Maestría en Economía

Economía Aplicada

Prof. Martin A. Rossi

Tutores: Paola Llamas y Tomás Pacheco

Problem Set 11: Matching

Garcia Ojeda, Juan
Hausvirth, Martina
Hayduk, Gaspar
Salvatierra, Elias Lucas D.

Fecha de entrega: 29 de noviembre de 2024

PROBLEM SET 11: MATCHING

GARCÍA OJEDA - HAUSVIRTH - HAYDUK - SALVATIERRA

El siguiente problem set está basado en "The Impact of Improving Access to Justice on Conflict Resolution" de Yuri Soares y Micaela Sviatschi, que analiza el impacto de mejorar el acceso a la justicia sobre la resolución de conflictos.

En particular, en él se evalúa el impacto de una intervención cuyo objetivo fue mejorar la cobertura del sistema judicial en zonas alejadas de los grandes centros urbanos. Para ello se crearon distintos centros de justicia o *justice modules* en zonas remotas preparados para ofrecer los servicios de justicia más importantes. La base de datos *"base_censo.dta"* contiene información sobre distintos distritos de Perú y la variable treated indica si en ese distrito se abrió o no un centro de justicia.

1. Para analizar si los distritos tratados y no tratados son similares en función de ciertas características observables se presenta la siguiente tabla, que muestra los resultados de los tests de medias para las variables Población, Principal Vía de Acceso y Pobreza entre los municipios tratados y los municipios no tratados. La hipótesis nula del test es que las medias son iguales para ambos grupos.

Table 1. Balance Table

	Treated	**Not Treated**	**P-value**
Población	19789.32	12795.78	0.0071
Principal Vía de Acceso	0.3144876	0.2314211	0.0028
Pobreza	960.9187	900.4413	0.0749
Observations	283	1,534	

Notas: Los valores reportados son las medias para las variables entre los municipios tratados y los no tratados. El p-valor corresponde a una prueba t para la diferencia de medias entre los grupos.

A partir de la Tabla 1, se observa que en cuanto a la variable Población, la diferencia entre los grupos es estadísticamente significativa al nivel del 1% ($p<0.01$), indicando que los municipios tratados tienen una población significativamente mayor que los no tratados.

En lo que refiere a la Principal Vía de Acceso, la diferencia entre los grupos es estadísticamente significativa al nivel del 1% ($p<0.01$); los municipios tratados tienen un mayor acceso promedio a vías principales en comparación con los no tratados.

Finalmente, respecto a la Pobreza, la diferencia entre los grupos no es estadísticamente significativa al nivel del 5% (p>0.05), pero podría considerarse marginalmente significativa al nivel del 10%; esto sugiere que no hay evidencia concluyente de que los niveles de pobreza sean diferentes entre los grupos.

Como conclusión, los municipios tratados tienen una población y acceso a vías principales significativamente mayores que los no tratados y no se observan diferencias significativas en los niveles de pobreza entre los dos grupos, aunque la significancia marginal podría indicar una ligera diferencia. Las diferencias entre los municipios tratados y los municipios no tratados en Población y Principal Vía de Acceso pueden indicar que la asignación no fue aleatoria y revelan la necesidad de usar métodos no experimentales. El problema es que no hay balance completo en características observables entre los municipios tratados y los municipios no tratados.

En este sentido, Matching es un método no experimental, cuyo objetivo es identificar, dentro del grupo de control, a las unidades que seas mas similares a las del grupo de tratamiento. Este método es útil cuando el set de características observables es largo y es probable que no haya balance en todas esas características observables. Para ello, el método de Matching calcula el Propensity Score (PS) para todas las unidades, el cual indica la probabilidad de recibir el tratamiento dadas las características observables pre-tratamiento. Así, si tengo dos unidades con un PS similar pero una unidad fue tratada y la otra no, la diferencia en el outcome entre esas unidades puede atribuirse al tratamiento.

2. En este inciso se calcule el PS, que es la probabilidad de que una unidad reciba el tratamiento dadas las características observables. Para ello, la ecuación no lineal a estimar es:

$$p = F(X^{'}\beta),$$

donde p denota la probabilidad de ser tratado, X son las características observables y β los coeficientes a estimar. La función $F(.)$ añade la no-linealidad con el objetivo de predecir probabilidades entre 0 y 1, y tiene las siguientes propiedades:

- $F(-\infty) = 0$
- $F(\infty) = 1$
- $f(z) = dF(z)/dz > 0$

En un modelo probit, $F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} dt$.

