors.

Propensity Score + Emparejamiento

Otra posibilidad es emparejar usando el propensity score.

El emparejamiento por vecino más cercano con el propensity score es tal vez el método más usado, especialmente en ciencias médicas.

De forma similar, podemos hacer **emparejamiento exacto discretizado**, que simplemente agrupa unidades por intervalos de PS.

Propensity Score + Ponderación

Otra cosa común es construir una muestra de control ponderando las unidades de control usando el propensity score.

Una forma de hacerlo es aplicar la misma idea del kernel matching al propensity score.

Propensity Score + IPW

La segunda idea es ponderar usando *pesos inversos de probabilidad* (IPW).

Es decir, cada unidad recibe un peso proporcional al *inverso* de la probabilidad de haber recibido la asignación que tuvo.

Supón que una unidad tiene PS de 0.8. Si fue tratada, su peso sería $\frac{1}{0.8}$. Si no fue tratada, su peso sería $\frac{1}{1-0.8} = \frac{1}{0.2}$.

¿Por qué tiene sentido el IPW?

Recuerda: el problema que queremos resolver es que hay asociación entre X y D. El tratamiento no está balanceado.

Tomemos un grupo con PS $= 0.8$. Si hay 10 unidades, 8 son tratadas. Una vez que aplicamos IPW, las unidades tratadas tienen un peso combinado de 1, y los controles también tienen un peso combinado de 1.

Como esto ocurre en todos los valores del PS, el tratamiento queda "balanceado".

El estimador IPW es entonces:

$$\hat{\tau} = \frac{1}{N} \sum \left[T_i \frac{1}{p(X_i)} Y_i - (1 - T_i) \frac{1}{1 - p(X_i)} Y_i \right]$$

Propensity Score + IPW

El estimador IPW a veces tiene problemas con valores extremos de p(X), ya que estamos dividiendo por un número muy pequeño.

- Es común lidiar con esto haciendo "trimming" de los valores con PS cercano a 0 o 1.

Este método también tiene una gran ventaja: no requiere un componente de emparejamiento, solo ponderar la muestra, así que es más fácil de usar.

Resumo

- **Regresión**: Depende de forma funcional correcta. Estima VWATE. Simple y transparente, pero VWATE no tiene relevancia clara.
- **Matching**: Depende de Soporte Común. Normalmente estima ATT. No necesita de modelaje. Dificuldades con muchas variables de controle. Muchas decisiones arbitratias.
- **Propensity Score**: Depende de estimar correctamente el PS, y de Soporte Común. Puede estimar ATE o ATT. Apenas una dimensión para controlar. Importante checar equilibrio.