PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULDAD DE AGRONOMÍA E INGENIERÍA FORESTAL

PROBLEMA 2

MATCHING

CURSO: EVALUACIÓN DE IMPACTO DE POLÍTICAS AMBIENTALES - AGE3802

PROFESOR: RODRIGO ARRIAGADA

DIOGO VALENTE POLÓNIA

03/11/2022

Problema 2: Matching

El objetivo de esta tarea es evaluar los efectos de la implementación de un programa de conservación de bosques en Costa Económica, un mítico país en Centroamérica. El análisis fue realizado con recurso a Stata y sigue anexado a este informe lo *.do file* creado.

1. Verificación de la asignación aleatoria a los grupos de tratamiento y control

Según la ley de los grandes números, a través de una asignación aleatoria de un dado programa, para muestras suficientemente grandes, todas las variables observables o no observables se distribuyen igualmente entre los grupos de control y tratamiento. Esto significa que no hay diferencias estadísticamente significativas entre los grupos, excepto la característica de tratamiento, es decir, que es posible inferir que las diferencias de productividad entre los grupos es causada solo por el tratamiento. Así, es importante evaluar si existen diferencias estadísticamente significativas entre las mismas variables observables de los distintos grupos, porque si no, entonces no se puede garantizar que la asignación aleatoria fué efectiva. Para esto, realizamos una prueba *t* para las medias de distribución de estas variables para los distintos grupos, con las siguientes hipótesis:

- H0: No hay diferencias entre las medias de cada variable entre los dos grupos
- H1: Hay diferencias entre las medias de cada variable entre los dos grupos

En el caso de las variables rechacen la hipótesis nula, i.e. valor p < alfa (0.05), habrá pruebas estadísticamente significativas de que la media de estas variables es diferente entre los grupos, por lo que no se podría comparar los grupos de tratamiento y control, sin controlar estas variables.

El *test t* fué hecho en Stata, con recurso a lo comando "ttest", agrupando las observaciones por valor de *consprog* y sus resultados se pueden ver en la siguiente tabla (con interpretación para un nivel de significancia de 5%):

Variable	**P-Value** (Pr(|T| > |t|))	**Interpretación**
pre_acres	0.0000	Rechaza H0
income	0.0000	Rechaza H0
popdensity	0.0000	Rechaza H0

Tabla 1 -Resultados del test t *entre grupos de control y tratamiento*

Como todas las variables rechazan la hipótesis nula (valor p < alfa), con valores p's muy bajos (0 para cuatro casas decimales), entonces hay pruebas estadísticamente significativas de

Problema 2: Matching

que la media de estas variables sea diferente entre los grupos tratados y no tratados (comunas que no participaron en el programa). Así que se puede concluir que la asignación de tratamiento no fué aleatoria. Los valores p son significativos al 0.01%, lo que permite una gran confianza estadística en el rechazo del test.

2. Verificación si la participación en el programa de conservación fue tan buena como si hubiera sido realizada en forma aleatoria condicional sobre el ingreso y la densidad poblacional

Si el programa hubiera sido asignado de manera aleatoria condicional sobre el ingreso y la densidad poblacional, los valores de la línea de base de bosques estarían distribuídos aleatoriamente entre observaciones con los mismos valores en las covariables. Por otras palabras, las covariables (*income* y *popdensity*) y la línea de base (*pre_acres*) no estarían correlacionadas.

Esto se puede verificar gráficamente (a través de gráficos de dispersión entre las variables) o corriendo una regresión entre las covariables y la línea base como variable dependiente. Lo que se espera es que no exista significancia una relación (coeficientes) entre *income*/ *popdensity* y *pre_acres*.

