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.

```
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.

```
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)

names(tabla1)[2]<-"Tratamiento"
  names(tabla1)[3]<-"Control"

kable(tabla1)
```

variable	Tratamiento	Control
age	25.8162162	25.0538462
educ	10.3459459	10.0884615
black	0.8432432	0.8269231
hisp	0.0594595	0.1076923
marr	0.1891892	0.1538462
nodegree	0.7081081	0.8346154
re74	2095.5736934	2107.0266512
re75	1532.0553131	1266.9090145
re78	6349.1435021	4554.8011202
u74	0.2918919	0.2500000
u75	0.4000000	0.3153846
sample_size	185.0000000	260.0000000

En la tabla 1 se muestra la media de todas la características de la muestra de datos utilizada para el análisis. Como se establecer en el paper, se muestra que el debido al selección aleatorio para el grupo de control y de tratamiento, no existen diferencias siginificativas de las variable entre los grupos.

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. (renglón 1 de la tabla 2).

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

kable(ATE)
```

ATE
1794.342

Se compara el renglón del Full CPS de la tabla 2:

```
Full_CPS <-nsw_dw_cpscontrol %>%
  filter(data_id == "CPS1") %>%
  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)%>%
  mutate(across(where(is.numeric), ~round(., 2))) %>%
  data.frame()

names(Full_CPS)[2]<- "Full CPS"

kable(Full_CPS)
```

variable	Full CPS
age	33.23
educ	12.03
black	0.07
hisp	0.07
marr	0.71
nodegree	0.30
re74	14016.80
re75	13650.80
re78	14846.66
u74	0.88
u75	0.89
sample_size	15992.00

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.

- Se define el modelo:

```
## NSW vs CPS

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

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

```
## treat ~ age + educ + black + hisp + marr + nodegree + re74 +
##     re75 + agesq + agecube + educsq + u74 + u75 + interaction1
## <environment: 0x55d33c5b75e0>
```

```
logit_cps <- nsw_dw_cpscontrol %>%
  filter(!(data_id == 'Dehejia-Wahba Sample' & treat == 0)) %>% #quitar los PSID, i.e. comparar NSW vs
  glm(frml, family = binomial(link = 'logit'),
            data = .)
logit_cps %>% summary()
```

```
##
## Call:
## glm(formula = frml, family = binomial(link = "logit"), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2701  -0.0433  -0.0151  -0.0059   3.7009
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.346e+01  3.766e+00  -8.886  < 2e-16 ***
## age           2.425e+00  3.501e-01   6.928 4.27e-12 ***
## educ          9.248e-01  2.500e-01   3.698 0.000217 ***
## black         3.851e+00  2.663e-01  14.461  < 2e-16 ***
## hisp          1.674e+00  4.099e-01   4.084 4.43e-05 ***
## marr         -1.556e+00  2.518e-01  -6.182 6.31e-10 ***
## nodegree      9.271e-01  3.254e-01   2.849 0.004391 **
## re74         -2.203e-04  1.086e-04  -2.028 0.042520 *
## re75         -1.969e-04  3.780e-05  -5.209 1.90e-07 ***
## agesq        -6.724e-02  1.113e-02  -6.041 1.53e-09 ***
## agecube       5.685e-04  1.113e-04   5.110 3.21e-07 ***
## educsq       -5.720e-02  1.362e-02  -4.200 2.67e-05 ***
## u74          -1.750e+00  2.897e-01  -6.039 1.56e-09 ***
## u75          -9.440e-03  2.575e-01  -0.037 0.970758
## interaction1  2.222e-05  9.076e-06   2.449 0.014344 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2022.14  on 16176  degrees of freedom
## Residual deviance:  808.32  on 16162  degrees of freedom
## AIC: 838.32
##
## Number of Fisher Scoring iterations: 10
```

Se calcula el *propensity score*:

```
pr_df <- nsw_dw_cpscontrol %>%
  filter(!(data_id == 'Dehejia-Wahba Sample' & treat == 0)) %>% #quitar los PSID, i.e. comparar NSW vs
  mutate(pr_score = logit_cps$fitted.values)


ps<- pr_df %>%
  group_by(data_id,treat) %>%
  summarise(across(pr_score, mean))
```

