Práctica Sesión 9

Mariana Lugo Mario Heredia

2022-10-22

Modelación en Ciencia de Datos

Matching

Para la elaboración de la práctica se utilizan las siguientes librerías:

```
library(tidyverse)
library(haven)
library(MatchIt)
library(knitr)
```

Para el ejercicio se utilizan los datos del experimento del mercado laboral del *National Supported Work*. El programa consistía en proveer experiencia laboral aquellos individuos que habían efrentado problemas económicos y sociales previos a su registro en el programa. Los participantes del experiemento fueron separados en un grupo de tratamiento y uno de control.

Por otro lado, se utilizan controles experiementales para obtener una estimación benchmark para el impacto del tratamiento uniendo a las unidades de tratamiento del experimento con la unidades de comparación de la *Population Survey of Income Dynamics (PSID)* y del *Current Population Survey (CPS)*.

Se cargan los datos que su utilizarán en la práctica y se contruye el set de datos utilizado en el paper agregando las variables necesarias para ajustar el modelo.

```
cps <- read_dta("https://raw.github.com/scunning1975/mixtape/master/cps_mixtape.dta")
psid <- read_dta("https://raw.github.com/scunning1975/mixtape/master/nsw_mixtape.dta")

nsw_dw_cpscontrol <- cps %>%
  bind_rows(psid) %>%
  mutate(agesq = age^2,
         agecube = age^3,
         educsq = educ^2,
         u74 = ifelse(re74 == 0,0,1),
         u75 = ifelse(re75 == 0,0,1),
         interaction1 = educ*re74
         )
```

1. Revise la descripción del programa que se realiza en la Sección I y replique la Tabla 1 de la muestra utilizada por los autores para describir a su conjunto de datos.

```r
tabla1 <- nsw_dw_cpscontrol %>%
  filter(data_id == "Dehejia-Wahba Sample") %>%
  select(-matches('sq|cube|interaction')) %>%
  group_by(treat = factor(treat, levels = c(1,0))) %>%
  summarise(across(where(is.numeric), mean),
            sample_size = n()) %>%
   pivot_longer(names_to = 'variable', values_to = 'values', cols = -treat) %>%
   pivot_wider(names_from = treat, values_from = values) %>%
  select(variable, 'Tratamiento' = 2, 'Control' = 3)
```

variable	Tratamiento	Control
age	25.82	25.05
educ	10.35	10.09
black	0.84	0.83
hisp	0.06	0.11
marr	0.19	0.15
nodegree	0.71	0.83
re74	2095.57	2107.03
re75	1532.06	1266.91
re78	6349.14	4554.80
u74	0.29	0.25
u75	0.40	0.32
sample_size	185.00	260.00

En la tabla 1 se muestra la media de todas la características de la muestra de datos utilizada para el análisis. Existen diferencias entre la tabla del paper y la tabla obtenida con estos datos únicamente en la variable `re75`: *Real earnings 12 months before training*. Vale la pena mencionar que en el paper cuentan con la variable de núm de hijos, mientras que en los datos provistos no está disponible.

En cuanto la comparación entre grupos (Treatment vs Control), las diferencias más importantes se aprecian en `no degree`: *Proportion of school dropouts*, `hisp`: *Proportion of Hispanics*, `marr`: *Married*.

2. Calcule el ATE del experimento.

Se obtiene que el efecto del NSW job-training program sobre los ingresos reales fue un incremento de $1,794.343, calculado como la diferencia de los promedios del ingreso real en 1978 (año de comparación) (renglón 1 de la tabla 2).

```r
ATE<-tabla1 %>%
  filter(variable == 're78') %>%
  summarise(ATE = Tratamiento - Control)
```

ATE
1794.34

Lo cual coincide con el renglón 1 de la tabla 2:

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]

3. Siguiendo el procedimiento de los autores, ahora utilice la información de la encuesta CPS como grupo de control no experimental. Agregue este conjunto de datos a los datos experimentales y estime el propensity score usando un modelo logit.

Los datos del CPS se agregaron desde el inicio, para evitar repetir el proceso de creación de variables. Sin embargo, es importante notar que es necesario eliminar los datos del PSID, es decir, los datos en donde `treat == 0` provenientes del dataset *Dehejia-Wahba Sample*

```
dta_mod <- nsw_dw_cpscontrol %>%
  filter(!(data_id == 'Dehejia-Wahba Sample' & treat == 0))
```

Definimos la fórmula con las variables:

$$Age, Age^2, Age^3, School, School^2, Married, Nodegree, Black, Hisp, RE74, RE75, U74, U75, School * RE74$$