Es importante destacar que en el contexto de Matching no nos interesan los coeficientes estimados $\hat{\beta}$, sino las probabilidades predichas, es decir, el Propensity Score.

3. La siguiente figura muestra la distribución del Propensity Score para los municipios tratados y los municipios no tratados. Podemos observar que la distribución para los municipios tratados está corrida a la derecha, indicando que los municipios tratados tienen mayor probabilidad de ser tratados. Por ahora, las distribuciones no están solapadas, indicando cierto sesgo de selección.

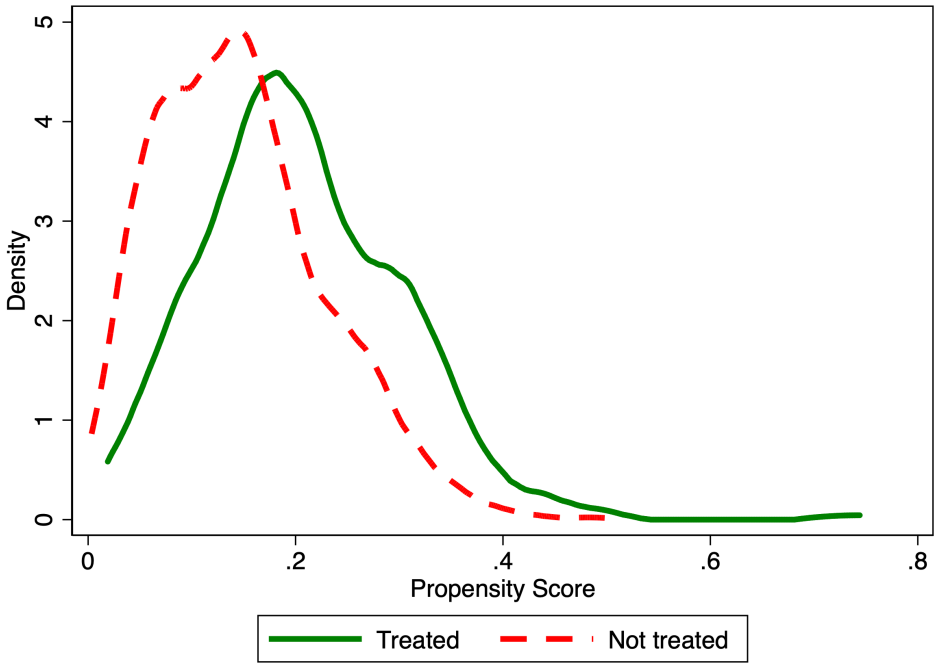

FIGURE 1. Distribucón del Propensity Score para los distritos tratados y para los no tratados

4. El soporte común se refiere a la región donde existe una superposición de las probabilidades de ser tratado entre los grupos de tratamiento y control. Garantizar el soporte común asegura que se puedan realizar comparaciones válidas entre estos grupos, ya que evita comparar unidades que son fundamentalmente diferentes.

En el do-file, se genera una variable binaria que valga 1 si la observación i está dentro del soporte común para luego eliminar las observaciones que no están dentro del él.

5. En el do-file, a partir de las lineas propuestas, se matchean distritos tratados y no tratados.

6. La siguiente figura muestra la distribución del Propensity Score entre municipios tratados y no tratados, pero considerando únicamente las observaciones utilizadas para hacer Matching. Al considerar las observaciones que recibieron matchs, estamos descartando unidades del grupo tratamiento y control que no tienen un buen contrafáctico basado en el PS.