```
## `summarise()` has grouped output by 'data_id'. You can override using the
## `.groups` argument.
```

```
kable(ps)
```

data_id	treat	pr_score
CPS1	0	0.0066476
Dehejia-Wahba Sample	1	0.4253567

Los resultados del propensity score, coinciden con los resultados del libro. Se puede observar que la mediana del propensity score es 0.4.

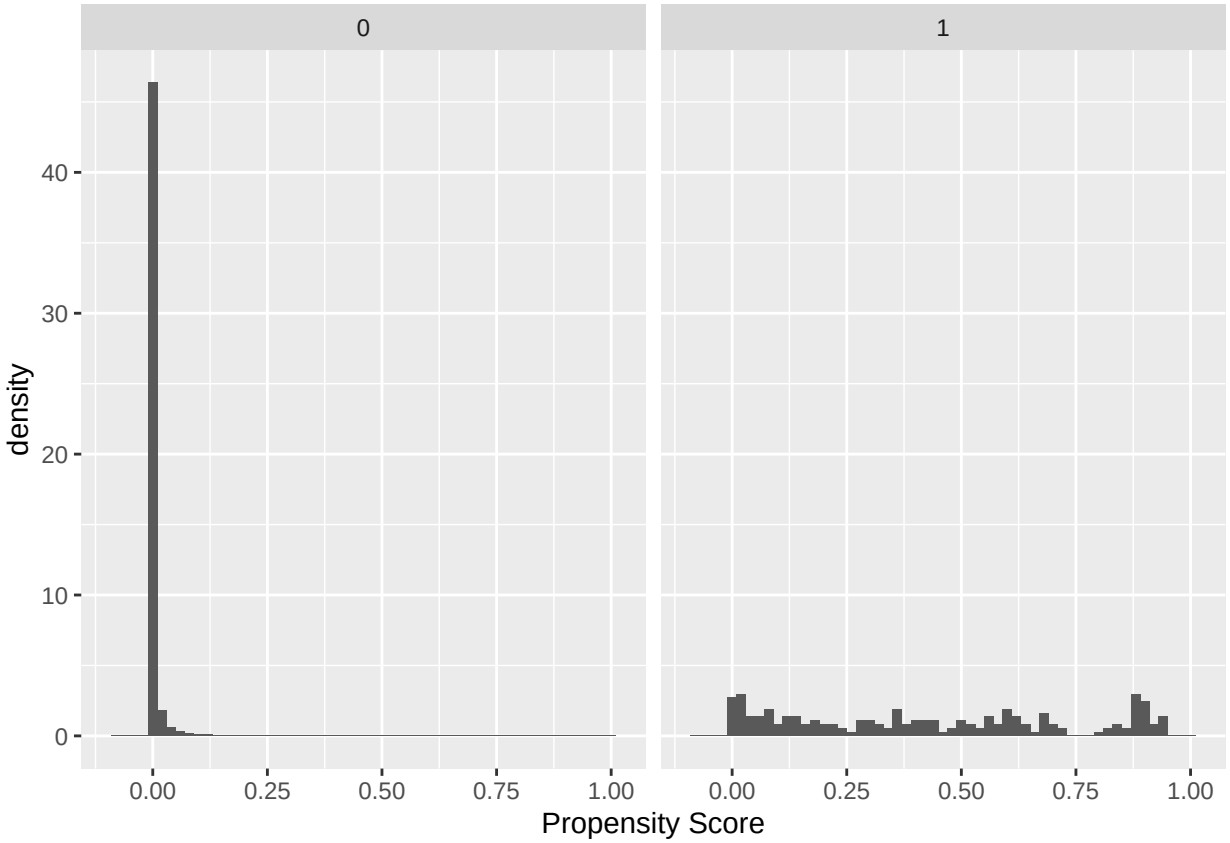
```
ps_percentil<-pr_df %>%
  group_by(data_id) %>%
  summarise(
    p01 = quantile(pr_score, 0.01),
    p05 = quantile(pr_score, 0.05),
    p10 = quantile(pr_score, 0.10),
    p25 = quantile(pr_score, 0.25),
    p50 = quantile(pr_score, 0.50),
    p75 = quantile(pr_score, 0.75),
    p90 = quantile(pr_score, 0.90),
    p95 = quantile(pr_score, 0.95),
    p99 = quantile(pr_score, 0.99)