Al igual que se realizó en el paper.

```
psid_cov <- dta_mod %>%
  select(age:re75, agesq:interaction1) %>%
  names()

frml <- paste0(psid_cov, collapse = '+') %>%
  paste('treat', ., sep = '~') %>%
  as.formula()

frml
```

```
## treat ~ age + educ + black + hisp + marr + nodegree + re74 +
##     re75 + agesq + agecube + educsq + u74 + u75 + interaction1
## <environment: 0x000001fa568f5a58>
```

Se define el modelo:

```
logit_cps <- glm(frml, family = binomial(link = 'logit'),
            data = dta_mod)
```

Se calcula el *propensity score* como las predicciones del modelo sobre el conjunto de datos con el que se ajustó:

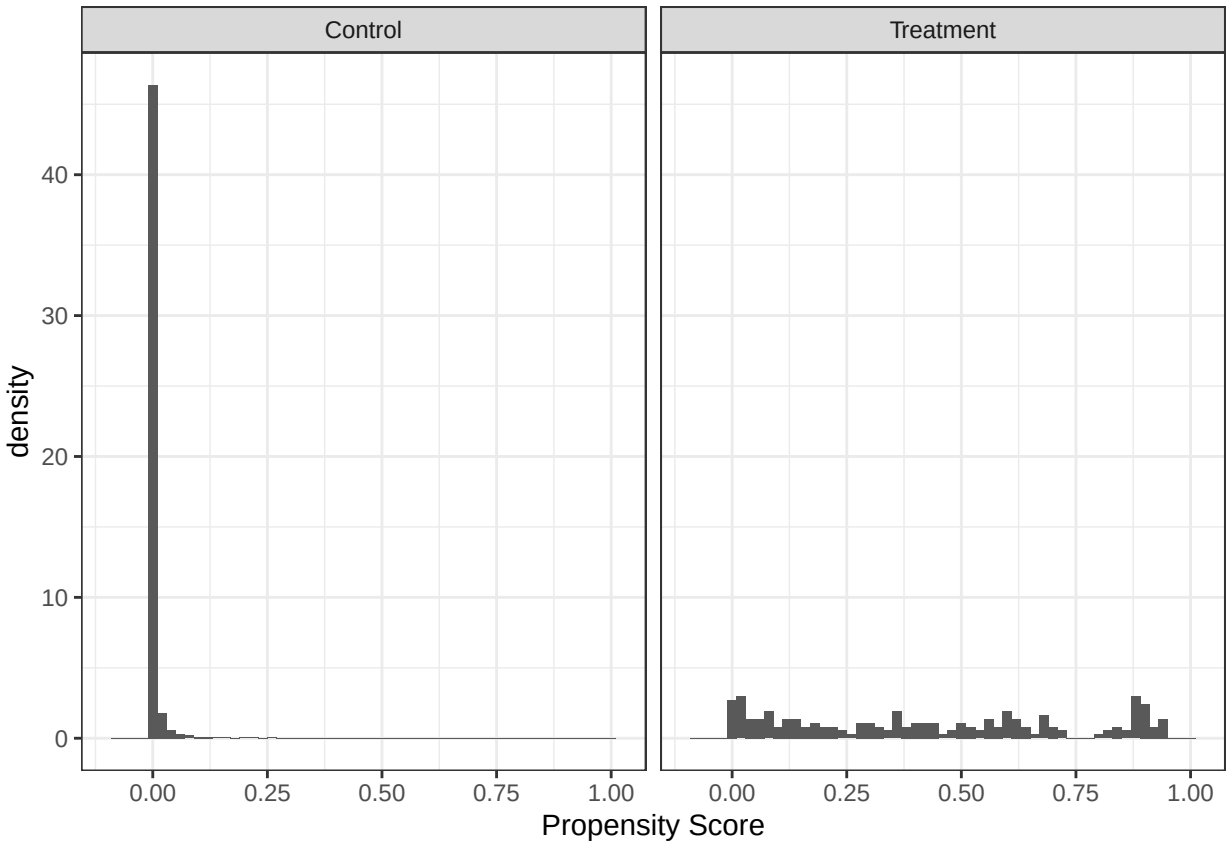
```
pr_df <- dta_mod %>%
  mutate(pr_score = logit_cps$fitted.values)
```

```
## # A tibble: 16,177 x 18
##    data_id treat   age  educ black  hisp  marr nodegree   re74   re75   re78
```

```
##    <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
##  1 CPS1        0    45    11     0     0     1        1 21517. 25244. 25565.
##  2 CPS1        0    21    14     0     0     0        0  3176.  5853. 13496.
##  3 CPS1        0    38    12     0     0     1        0 23039. 25131. 25565.
##  4 CPS1        0    48     6     0     0     1        1 24994. 25244. 25565.
##  5 CPS1        0    18     8     0     0     1        1  1669. 10728.  9861.
##  6 CPS1        0    22    11     0     0     1        1 16366. 18449. 25565.
##  7 CPS1        0    48    10     0     0     1        1 16805. 16355. 18059.
##  8 CPS1        0    18    11     0     0     0        1  1144.  3620. 15739.
##  9 CPS1        0    48     9     0     0     1        1 25862. 25244. 25565.
## 10 CPS1        0    45    12     0     0     1        0 25862.     0   3925.
## # ... with 16,167 more rows, and 7 more variables: agesq <dbl>, agecube <dbl>,
## #   educsq <dbl>, u74 <dbl>, u75 <dbl>, interaction1 <dbl>, pr_score <dbl>
```