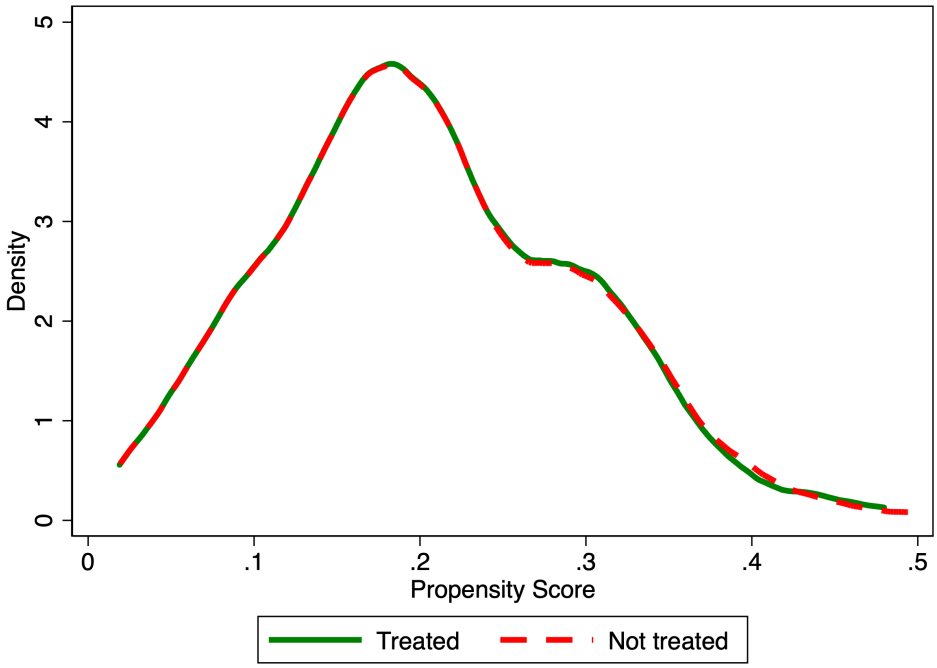

FIGURE 2. Distribucón del Propensity Score para los distritos tratados y para los no tratados considerando la submuestra con matches

Podemos observar que las distribuciones del Propensity Score para municipios tratados y municipios no tratados se superponen completamente, indicando que el Matching seleccionó unidades no tratadas que son comparables a las tratadas en términos de las características observables (ajustadas a través del PS).

7. La siguiente tabla muestra los resultados de los tests de medias para las variables Población, Principal Vía de Acceso y Pobreza entre los municipios tratados y los municipios no tratados, pero considerando las observaciones que recibieron un match. La hipótesis nula del test es que las medias son iguales para ambos grupos. Podemos observar en ella que la cantidad de observaciones es igual para ambos grupos, esto ocurre debido a que especificamos que cada unidad de control puede ser emparejada con una unidad tratada una sola vez.

TABLE 2. Balance Table

	Treated	**Not Treated**	**P-value**
Población	13354.76	16622.53	0.3513
Principal Vía de Acceso	0.2915129	0.2915129	1.0000
Pobreza	939.1033	921.7196	0.6907
Observations	271	271	

Notas: Los valores reportados son las medias para las variables entre los municipios tratados y los no tratados. El p-valor corresponde a una prueba t para la diferencia de medias entre los grupos.

De dicha tabla, apreciamos que las diferencias entre los grupos no son significativas al nivel del 5% (p>0.05), por lo que no rechazamos la hipótesis nula de medias iguales entre los municipios tratados y los municipios no tratados. Esto sugiere que el Propensity Score Matching logró balancear las características observables entre los dos grupos.

8. Si estimásemos una regresión de la forma $y_i = \beta_0 + \beta_1 Treated_i + \beta_2 X_i$ sobre la muestra matcheada para obtener el impacto causal del tratamiento, $\hat{\beta}_1$, el supuesto de identificación sería que no hay factores inobservables que afectan tanto al outcome como a la probabilidad de ser tratado. Este supuesto se conoce como Conditional Independence Assumption (CIA) y establece que, una vez que se controlan las covariables utilizadas para calcular el Propensity Score (PS), el tratamiento es independiente de los outcomes potenciales. Por lo que este supuesto implica que cualquier diferencia en el resultado promedio entre tratados y no tratados en la muestra emparejada se puede atribuir al tratamiento.

