Condicionando en Observables

2025-04-23

Introducción

Para la gran mayoría de las preguntas de investigación, es dificil de usar experimentos aleatorizados.

Pero aún así queremos aprender sobre efectos de tratamiento.

Supuesto de Independencia Condicional

El primer paso es usar teoría y conocimiento del contexto para entender que variables, X, poden causar un sesgo.

Si creemos que al condicionar en X se cierran todos los caminos de backdoor, entonces podemos obtener los efectos causales. Cuando condicionamos en la variable X, escribimos "|X".

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.
4. No existen caminos de backdoor después que cierramos los caminos por X.

Condicionar en X

¿Como podemos usar ese supuesto?

Vamos imaginar un programa del gobierno, que te encargan de evaluar. Para tanto tiene solamente la seguinte información:

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

¿Cuál es tu estimativa del efecto de tratamiento? ¿Bajo cuál suspuesto?

Condicionar en X

Ahora supón que el programa fue administrado en dos tipos de areas: urbanas y rurales, y las comunas urbanas recibiron 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?

Condicionar en X

Ahora vamos cambiar un poco los datos. Y si los datos fueran así:

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

¿Cuál es el efecto promedio de tratamiento en ese caso? ¿Y si queremos alicar eses resultados para todo Chile, donde 88% de la población es urbana?

Condicionar en X

Para cada valor de x, calculamos:

$$\hat{\tau}_x = \frac{\sum Y \cdot 1(D=1, X=x)}{\sum 1(D=1, X=x)} - \frac{\sum Y \cdot 1(D=0, X=x)}{\sum 1(D=0, X=x)}$$

Lo que queremos calcular es:

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

Quales probabilidades usamos ($Pr(X=x)$) determinan sobre cual población estamos hablando.

Estimación

Este caso simple con una unica covariada binária es fácil, pero muchas veces tenemos problemas más complejos.

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 exatamente lo hicimos antes.

- 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:

$$\hat{\tau}_x = \frac{\sum Y \cdot 1(D = 1, X = x)}{\sum 1(D = 1, X = x)} - \frac{\sum Y \cdot 1(D = 0, X = x)}{\sum 1(D = 0, X = x)}$$

- Agregamos según la distribución de X.

Subclasificación

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

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

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

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

Subclasificación

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 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 urbano/rural, pero también edad y sexo, 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 contínuos

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

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.

Regresión

La regresión recupera un único efecto de tratamiento $\hat{\tau}$. ¿Es un promedio que corresponde a que población?

¿Es el ATE? ¿El ATT? ¿Otra cosa?

Regresión

Cuando usamos regresión para controlar covariadas, 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:

$$\hat{\tau} = \sum \hat{\tau}_x w_x$$

$$w_x \propto n_x \text{Var}(D|X = x) = n_x p_x (1 - p_x)$$

Si grupos A y B tienen el mismo tamaño, pero A tiene proporcion del tratamiento más cerca de 50%, entonces A recibe más peso.

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

Si especificamos mal la forma funcional del control, puede ser 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 unidades 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: 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: Tratados (izquierda) y no-tratados (derecha)