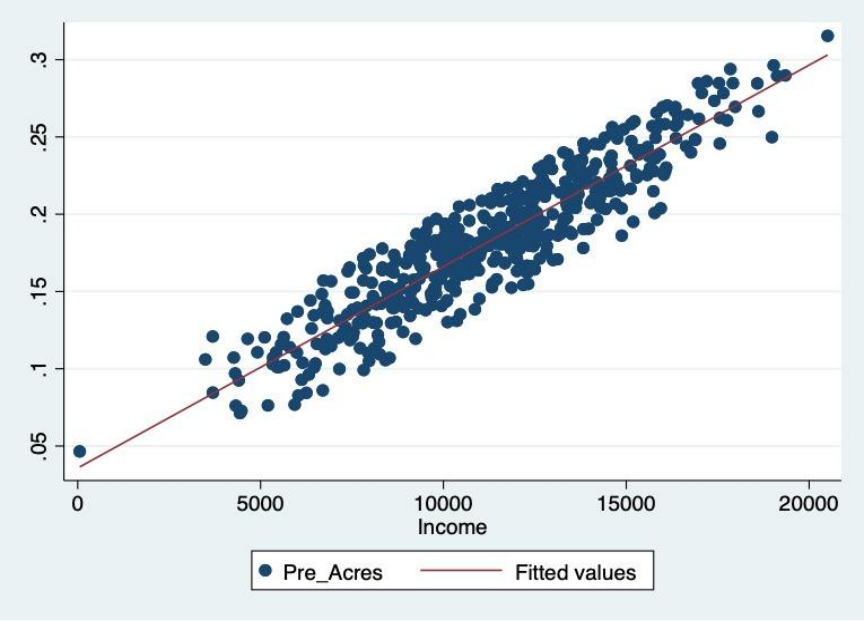

Figura 1 - Dispersión entre Income y Pre_Acres (azul), y línea de regresión lineal (rojo)

Problema 2: Matching

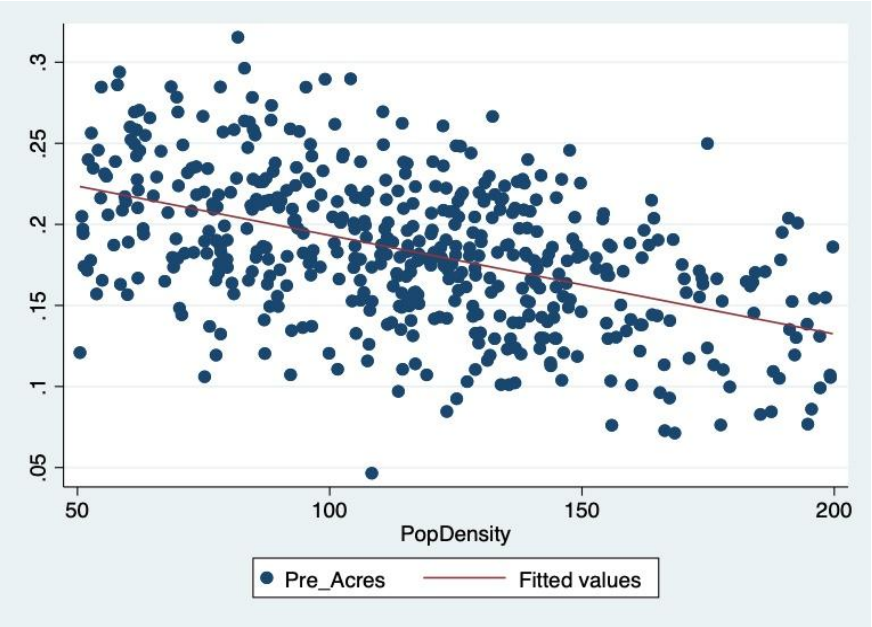

Figura 2 - Dispersión entre PopDensity y Pre_Acres (azul), y línea de regresión lineal (rojo)

De los gráficos de dispersión se desprende que existe una relación relevante entre el *income*/*popdensity* y *pre_acres*. En particular, la relación entre *income* y *pre_acres* es más fuerte (porque hay menos dispersión de los valores de *pre_acres*), cuanto mayor el ingreso más área conservada tenía la comuna antes del programa. Por outro lado, la relación entre *popdensity* y *pre_acres* es negativa, o sea, las comunas con menor densidad poblacional ya tenían más área conservada. Estas relaciones son bastante lógicas y muestran la necesidad de un buen análisis causal de los programas de conservación, que debe combatir estos problemas de endogeneidad.