4. Mediante la construcción del histograma para ambos grupos, analice la región del *common support*.

```
pr_df %>%
  mutate(treat = ifelse(treat==0, 'Control', 'Treatment')) %>%
  ggplot(aes(pr_score, after_stat(density))) +
  geom_histogram(binwidth = 0.02)+ #el valor del eje y cambia de acuerdo al valor de binwidth
  facet_wrap(~treat)+
  xlim(c(-0.1,1.02)) + xlab("Propensity Score")+
  theme_bw()
```



Los *propensity scores* o *p-scores* de las observaciones del grupo de control se distribuyen de manera relativamente uniforme a lo largo del intervalo $[0, .95]$. En contraste con el grupo de control, en el cual la gran mayoría de los *p-scores* se concentran en $[0, 0.01]$.

Esto quiere decir que hay muy pocas observaciones que compartan *p-scores* entre ambos grupos.

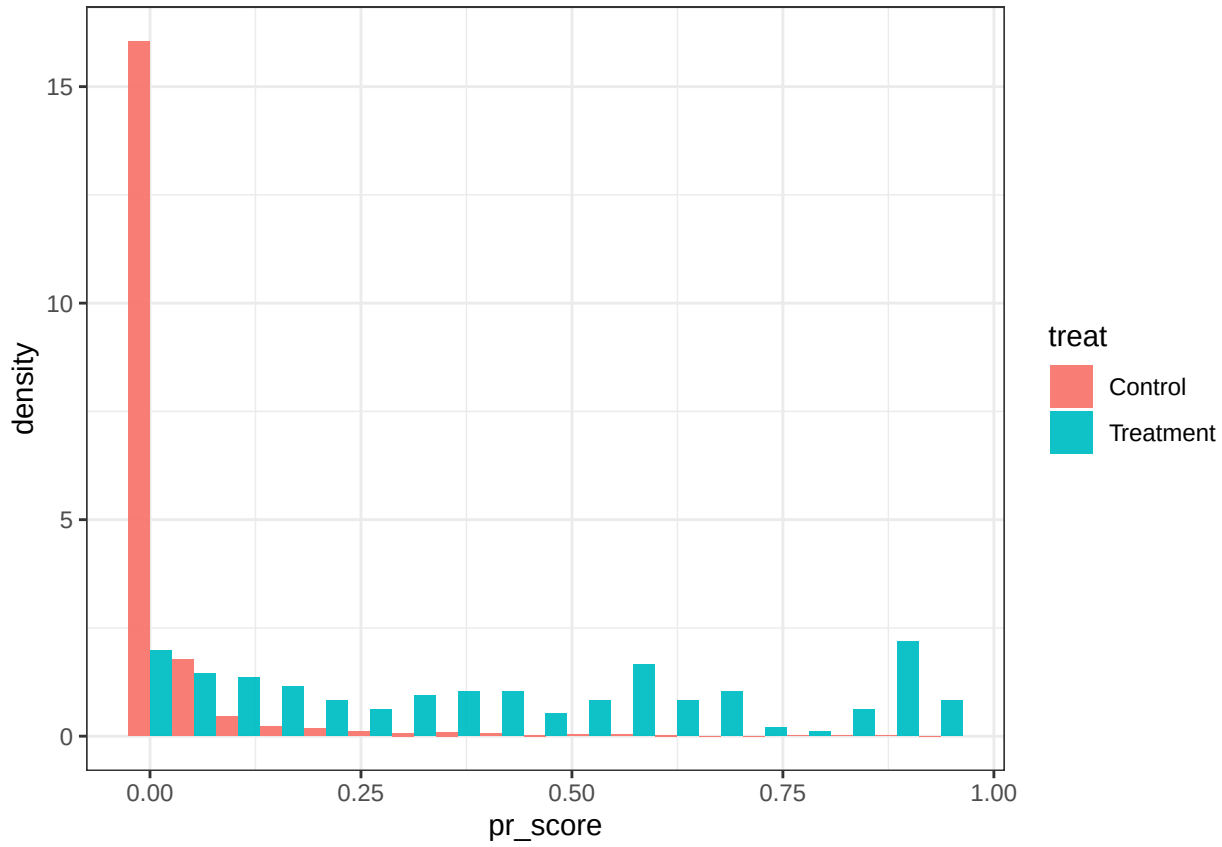
En el paper, descartan todas las observaciones del grupo de control cuyo *p-score* es menor que el valor mínimo de *p-scores* del grupo de control.

