AYUDANTÍA 2 STATA: MATCHING

Economía y Evaluación de las Políticas Sociales

Francisca Cuadros (francisca.cuadros@uc.c)

En base a la ayudantía de Vicente Munita 2020

7 de septiembre 2022

INTRODUCCIÓN

- Paper en que se basa la ayudantía: The Costs of Low Birth Weight - Almond et al. (2005)
- Buscan estimar si el que las madres fumen durante el embarazo afecta el peso al nacer de sus hijos.
- ¿Por qué? Los niños/as con peso al nacer de menos de 2,500 gramos presentan diversas dificultades de salud y desarrollo, lo cual puede imponer costos a la sociedad.
- Además, la literatura señala que el bajo peso al nacer esta asociado a niveles educativos, de empleabilidad e ingresos más bajos.
- Por lo tanto, quieren testear si:

Fumar en el embarazo → Menor peso al nacer

1) Es menos directamente atribuible a la genética.
2) No es una consecuencia de la salud inherente del feto
3) Principal causa modificable de BPN en EE. UU

BASE DE DATOS

- 4.642 observaciones de recién nacidos en Pennsylvania.
- Variables principales:
  - bweight: peso al nacer
  - mbsmoke: =1 si la madre fumaba en el embarazo, =0 si no.
- STATA: Abriremos la base matching.dta

EFECTO CAUSAL

- ¿Que queremos estimar? El efecto causal de fumar sobre el peso de recién nacidos:

$$\tau = E[Peso_{fumar} \mid fumo] - E[Peso_{no\ fumar} \mid fumo]$$

Contrafactual

- No observamos el contrafactual, pero si observamos los pesos de lo recién nacidos de madres que no fumaron: $E\left[Peso_{no\ fumar} \mid no\ fumo\right]$.
- Por lo tanto, tenemos:

Peso de recién nacidos de madres fumadoras

Peso de recién nacidos de madres no fumadoras

¿Entonces como identificamos el efecto causal?

OPCIONES

1) MCO:

- Es necesario que se cumpla el supuesto de $E(u_i/x_i) = 0$
- Problema de variables omitidas.

2) Aleatorización:

- Ética
- Necesitaríamos balance en observables → test de medias
- STATA: crearemos una matriz que nos indique la media en el grupo control, la media en el grupo tratamiento y el valor p para cada variable.

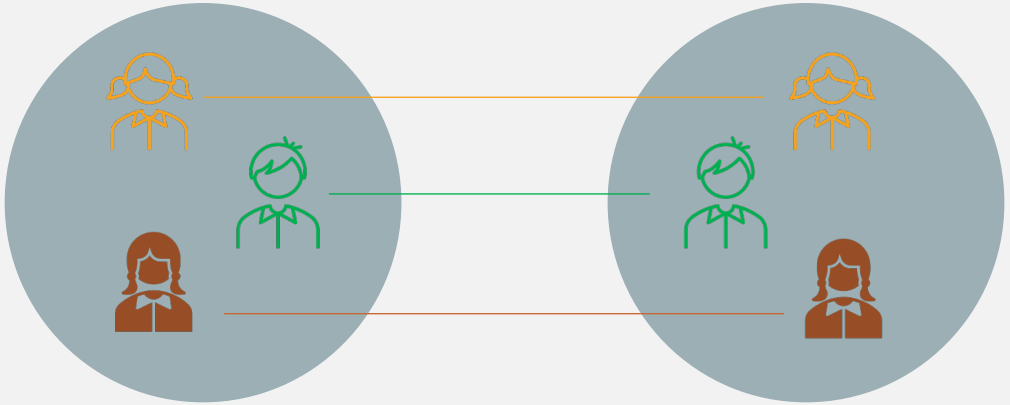
TEST DE MEDIAS

```
.
.          mat list A // nos muestra la matriz que creamos

A[16,3]
           Media Ctrls   Media Trat      p-value
  bweight    3412.9116    3137.6597    4.684e-37
 mmarried     .7514558     .47337963    7.918e-60
    mhisp    .03626257    .02430556    .08038386
    fhisp    .03785071    .03356481    .54748828
  foreign    .05982001    .02546296    .00005038
  alcohol    .01879301    .09143519    5.881e-28
     mage    26.810482    25.166667    7.158e-15
     medu    12.929857    11.638889    8.048e-43
     fage    27.844362    24.743056    1.075e-18
     fedu    12.673902    10.703704    1.293e-46
 nprenatal    10.962943    9.8622685    1.795e-15
    mrace    .84780307    .80902778    .00496438
    frace    .82689254    .75578704    1.252e-06
 prenatal    1.1776072    1.3078704    9.417e-12
  mage_sq     750.6649     661.43287    7.460e-15
  fage_sq    852.62361    736.35185    1.749e-12
```