    ) %>%
  pivot_longer(names_to = 'percentiles', values_to = 'values', cols = -data_id) %>%
  pivot_wider(names_from = data_id, values_from = values)

kable(ps_percentil)
```

percentiles	CPS1	Dehejia-Wahba Sample
p01	0.0000006	0.0017390
p05	0.0000017	0.0081935
p10	0.0000036	0.0265335
p25	0.0000193	0.1322174
p50	0.0001187	0.4001992
p75	0.0009634	0.6706164
p90	0.0066317	0.8866026
p95	0.0163109	0.9021386
p99	0.1549808	0.9372250

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

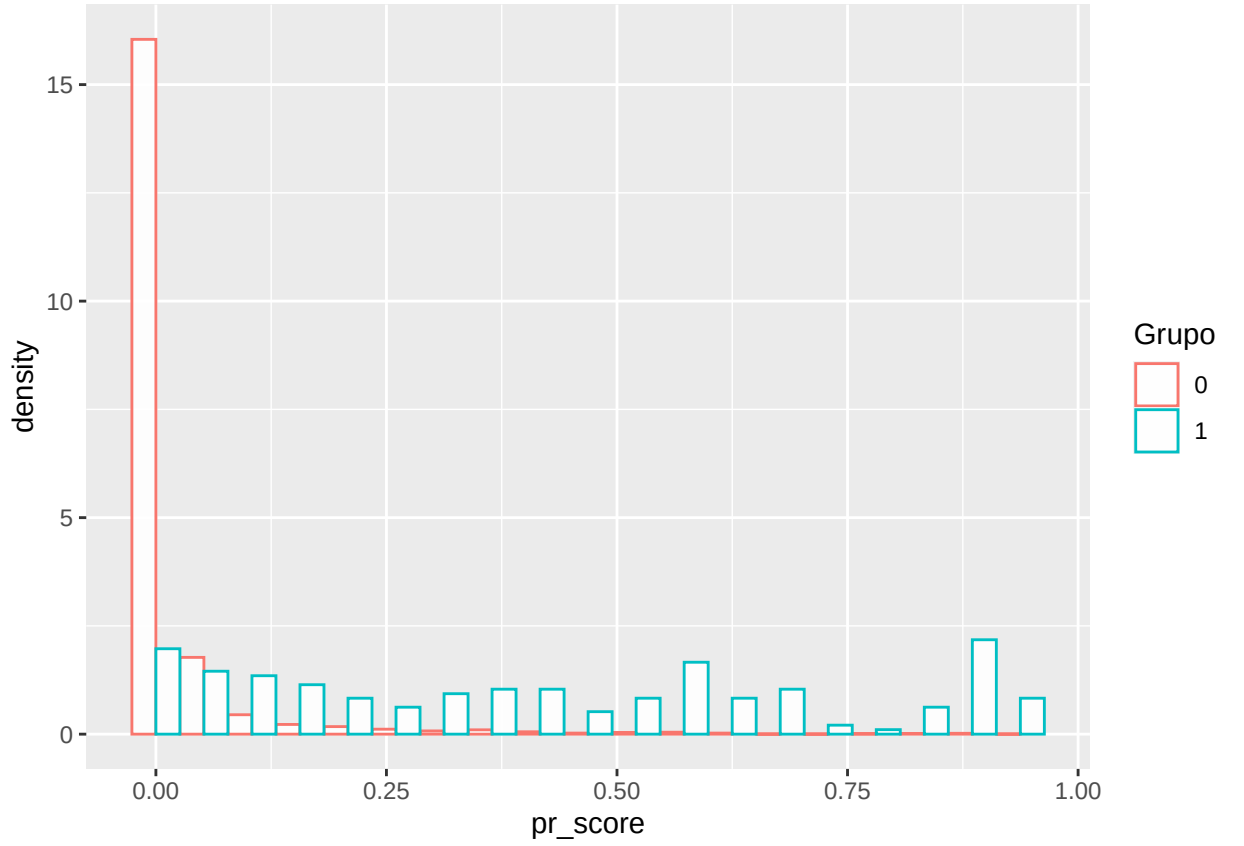
```
pr_df%>%
  ggplot(aes(pr_score, after_stat(density)), fill= factor(treat)) +
  geom_histogram(binwidth = 0.02)+ #el valor del eje y cambia de acuerdo al valor de binwidth
```