Además, esto indica que la participación en el programa de conservación no fue tan buena como si hubiera sido realizada en forma aleatoria condicional sobre el ingreso y la densidad poblacional. En ese caso, no debería haber ninguna relación entre estas variables, o sea la línea de regresión debería ser horizontal y las observaciones (puntos) deberían estar dispersas por todo el gráfico de forma uniforme.

A continuación, se presentan los resultados de la regresión lineal, a través del comando "reg" en Stata, ocupando la variable *pre_acres* cómo Y (variable dependiente), y las variables *income* y *popdensity* cómo X 's (variables independientes). Como esperado, los coeficientes, aunque bajos, son muy significativos, mismo al 0.1% (como se puede ver pelos valores a rojo). Además, la regresión tuve muy buenos resultados de estimación (en amarillo), con un R^2 de 100%, o sea, toda la variación de la variables dependiente (*pre_acres*) es explicada por las variables dependientes, que es el contrario de que debería acontecer en el caso si la

Problema 2: Matching

asignación hubiera sido realizada en forma aleatoria condicional sobre el ingreso y la densidad poblacional.

Se concluye que la participación en el programa de conservación no fue tan buena como si hubiera sido realizada en forma aleatoria condicional sobre el ingreso y la densidad poblacional.

```
. reg pre_acres income popdensity
```

Source	SS	df	MS	Number of obs	=	500
				F(2, 497)	>	99999.00
Model	1.05564405	2	.527822027	Prob > F	=	0.0000
Residual	1.9805e-14	497	3.9849e-17	R-squared	=	1.0000
				Adj R-squared	=	1.0000
Total	1.05564405	499	.002115519	Root MSE	=	6.3e-09

pre_acres	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
income	.0000125	8.75e-14	1.4e+08	0.000	.0000125	.0000125
popdensity	-.0005	7.85e-12	-6.4e+07	0.000	-.0005	-.0005
_cons	.1	1.44e-09	7.0e+07	0.000	.1	.1

Figura 3 - Resultados de la regresión lineal

$(preacres = \alpha \cdot income + \beta \cdot popdensity + \varepsilon)$

3. Estimación del impacto causal con regresión lineal

Haciendo la regresión a través de lo comando *reg* en Stata, ocupando la variable *post_acres* como *outcome* (variable dependiente), *consprog* cómo *treatment* y *pre_acres*, *income*, *popdensity* como controles (factores), se obtienen los resultados de la Figura 4.

Obtenemos un impacto estimado (visto a través del coeficiente de la variable *consprog*) de 0.045641, o sea, se podría concluir que el programa de conservación aumenta la porcentaje de área conservada en 0.045641 puntos porcentuales, o sea, una comuna con 10% de conservación, después de participar teria 10.045% de area conservada. Este coeficiente, es estadísticamente significativo (valor p = 0.0%). Además, las covariables *income* y *popdensity* tienen también coeficientes significativos, validando aún más la necesidad de controlarlos. Por último, la variable *pre_acres* fue omitida por colinealidad, debido a una dependencia entre las variables independientes. De hecho, como vimos en la pregunta anterior la variable *pre_acres* depende significativamente de las covariables *income* y *popdensity*.

Problema 2: Matching

```
. reg post_acres consprog $controles
note: pre_acres omitted because of collinearity.
```

Source	SS	df	MS		
Model	.753414493	3	.251138164	Number of obs	= 500
Residual	1.21550503	496	.002450615	F(3, 496)	= 102.48
				Prob > F	= 0.0000
Total	1.96891952	499	.003945731	R-squared	= 0.3827
				Adj R-squared	= 0.3789
				Root MSE	= .0495

post_acres	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
consprog	.045641	.0055434	8.23	0.000	.0347496	.0565324
pre_acres	0	(omitted)				
income	.0000123	7.25e-07	17.03	0.000	.0000109	.0000138
popdensity	-.0003914	.0000714	-5.48	0.000	-.0005317	-.0002512
_cons	.0900096	.0112966	7.97	0.000	.0678144	.1122048