- No hay balance de observables
- La asignación no es experimental → **Solución: usar matching**

MATCHING



recién nacidos de madres fumadoras

recién nacidos de madres no fumadoras

Idea: para cada persona nacida, cuya madre fumó durante su embarazo, buscamos al mejor par según características observables. (misma edad de la madre, misma raza, mismo estado civil, etc.)

- Aplicación por default: 1 "vecino" mas cercano.
- Una vez encontrado el "clon", podemos usar ese peso al nacer de la persona del grupo control y comparamos con la persona tratada. La resta sería el impacto para esa persona y si hacemos esto para cada persona del grupo tratamiento y promediamos el impacto se obtiene el ATT (Averege Treatment Effect on Treated) .

SUPUESTOS DE IDENTIFICACIÓN

Ignorabilidad Fuerte:

1) **Independencia condicional**: pertenecer al grupo tratamiento solo se explica por características observables:

$$\{PesoFumar , PesoNoFumar \} \perp Fumar|OBSERVALES$$

2) **Overlap:** Para cada vector de observables, cualquier persona podría ser parte del tratamiento o del control:

$$0 < Pr (Fumar = 1|OBSERVABLES) < 1$$

Pero… es casi imposible encontrar "clones" perfectos: maldición de la dimensionalidad
Solución → **Propensity Score Matching**

PROPENSITY SCORE MATCHING

- **Idea**: Para cada unidad del grupo de tratamiento y del grupo control se computa la probabilidad de que esta unidad sea parte del tratamiento (puntaje de propensión) dadas las características observables.
- El propensity score se define como la probabilidad de que el individuo i haya sido sujeto de tratamiento, definida como: **$p(X) \equiv Prob(Fumar = 1|X)$**, donde X son las características observables.
- Mismos supuestos: Ignorabilidad Fuerte

Se reduce el problema a una dimensión: ahora en vez de hacer matching en base a todas las características observables, se hace solo con el pscore:

$\{PesoFumar, PesoNoFumar\} \perp Fumar|X \rightarrow \{PesoFumar, PesoNoFumar\} \perp Fumar|p(X)$

PSM EN STATA

Pasos:

1) Elegir características observables significativas al 5% de significancia (regla del pulgar) → hace sentido usarlas para calcular la probabilidad de que una madre fume en el embarazo.
2) Elegir modelo Probit o Logit (modelos no lineales)
3) Estimar el pscore
4) Soporte Común
5) Estimación ATT

PSM EN STATA

Paso 1:

Elegir las significativas
(p <0.05)

mbsmoke	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
mmarried	-.6562897	.0652798	-10.05	0.000	-.7842357	-.5283438
mhisp	-.4855616	.2090439	-2.32	0.020	-.8952802	-.0758431
fhisp	-.1049293	.1945735	-0.54	0.590	-.4862863	.2764277
foreign	-.4540175	.1368741	-3.32	0.001	-.7222858	-.1857492
alcohol	.9300851	.114736	8.11	0.000	.7052067	1.154964
mage	.2084491	.0365554	5.70	0.000	.1368018	.2800963
medu	-.0830806	.0117584	-7.07	0.000	-.1061266	-.0600346
fage	-.0070819	.0077821	-0.91	0.363	-.0223345	.0081707
fedu	-.0297242	.0085465	-3.48	0.001	-.046475	-.0129734
nprenatal	-.0172614	.0066032	-2.61	0.009	-.0302035	-.0043194
mrace	.2540817	.1276235	1.99	0.046	.0039443	.5042192
frace	.1440847	.1258745	1.14	0.252	-.1026247	.3907941
prenatal	.0861809	.0450463	1.91	0.056	-.0021082	.17447
mage_sq	-.0038662	.0006758	-5.72	0.000	-.0051908	-.0025416
fage_sq	.0002416	.0001675	1.44	0.149	-.0000866	.0005699
_cons	-2.047726	.4776453	-4.29	0.000	-2.983893	-1.111558

PSM EN STATA

Paso 2: elegir modelo según criterios de información de Akaike (AIC) o Bayesiano (BIC).

```
.
.         qui probit mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, r

.         estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	4,642	-2230.748	-1971.864	11	3965.728	4036.6

```
.
.         qui logit mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, r

.         estat ic

Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	4,642	-2230.748	-1975.93	11	3973.861	4044.733

Elegir modelo con menores AIC y BIC: en este caso es **PROBIT**

PSM EN STATA

Paso 3: estimar pscore

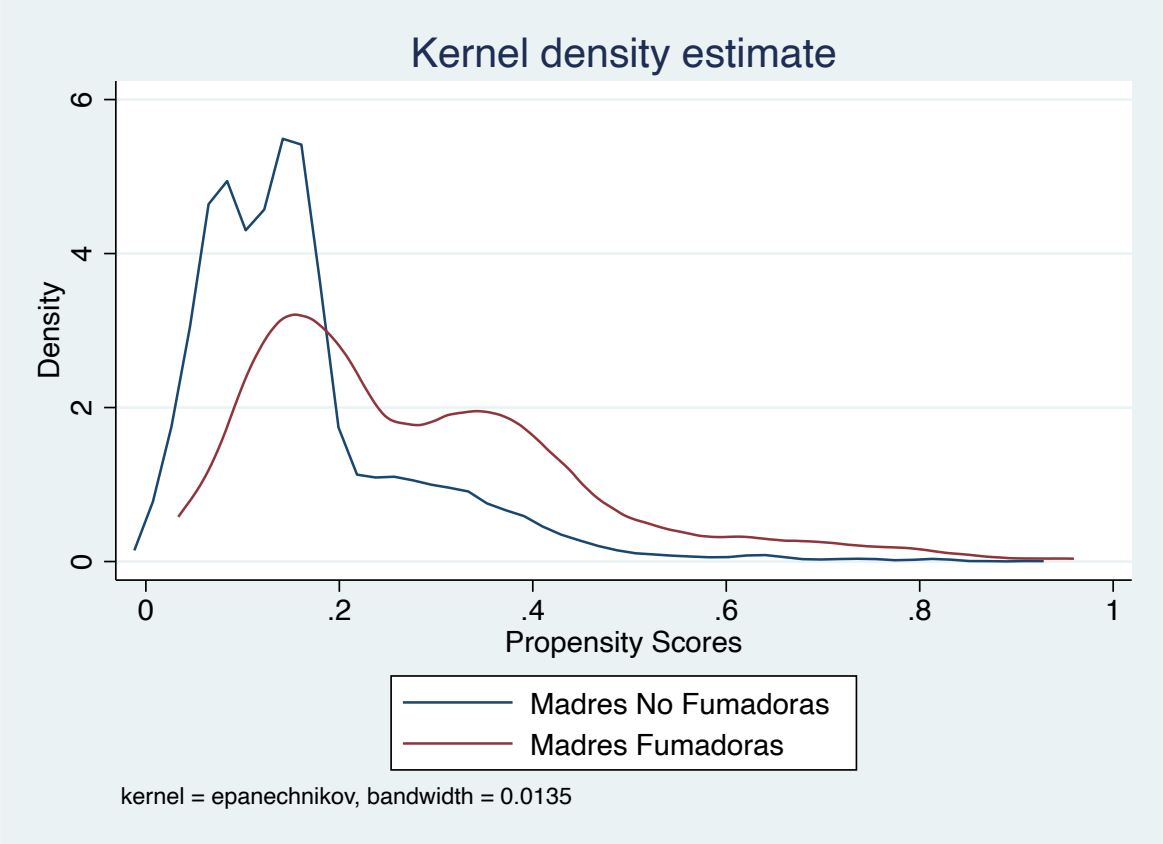
Modelo (probit en este caso) + predict:

```
// PASO 3: estimación pscores
probit mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, r
predict pscore // En vez de pscore pudimos haber escrito cualquier cosa
```