```
  facet_wrap(~treat)+
  xlim(c(-0.1,1.02)) + xlab("Propensity Score")
```



Se puedo observar que para el intervalo del propensity score existen unidades en el grupo de tratamiento con un propensity score de mayor a 0.25, mientras que no se observan unidades en ese intervalo en el grupo de control. Es decir, en general no se muestra una superposición de los propensity scores de los grupos.

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

pr_df %>%
   filter(pr_score >= min_treated) %>% ##excluimos 12136 con este filtro
  ggplot(aes(pr_score, after_stat(density), colour = factor(treat), group = factor(treat)))+
  geom_histogram(fill="white", position = "dodge",bins = 19, alpha = 0.95)+  guides(fill = guide_legend
         colour = guide_legend(title = "Grupo"))
```



En la gráfica anterior se muestra que filtrando el propepensity score mínimo del grupo de tratamiento, se observa un common support entre los grupos. Hay más sobreposición entre los grupos.

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))
```

```
N <- nrow(pr_df)

penalty_df <- pr_df %>%
  mutate(d1 = treat/pr_score, #0 cuando treat=0, grande cuando treat=1 y pr_score~0 (penaliza FN)
         d0 = (1-treat)/(1-pr_score)) #0 cuando treat=1, grande cuando treat=0 y pr_score~1 (penaliza F

# non-no}rmalized weights ------------------------------------------------

wt_df <- penalty_df %>%
  mutate(y1 = d1*re78, #d1 * re78 (penalty de FN mult por el salario resultante)
         y0 = d0*re78, #d0 * re78
```

```
         ht = y1 - y0) #ATT

wt_df %>%
  pull(ht) %>%
  mean()
```

a) Weighting on the propensity score

```
## [1] -11876.79
```

```
# Normalized weights ------------------------------------------------------

s1 <- sum(penalty_df$d1)
s0 <- sum(penalty_df$d0)

wt_norm_df <- penalty_df %>%
  mutate(y1 = (d1*re78)/(s1/N),
         y0 = (d0*re78)/(s0/N),
         norm = y1 - y0)


wt_norm_df %>%
  pull(norm) %>%
  mean()
```

```
## [1] -7238.14
```

```
# trimming propensity score -----------------------------------------------

penalty_trimmed <- pr_df %>%
  filter(between(pr_score, 0.1,0.9))%>%
  mutate(d1 = treat/pr_score, #0 cuando treat=0, grande cuando treat=1 y pr_score~0 (penaliza FN)
         d0 = (1-treat)/(1-pr_score))

# Non normalized
wt_df_trimmed <- penalty_trimmed %>%
  mutate(y1 = d1 * re78, #d1 * re78 (penalty de FN mult por el salario resultante)
         y0 = d0 *re78, #d0 * re78
         ht = y1 - y0)

wt_df_trimmed %>%
  pull(ht) %>%
  mean()
```

```
## [1] 2006.365
```

```
#Normalized

N_tr <- nrow(penalty_trimmed)
s1_tr <- sum(penalty_trimmed$d1)
s0_tr <- sum(penalty_trimmed$d0)
```

```
wt_norm_df_trimmed <- penalty_trimmed %>%
  mutate(y1 = (d1*re78)/(s1_tr/N_tr),
         y0 = (d0*re78)/(s0_tr/N_tr),
         norm = y1 - y0)


wt_norm_df_trimmed %>%
  pull(norm) %>%
  mean()
```

```
## [1] 1806.73
```

```
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()

kable(N_ATE)
```

b) Nearest-neighbor matching

ATE
1055.04

```
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()

kable(CEM_ATE)
```

c) Coarsened exact matching

ATE
2668.945

Recordemos que el efecto causal real usando los datos experimentales es de $1,794. Consideramos que el método de weighting on propensity scores, una vez que se normalizan los datos, tiene el valor más similiar el efecto real causal. Asímismo, el método de VecinosMás cercanos estima resultados similares al valor real. 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