De esta manera, se excluye la gran mayoría de las observaciones que se concentraban en la primer barra del histograma:

```
(min_treated <- pr_df %>%
  filter(treat == 1) %>%
  summarise(min = min(pr_score)) %>%
  pull(min))
```

```
## [1] 0.001061388
```

```
pr_df %>%
  mutate(treat = ifelse(treat==0, 'Control', 'Treatment')) %>%
  filter(pr_score >= min_treated) %>% ##excluimos 12136 con este filtro
  ggplot(aes(pr_score, after_stat(density), fill = treat, group = treat))+
  geom_histogram(position = "dodge",bins = 19, alpha = 0.95)+
  theme_bw()
```



Incluso con el filtro aplicado, se observa que aún hay muy pocas observaciones del grupo de control que tengan un *p-score* de 0.01 o más.

5. Utilice el procedimiento de matching bajo los siguientes criterios para calcular el ATE, compare sus resultados y concluya el efecto del tratamiento en el contexto del problema.

Datos:

```
new_dta <-nsw_dw_cpscontrol %>%
  filter(!(data_id == 'Dehejia-Wahba Sample' & treat == 0))
```

a) Weighting on the propensity score Partimos de que:

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))}\right]\end{aligned}$$

La prueba de esta igualdad:

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))}\right]\end{aligned}$$

De esta manera, entonces, el estimador del ATE utilizando estimadores muestrales es:

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))}\right]\end{aligned}$$

Donde:

- $N =$ `tamaño de muestra`
- $D =$ `treat`
- $Y =$ `re78`
- $\hat{p}(X_i) =$ `propensity score`

De esta manera, calculamos el ATE con nuestra muestra:

```r
ATE_wt <- pr_df %>%
  select(treat, pr_score, re78) %>%
  mutate(ate_i = re78*(treat - pr_score)/(pr_score * (1-pr_score))) %>%
  summarise(ATE = sum(ate_i)/nrow(.))
```

ATE
-11876.79

Tenemos que -11876.79 está muy alejado del 1794 que fue reportado en el paper. Esto se debe a que, al estar utilizando el inverso de las probabilidades, cuando `p-score` es muy cercano a 0 o a 1, el valor del ponderador crece excesivamente, haciendo que la suma final se haga con valores extremos, desestabilizando el cálculo en general.

Una manera de lidiar este problema es *recortando* los valores más extremos, es decir, conservando los valores del *p-score* que se encuentren en el intervalo $[0.1, 0.9]$

De esta manera, lo calculamos como:

```r
ATE_wt <- pr_df %>%
  select(treat, pr_score, re78) %>%
  filter(between(pr_score, 0.1,0.9)) %>%
  mutate(ate_i = re78*(treat - pr_score)/(pr_score * (1-pr_score))) %>%
  summarise(ATE = sum(ate_i)/nrow(.))
```

ATE
2006.37

Lo cual nos da un valor mucho más cercano al 1794 reportado anteriormente.

b) Nearest-neighbor matching Este método esta precargado en la librería `matchit`.

```r
mod_match_nn <- matchit(frml, method = 'nearest', data = new_dta,
                        distance = 'glm', link = 'logit')

dta_matched_nn <- match.data(mod_match_nn)

N_ATE<- dta_matched_nn %>%
  group_by(treat = factor(treat, levels = c(1,0))) %>%
  summarise(across(where(is.numeric), mean),
            sample_size = n()) %>%
  select(treat, re78) %>%
  summarise(ATE = re78 -lead(re78)) %>%
  drop_na()
```

ATE
1055.04

Con el método del vecino más cercano, tenemos un ATE bastante cercano al valor reportado en el paper.

c) Coarsened exact matching Este método esta precargado en la librería `matchit`.

```
mod_match_cem <- matchit(frml, method = 'cem', data = new_dta,
                         distance = 'glm', link = 'logit', estimand = 'ATE')

dta_matched_cem <- match.data(mod_match_cem, distance = 'pr_score')

CEM_ATE<- dta_matched_cem %>%
  group_by(treat = factor(treat, levels = c(1,0))) %>%
  summarise(across(where(is.numeric), mean),
            sample_size = n()) %>%
  select(treat, re78) %>%
  summarise(ATE = re78 -lead(re78)) %>%
  drop_na()
```

ATE
2668.945

El método de Coarsened Exact Matching estima un resultado mayor al valor real.

Bibliografía

- Causal Inference: The mixtape
- Rajeev H. Dehejia, Sadek Wahba; Propensity Score-Matching Methods for Nonexperimental Causal Studies. The Review of Economics and Statistics 2002; 84 (1): 151–161. doi: https://doi.org/10.1162/003465302317331982