PSM EN STATA

Paso 4: Restringir al soporte común

Soporte común: zona en donde las distribuciones de los pscores del grupo tratado y control están presentes.



PSM EN STATA

Paso 4: Restringir al soporte común

Cotas:

Elegimos el valor mayor de los mínimos

```
. sum pscore if mbsmoke==0

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |      3,778    .1633588    .1164466   .0016356   .9146181

. sum pscore if mbsmoke==1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |        864    .2814133    .1692832   .0335256   .9593767
```

Elegimos el valor menor de los máximos

Ojo: Restringir la muestra al SC no tiene un fundamento estadístico claro. Una cosa es cumplir el supuesto de overlap y otra muy distinta es restringir la muestra al SC. No hay consenso en la literatura sobre si es una buena o mala práctica.

PSM EN STATA

Paso 4: ¿Se pierde representatividad?

```
. gen sop_comun=(pscore>=0.0335256&pscore<=0.9146181)

. sum sop_comun
```

Variable	Obs	Mean	Std. dev.	Min	Max
sop_comun	4,642	.9646704	.1846314	0	1

Más del 96% de la muestra esta dentro del soporte común, por lo tanto hay muy poca perdida de representatividad.

PSM EN STATA

Paso 5: Estimación ATT

teffects psmatch (var dep) (vart var1 var2 ... vark, modelo) if soporte comun==1, atet + otras opciones

- var dep: variable dependiente, en nuestro caso es bweight.
- vart: variable que indica el tratamiento, en nuestro caso es mbsmoke.
- var1 var2 ... Vark: es la lista de variables explicativas que se incluyeron en el modelo de probabilidad no lineal.
- modelo: especifca el modelo de probabilidad que usamos. El que va por defecto es logit, por lo que tendremos que escribir probit.
- if soporte comun==1 restringe la muestra al SC (podría no incluirse)
- atet indica que queremos el ATT (el efecto promedio sobre los tratados).
- En otras opciones podemos indicar el método de emparejamiento. Por defecto es 1 vecino más cercano y por el momento lo dejaremos así.

PSM EN STATA

Paso 5: Estimación ATT

```
. teffects psmatch (bweight) (mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, probit) if sop_comun==1, atet

Treatment-effects estimation                   Number of obs      =      4,478
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: probit                                    max =         20
```

bweight	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET						
mbsmoke (smoker vs nonsmoker)	-274.9279	33.24959	-8.27	0.000	-340.0959	-209.7599

Los hijos de madres que fuman durante su embarazo pesan, en promedio, 275 gramos menos que los hijos de las madres que no fumaron.

MÁS VECINOS CERCANOS

N vecinos más cercanos: opción nn(#) en teffects psmatch. Por defecto # = 1.

- ¿Qué ganamos si # > 1? → Precisión
- ¿Qué perdemos? → Consistencia (o insesgamiento), ya que comparamos con más personas, pero más distintas.