Figura 4 - Resultados de la regresión lineal entre la variable post_acres y consprog controlando por las covariables

Como vimos en clase, una relación causal consiste de tres partes: (1) la causa precede al efecto, (2) la causa fue relacionada con el efecto (correlación) y (3) no podemos encontrar una explicación alternativa plausible para el efecto que no sea la causa, o si podemos encontrar, entonces debemos controlar por estos factores. El punto 3 es, normalmente, lo más difícil de confirmar. En la inferencia causal podemos pensar en la identificabilidad (estrategia de identificación) como la condición que permite medir el efecto causal a partir de los datos observados, o sea, confirmar los 3 puntos en conjunto.

En las preguntas anteriores vimos que las variables *pre_acres*, *income*, *popdensity*, sí, influyen en *post_acres*, pudiendo ser explicaciones alternativas plausibles del efecto además del programa (*consprog*), pués la asignación no fué aleatória y es probable que las comunas con mayor propensión a conservar fueron las que participaron en el programa. Esta identificación que nos leva a controlar la regresión por estas variables, es lo que permite darle una interpretación causal al parámetro estimado (0.045641) que está asociado a la variable de participación en el programa (*consprog*), que es, además, significativo, cuando vemos el valor p de la regresión.

En la figura 5 podemos ver un diagrama de la relación causal identificada que permitiría darle a la regresión una interpretación causal, cumpliendo las 3 partes de la inferencia causal.

Problema 2: Matching

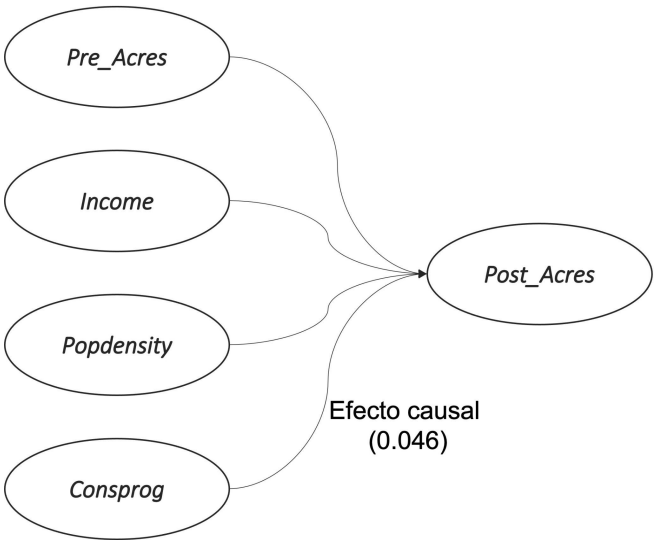

Figura 5 - Diagrama de relación causal que permitiría darle una interpretación causal a la regresión

Sin embargo, esta estrategia de identificación no es satisfactoria, ya que además de influir en el *post_acres*, las covariables también están influyendo en la propia asignación del tratamiento y tienen dependencias entre ellas. Esto lleva a la conclusión de que el efecto causal calculado, aunque significativo, no es válido.

4. Estimación del impacto a través de Matching

El Average Treatment Effect (ATE) es la media de los efectos individuales del tratamiento de la población considerada. Por otro lado, el Average Treatment Effect on the Treated (ATT) es la media de los efectos individuales del tratamiento de los tratados (por tanto, no de toda la población). En otras palabras, el ATE mide el efecto medio que tuvo el programa en las comunas tratadas y el que tendría si hubieran participado las no tratadas, mientras que el ATT sólo mide el efecto en las tratadas.

Debido a que hemos visto pruebas de las comunas tratadas son muy distintas de las no tratadas (con más propensión para la conservación), se espera que el ATT sea mayor que el ATE (potencialmente el programa tendría un efecto en las comunas no tratadas mayor que tuve en las tratadas).

El emparejamiento se realiza con 1 y 4 vecinos. También se espera que los efectos con 4 vecinos sean mayores que con 1, porque las observaciones serán más diferentes. A continuación, se presentan los resultados para las 4 análisis.