El supuesto de identificación usando Matching es el mismo supuesto necesario para hacer una regresión con controles: se asume que las covariables observadas capturan todas las fuentes de sesgo potencial. A su vez, el proceso Matching nos obliga a hacer nuevos supuestos relacionados a, por ejemplo, si usar un probit o un logit, si quedarnos con las observaciones del soporte común o no, si usar reposición o no a la hora de matchear, etc.

```
/*****************************************************************************
                            Problem Set 11: MATCHING
                            Universidad de San Andrés
                               Economía Aplicada
*****************************************************************************/
* Gaspar Hayduk; Juan Gabriel García Ojeda; Elias Lucas Salvatierra; Martina Hausvirth

/*****************************************************************************/

* 0) Set up environment
*============================================================================*
clear all
global main "/Users/gasparhayduk/Desktop/Economía Aplicada/ps11"
global output "$main/output"
global input "$main/input"
cd "$output"


*============================================================================*

* Abrimos la data:
use "$input/base_censo.dta", clear


*============================================================================*
* Inciso 1: test de medias para pobl_1999, via1, ranking_pobr
ttest pobl_1999, by(treated)
ttest via1, by(treated)
ttest ranking_pobr, by(treated)
*La hipotesis nula del test es que las medias son iguales. Los pvalores son 0.0071, 0.0028 y 0.07.
* como todos son menores a 0.10, rechazo H0.


*============================================================================*
```

```
35  *================================================================
    ============*
36  * Inciso 2: calculo del PS.
37  * Calculate propensity score
38  probit treated ind_abs_pobr ldens_pob prov_cap pob_1 pob_2 pob_3
    pob_4 km_cap_prov via3 via5 via7 via9 region_2 region_3 laltitud
    tdesnutr deficit_post deficit_aulas
39  predict p_score
40
41  * Hay missing values. los dropeamos:
42  drop if  p_score==.
43
44
45  *================================================================
    ============*
46  * Inciso 3: Distribucion del PS para tratados y no tratados:
47
48  *Falta guardar la figura. Esta es la que va
49  twoway (kdensity p_score if treated==1, lwidth(thick) lpattern(
    solid) lcolor(green)) ///
50         (kdensity p_score if treated==0, lwidth(thick) lpattern(
    "_####_####") lcolor(red)) ///
51         , scheme(s1mono) legend(lab(1 "Treated") lab(2 "Not
    treated")) ///
52        xtitle("Propensity Score") ytitle("Density")
53
54  * Exporto el grafico
55  graph export "$output/figura1.png", replace
56
57  *================================================================
    ============*
58  * Inciso 4: Generar una dummy que valga 1 si la obs esta dentro
    del common support
59  bysort treated: summ p_score
60  egen x = min(p_score) if treated==1
61  egen psmin = min(x)
62  *min PS for treated group
63  egen y = max(p_score) if treated==0
64  egen psmax=max(y)
65  *max PS for treated group
66  drop x y
67  gen common_sup=1 if (p_score>=psmin & p_score<=psmax) & p_score!=.
68  *genero una dummy que indique si cada obs esta o no dentro del
    common support
```

```
*genero una dummy que indique si cada obs esta o no dentro del common support
replace common_sup=0 if common_sup==.


* Dropeamos las obs que estan fuera del CS:
drop if common_sup==0


*================================================================*
* Corremos lineas para matchear distritos tratados con distritos no tratados:
* ssc install psmatch2
psmatch2 treated if common_sup==1, p(p_score) noreplacement
gen matches=_weight
replace matches=0 if matches==.



*================================================================*
* Inciso 6: graficar nuevamente la distribucion del PS para ambos grupos pero considerando la submuestra matches==1

preserve
keep if matches == 1


* Falta guardar la figura. Esta es la que va
twoway (kdensity p_score if treated==1, lwidth(thick) lpattern(solid) lcolor(green)) ///
       (kdensity p_score if treated==0, lwidth(thick) lpattern("_####_####") lcolor(red)) ///
       , scheme(s1mono) legend(lab(1 "Treated") lab(2 "Not treated")) ///
       xtitle("Propensity Score") ytitle("Density")

       * Exporto el grafico
graph export "$output/figura2.png", replace

restore

*================================================================*
```

```
*=================================================================
=============*
* Inciso 7: Repetir los tests de medias pero solo para la
submuestra matches==1


*Testeamos
ttest pobl_1999 if matches == 1, by(treated)
ttest via1 if matches == 1, by(treated)
ttest ranking_pobr if matches == 1, by(treated)



* Ahora sí no rechazamos H0 de medias iguales
```