BALANCE POST MATCHING

Luego del matching, las características observables, entre los tratados y sus clones, debiesen estar balanceadas.

→ En general, matching funcionó. Esta tabla es una forma de validar el procedimiento.

A[15,3]

	Control	Tratamiento	P-Valor
mhisp	.46457607	.47337963	.60460757
fhisp	.01509872	.02430556	.07938684
foreign	.02787456	.03356481	.35360024
alcohol	.02671312	.02546296	.81570756
mage	.05691057	.09143519	.00045609
medu	25.074332	25.166667	.6088112
fage	11.557491	11.638889	.27002043
fedu	24.315912	24.743056	.26037169
nprenatal	10.490128	10.703704	.12581982
mrace	9.8106852	9.8622685	.71867069
frace	.81068525	.80902778	.90144285
prenatal	.757259	.75578704	.9198508
mage_sq	1.3263647	1.3078704	.38814874
fage_sq	656.49593	661.43287	.60492568
fage_sq	714.58304	736.35185	.1730174

. }

TEST PLACEBO

La idea es calcular el efecto sobre una variable que creemos no debería ser significativa.

```
. teffects psmatch (fhisp) (mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, probit) if sop_comun==1, atet

Treatment-effects estimation                Number of obs      =      4,478
Estimator      : propensity-score matching  Matches: requested =          1
Outcome model  : matching                               min =          1
Treatment model: probit                                 max =         20
```

	fhisp	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET							
	mbsmoke						
	(smoker vs nonsmoker)	**.0061169**	**.0082328**	**0.74**	**0.457**	**-.0100191**	**.022253**

No es significativo!

¿SE CUMPLEN SUPUESTOS?

1) Independencia condicional: se responde de manera conceptual. ¿Creen que hay otras variables que afectan la asignación del tratamiento y el peso de las personas al nacer?
2) Overlap: se puede testear viendo las distribuciones de los pscores.

Ojo: restringir la muestra al SC no garantiza que se cumpla el supuesto de Overlap.

MCO

```
. reg bweight mbsmoke, r

Linear regression                               Number of obs     =      4,642
                                                F(1, 4640)        =     168.33
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0343
                                                Root MSE          =     568.88

------------------------------------------------------------------------------
             |               Robust
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -275.2519   21.21501   -12.97   0.000    -316.8434   -233.6604
       _cons |   3412.912   9.285455   367.55   0.000     3394.708    3431.115
------------------------------------------------------------------------------
```

```
. reg bweight mbsmoke mmarried mhisp foreign alcohol mage medu fedu nprenatal mrace mage_sq, r

Linear regression                               Number of obs     =      4,642
                                                F(11, 4630)       =      40.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1040
                                                Root MSE          =     548.55

------------------------------------------------------------------------------
             |               Robust
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -222.1905   22.44463    -9.90   0.000    -266.1927   -178.1883
    mmarried |   41.61622   24.71439     1.68   0.092    -6.835765    90.0682
       mhisp |  -4.212862   44.16398    -0.10   0.924     -90.7953    82.36958
     foreign |  -8.613031   40.46731    -0.21   0.831    -87.94824    70.72218
     alcohol |  -9.751678   45.86251    -0.21   0.832    -99.66405    80.16069
       mage |   10.92957   12.29789     0.89   0.374    -13.18015    35.03928
       medu |  -1.881357   3.923223    -0.48   0.632    -9.572744    5.81003
       fedu |   1.123669   2.963235     0.38   0.705    -4.685684    6.933022
  nprenatal |   26.11472   2.810171     9.29   0.000     20.60545     31.624
      mrace |   220.6861   28.40962     7.77   0.000     164.9897    276.3825
     mage_sq |  -.1560589   .2214041    -0.70   0.481    -.5901165    .2779987
       _cons |   2743.303   163.2404    16.81   0.000     2423.274    3063.332
------------------------------------------------------------------------------
```

No es mucha la diferencia…