Problema 2: Matching

```
. teffects nnmatch (post_acres $controles) (consprog), ate nneighbor(1)
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        500
Estimator      : nearest-neighbor matching      Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
consprog (1 vs 0)	.0410994	.0076203	5.39	0.000	.0261639	.0560349

Figura 6 - Resultados del Matching (ATE) para el vecino más cercano

```
. teffects nnmatch (post_acres $controles) (consprog), ate nneighbor(4)
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        500
Estimator      : nearest-neighbor matching      Matches: requested =          4
Outcome model  : matching                                  min =          4
Distance metric: Mahalanobis                               max =          4
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
consprog (1 vs 0)	.0393229	.006515	6.04	0.000	.0265537	.052092

Figura 7 - Resultados del Matching (ATE) para los 4 vecinos más cercanos

Problema 2: Matching

```
. teffects nnmatch (post_acres $controles) (consprog), atet nneighbor(1)
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        500
Estimator      : nearest-neighbor matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Distance metric: Mahalanobis                                max =          1
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET						
consprog (1 vs 0)	**.0363754**	**.0098018**	**3.71**	**0.000**	**.0171642**	**.0555866**

Figura 8 - Resultados del Matching (ATT) para el vecino más cercano

```
. teffects nnmatch (post_acres $controles) (consprog), atet nneighbor(4)
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        500
Estimator      : nearest-neighbor matching      Matches: requested =          4
Outcome model  : matching                                   min =          4
Distance metric: Mahalanobis                                max =          4
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET						
consprog (1 vs 0)	**.0346162**	**.007426**	**4.66**	**0.000**	**.0200614**	**.0491709**

Figura 9 - Resultados del Matching (ATT) para los 4 vecinos más cercanos

El efecto del programa, con Matching, sigue siendo significativo, pero disminuye. Esto indica que el efecto calculado anteriormente (con regresión) está sobreestimado, probablemente debido a las grandes diferencias entre los grupos. Además, el ATT es menor que el ATE, lo que refuerza la afirmación de que las comunas tratadas ya tenían más propensión a aumentar su área de conservación. La variable *popdensity* también fué omitida, debido a la colinealidad entre *pre_acres*, *income* y *popdensity*.

Problema 2: Matching

5. Estimación del impacto a través de Matching, con caliper

Según Andam et al. (2008), el emparejamiento con calipers mejora el equilibrio de covariables. Los calipers definen un nivel de tolerancia para juzgar la calidad de las coincidencias; si una comuna tratada no tiene una coincidencia dentro del caliper (es decir, los controles disponibles no son buenas coincidencias), debe eliminarse de la muestra. Esto significa que los calipers reducen el sesgo, pero a costa de estimar la mejora de la conservación en una submuestra que puede no ser representativa de la población de comunas.

Se ocupó el siguiente comando de Stata:

teffects nnmatch (post_acres $controles) (consprog), ate caliper(0.5)

Pero, si obtiene el siguiente error: "*no nearest-neighbor matches for observation 1 within caliper 0.5; use option osample() to identify all observations with deficient matches*". O sea, el comando no elimina las observaciones sin coincidencia de forma automática. Cuando se ocupa la opción *osample()* y luego se aplica el *teffects* sin las observaciones identificadas, obtenemos el mismo error. Se intentaron 10 iteraciones de este proceso sin lograr resultados, así que no fué posible ocupar el caliper en Stata. Los comandos usados son los siguientes:

1. *teffects nnmatch (post_acres $controles) (consprog), ate caliper(0.5) osample(unmatched)*

2. *teffects nnmatch (post_acres $controles) (consprog) if unmatched != 1, ate caliper(0.5)*