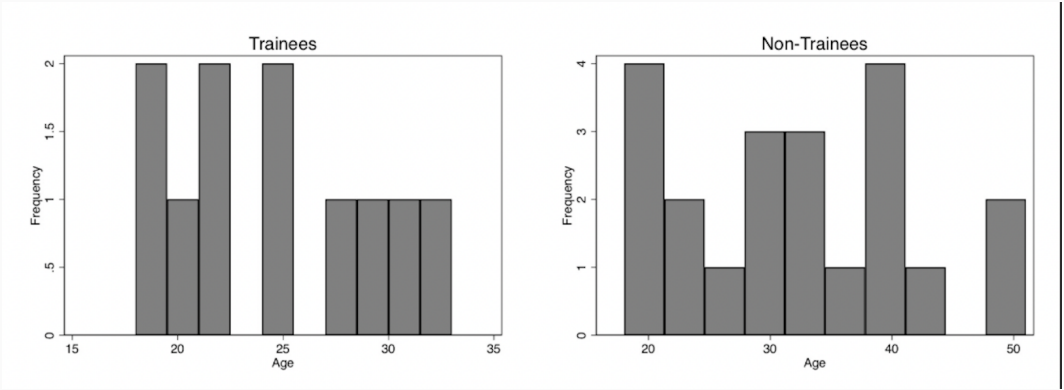
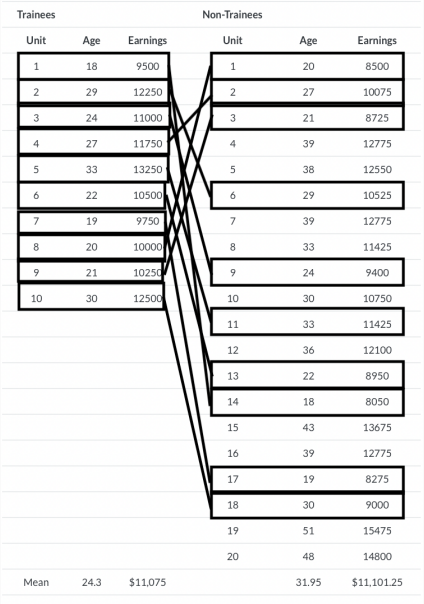


Figure 2: La distribución de edad no está balanceada



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: Matching exacto

Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	14	18	8050
2	29	12250	6	29	10525
3	24	11000	9	24	9400
4	27	11750	8	27	10075
5	33	13250	11	33	11425
6	22	10500	13	22	8950
7	19	9750	17	19	8275
8	20	10000	1	20	8500
9	21	10250	3	21	8725
10	30	12500	10,18	30	9875
Mean	24.3	$11,075		24.3	$9,380

Figure 4: Tratados (Left) y controles emparejados (derecha)

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".

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

$$d_E(X_i, X_j) = \sqrt{\sum_k (X_i^k - X_j^k)^2}$$

Y escogemos la unidad de control con menor distancia.

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.

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

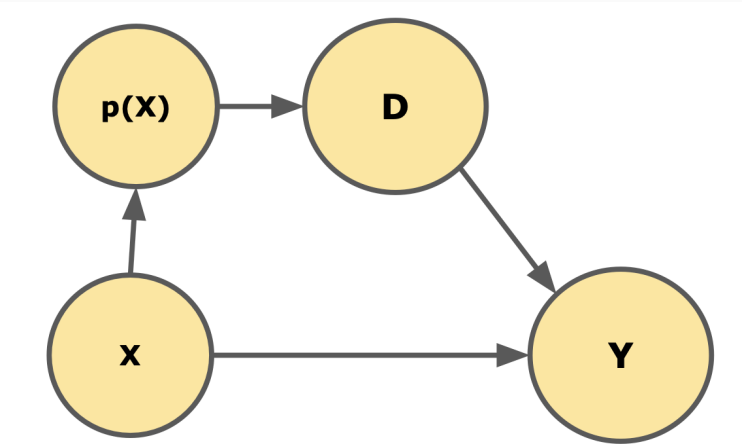
Propensity Score

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).

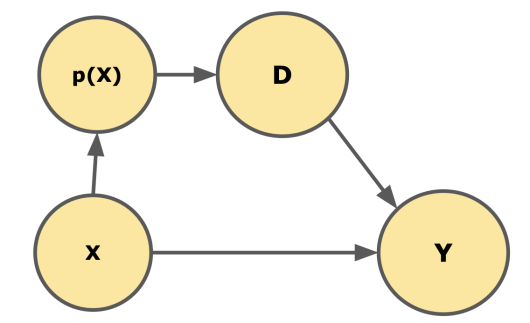
Hay varias formas de hacer el paso 2.

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.

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.