Es posible ejecutar el comando con un calibre superior o igual a 1,45, pero esto no elimina ninguna observación, por lo que no aporta nueva información.

```
. teffects nnmatch (post_acres $controles) (consprog), ate caliper(1.45)
note: popdensity omitted because of collinearity.

Treatment-effects estimation                   Number of obs      =        500
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
consprog (1 vs 0)	**.0410994**	**.0076203**	**5.39**	**0.000**	**.0261639**	**.0560349**

Figura 9 - Resultados del Matching (ATE) con caliper = 1.45

Problema 2: Matching

6. Estimación del impacto a través de Propensity Score Matching

El Propensity Score Matching (PSM) es un método cuasiexperimental en el que se construye un grupo de control artificial emparejando cada unidad tratada con una unidad no tratada de características similares. En concreto, el PSM calcula la probabilidad de que una unidad se inscriba en un programa en función de las características observadas. Esta es la puntuación de propensión (*pscore*). Luego, el PSM empareja las unidades tratadas con las no tratadas basándose en la puntuación de propensión. El PSM se basa en el supuesto de que, en función de algunas características observables, las unidades no tratadas pueden compararse con las unidades tratadas, como si el tratamiento hubiera sido totalmente aleatorio. De este modo, el PSM trata de imitar la aleatorización para superar los problemas de sesgo de selección que afectan a los métodos no experimentales.

Así que la condición de identificación fundamental del método de matching (y PSM) es:

$$E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$$

O sea, en este caso, que la área conservada (en promedio) antes del programa (línea base) de las comunas tratadas con las variables observables de valores X, es la misma que la área conservada (en promedio) de las comunas no tratadas que también tengan las variables observables de valores X.

A medida que el número de características aumenta, las chances de encontrar un match se reducen (ej. si X contiene n covariables todas dicotómicas, el número posibles matches será 2^n. Así que el PSM es muy útil para alcanzar estimaciones consistentes con muchas variables.

Se empieza por estimar la probabilidad individual de estar en el grupo de tratamiento para todos los individuos en los grupos de tratamiento y control, a través de un *probit*. En la figura 10, se puede observar que una vez más, la variable *popdensity* fué omitida por la elevada colinealidad entre las covariables. Después, se creó la variable *pscore*, que contiene las probabilidades individuales de cada comuna pertenecer al grupo de tratamiento, a través del comando *predict*.

El segundo paso es garantizar que vamos a comparar observaciones similares, es decir, tenemos que restringir la muestra para las observaciones en el soporte común. No-participantes fuera de los extremos de la región de soporte común no pueden encontrar sus pares. Gráficamente, el soporte común es la región superpuesta entre las observaciones tratadas y las no tratadas, en términos de *pscore*. Se puede ver este gráfico en la figura 11, creada con el comando *kdensity*.

Problema 2: Matching

```
. probit consprog $controles, r

note: popdensity omitted because of collinearity.
Iteration 0:   log pseudolikelihood = -336.50583
Iteration 1:   log pseudolikelihood =  -236.5467
Iteration 2:   log pseudolikelihood = -236.31291
Iteration 3:   log pseudolikelihood = -236.31282
Iteration 4:   log pseudolikelihood = -236.31282

Probit regression                                 Number of obs =    500
                                                  Wald chi2(2)  = 158.04
                                                  Prob > chi2   = 0.0000
Log pseudolikelihood = -236.31282                 Pseudo R2     = 0.2977
```

consprog	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
pre_acres	-46.59549	3.886711	-11.99	0.000	-54.21331	-38.97768
income	.0004254	.0000483	8.80	0.000	.0003307	.0005202
popdensity	0	(omitted)				
_cons	3.380072	.3359286	10.06	0.000	2.721664	4.03848

Figura 10 - Resultados del probit

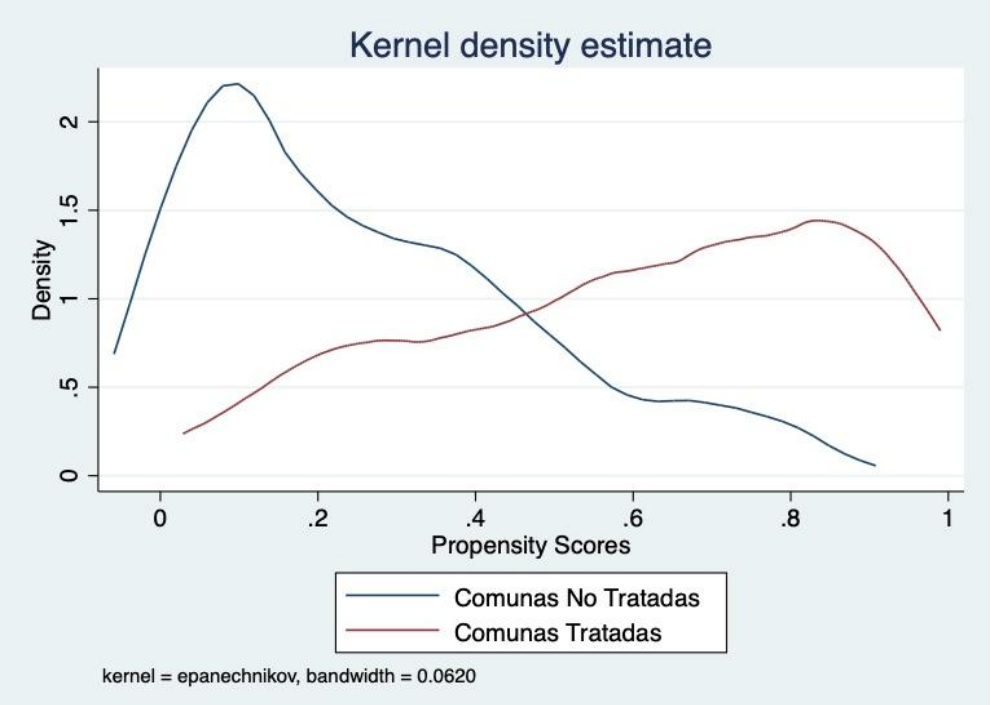

Figura 11 - Gráfico de suporte común

Problema 2: Matching

Podemos ver que la región de soporte común se sitúa entre 0 y 0.9, pero podemos especificarlo aún más seleccionando las observaciones con un *pscore* superior al *pscore* mínimo del grupo tratado (rojo), 0.288198, y un *pscore* inferior al *pscore* máximo del grupo no tratado (azul), 0.845838.

```
. // Comunas No Tratadas
. sum pscore if consprog==0
```

Variable	Obs	Mean	Std. dev.	Min	Max
pscore	300	.2620113	.2156747	.0032378	.845939

```
.
.
. // Comunas Tratadas
. sum pscore if consprog==1
```

Variable	Obs	Mean	Std. dev.	Min	Max
pscore	200	.6127921	.2620551	.0288198	.9899962

Figura 12 - Summary de la variable pscore

La variable dicotómica *sop_comun*, fue creada para contener las observaciones en el soporte común, a través del comando:

gen sop_comun=(pscore>=0.0288198 & pscore<=0.845939)

El último paso es estimar los efectos ATE y ATT, usando el comando *teffects*. Los resultados se presentan a continuación en las figuras 13 y 14.

Problema 2: Matching

```
. teffects psmatch (post_acres) (consprog $controles, probit) if sop_comun==1, ate
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        422
Estimator      : propensity-score matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Treatment model: probit                                     max =          1
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATE						
consprog (1 vs 0)	.0424644	.0069703	6.09	0.000	.0288029	.0561258

Figura 13 - Resultados del PSM (ATE)

```
. teffects psmatch (post_acres) (consprog $controles, probit) if sop_comun==1, atet
note: popdensity omitted because of collinearity.

Treatment-effects estimation                    Number of obs      =        422
Estimator      : propensity-score matching      Matches: requested =          1
Outcome model  : matching                                   min =          1
Treatment model: probit                                     max =          1
```

post_acres	Coefficient	AI robust std. err.	z	P>|z|	[95% conf.	interval]
ATET						
consprog (1 vs 0)	.0436139	.0078417	5.56	0.000	.0282445	.0589834

Figura 14 - Resultados del PSM (ATT)

Los resultados con PSM son mayores que los con Full Matching y, ahora, el ATT es superior al ATE, indicando que el programa tiene un efecto mayor en las comunas tratadas que tendría en las comunas no tratadas. Esto puede indicar que este método es lo mejor hasta ahora para estimar los efectos causales, porque es de esperar que el efectos sea más grande en comunas con más propensión a conservar (que son las que sí fueron tratadas).

Como vimos anteriormente, este método permite estimar efectos causales consistentes y no-sesgados debido a la estrategia de identificación que identifica que los grupos comparados

Problema 2: Matching

(emparejados) tienen la misma línea base de conservación y son iguales en todas las variables observables, excepto el tratamiento.

7. Representación gráfica de las relaciones causales

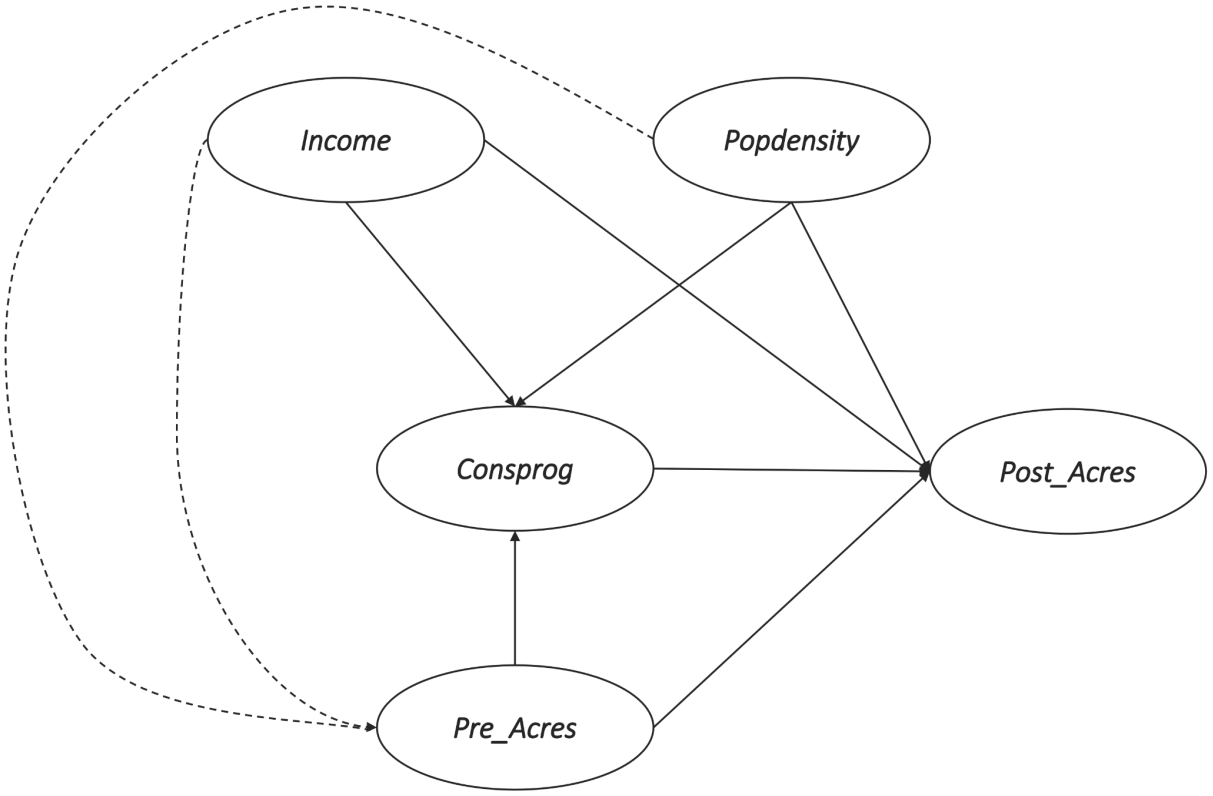

Figura 15 - Relaciones causales completas

En verdad, las covariables *pre_acres*, *income* y *popdensity* influyen en *post_acres* y en el tratamiento. Haciendo el *Matching*, lo que estamos haciendo es eliminar la influencia de las covariables en el tratamiento, para que no se confundan con *consprog*, permitiendo que los grupos de tratamiento y control siguan la relación del gráfico de la figura 5.