

Matching

Manoel Galdino

2024-03-07

Introdução

Na aula de hoje, iremos aprender sobre a principal estratégia de “seleção em não-observável”, que é matching. Mas antes, vamos falar de subclassificação ,que é uma técnica mais simples e é útil para introduzir a ideia de matching.

Subclassificação

Subclassificação é um método para para cumprir os requisitos de ajustamento de “backdoor” de um DAG, por meio de estratificação e uso dos pesos de cada estrato. A ideia é que em vez de comparar a diferença na média entre o tratamento e controle em um estudo observacional, iremos comparar a diferença na média por estratos, responderando pelos pesos dos estratos. Este método consegue produzir equilíbrio (balancing) entre tratamento e controle em termos do controle observável.

O artigo clássico que é a referência no método é de Cochran (1968). O exemplo que ele usa para ilustrar o procedimento é a investigação sobre se cigarro causa câncer de pulmão. À época, ainda era um assunto bastante debatido se cigarro causava câncer. Fisher e Neyman, por exemplo, eram críticos das evidências de que cigarro causaria câncer.

```
library(knitr)
# Data for Table 5.1 (with NA for missing values)
data <- data.frame(
  Group = c("Non-smokers", "Cigarette smokers", "Cigar/pipe smokers"),
  Canada = c(20.2, 20.5, 35.5),
  UK = c(11.3, 14.1, 20.7),
  US = c(13.5, 13.5, 17.4)
)

# R Markdown code for the table

kable(data, caption = "Table 5.1: Death rates per 1,000 person-years Cochran 1968, apud Cunnigham 2022")
```

Table 1: Table 5.1: Death rates per 1,000 person-years Cochran 1968, apud Cunnigham 2022

Group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarette smokers	20.5	14.1	13.5
Cigar/pipe smokers	35.5	20.7	17.4

```
data <- data.frame(
  Group = c("Non-smokers", "Cigarette smokers", "Cigar/pipe smokers"),
  Canada = c(54.9, 50.5, 65.9),
  UK = c(79.1, 49.8, 55.7),
  US = c(57.0, 53.2, 59.7)
)

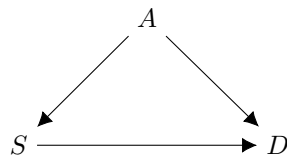
# R Markdown code for the table

kable(data, caption = "Idade média por grupo - Cochran 1968, apud Cunnigham 2022")
```

Table 2: Idade média por grupo - Cochran 1968, apud Cunnigham 2022

Group	Canada	UK	US
Non-smokers	54.9	79.1	57.0
Cigarette smokers	50.5	49.8	53.2
Cigar/pipe smokers	65.9	55.7	59.7

Tudo se passa como se tivéssemos o seguinte DAG:



Ou seja, precisamos controlar para idade (violação da CIA, ignorability, backdoor aberto etc.). Uma forma em que isso aparece é desbalanceamento das covariáveis. Ou seja, não temos mesma média de idade entre os grupos. Na verdade, balanceamento tem a ver com ter distribuição similar nas covariáveis entre diferentes níveis de tratamento.

A ideia da subclassificação é então ajustar o desbalanceamento na idade, de modo que tenha a mesma distribuição de idade (no caso, média) entre os três grupos. Digamos que tivéssemos três faixas etárias. Então, para cada faixa etária, nós usamos como peso a proporção de pessoas em cada faixa etária do grupo de controle e recalculamos a média (responderada), ajustando para idade.

Vamos fazer um exemplo (fictício) no R para entender isso.

```
library(fabricatr)
```

```
## Warning: package 'fabricatr' was built under R version 4.3.3
```

```
library(arm)
```

```
## Warning: package 'arm' was built under R version 4.3.2
```

```
## Carregando pacotes exigidos: MASS
```

```
## Carregando pacotes exigidos: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.3.3
```

```

## Carregando pacotes exigidos: lme4

## Warning: package 'lme4' was built under R version 4.3.1

##
## arm (Version 1.13-1, built: 2022-8-25)

## Working directory is C:/Users/3134777/Documents/DCP/Curso/Causalidade/Causality

library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'purrr' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.2

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.2      v readr      2.1.4
## v forcats    1.0.0      v stringr    1.5.0
## v ggplot2    3.4.4      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.0
## v purrr      1.0.2

## -- Conflicts ----- tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x tidyr::pack()    masks Matrix::pack()
## x dplyr::select() masks MASS::select()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

set.seed(123)
smoking <- fabricate(
  smoke_status = add_level(
    N = 3, status = c("não-fumante", "p-cachimbo-charuto", "s-cigarro"),
    mean_age = c(55, 60, 50),
    qtde_pessoas = c(15000, 6000, 10000)
  ),
  person_id = add_level(
    N = qtde_pessoas, age = rpois(N, mean_age),
    death = draw_binary(N = N, prob = invlogit(-4.5 + .1*(age - mean(age)) + .525*ifelse(status == "s-cigarro", 1, 0)))
  )
)
smoking <- smoking %>%
  mutate(faixa_etaria = case_when(
    age < 50 ~ "30-50",
    age < 60 ~ "51-60",
    age < 70 ~ "61-70",
    .default = "70+"
  ))
# death rate

```

```
# unique(as.factor(smoking$status))
# as.numeric(unique(as.factor(smoking$status)))
```

```
smoking %>%
  group_by(status) %>%
  summarise(death_rate = 1000*sum(death)/n())
```

```
## # A tibble: 3 x 2
##   status      death_rate
##   <chr>         <dbl>
## 1 não-fumante      15.1
## 2 p-cachimbo-charuto 23.5
## 3 s-cigarro       15.3
```

```
aux <- smoking %>%
  group_by(faixa_etaria, status) %>%
  summarise(death_per_age = sum(death),
            num_per_age = n(),
            death_age_thousands = death_per_age/num_per_age)
```

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

```
adustment_table <- aux %>%
  filter(status == "não-fumante") %>%
  ungroup() %>%
  mutate(total = sum(num_per_age)) %>%
  group_by(faixa_etaria) %>%
  summarise(num = num_per_age,
            total = total,
            prop_grupo_controle = num_per_age/total)

tabela <- aux %>%
  inner_join(adustment_table, by = join_by(faixa_etaria)) %>%
  ungroup() %>%
  group_by(status) %>%
  summarise(death_age_adjusted = 1000*sum(prop_grupo_controle*death_age_thousands),
            alternative_computation = 1000*weighted.mean(x=death_age_thousands, w=prop_grupo_controle))

kable(tabela, caption = "Subclassificação - exemplo simulado")
```

Table 3: Subclassificação - exemplo simulado

status	death_age_adjusted	alternative_computation
não-fumante	15.06667	15.06667
p-cachimbo-charuto	15.21693	15.21693
s-cigarro	22.47793	22.47793

```
data <- data.frame(
  Group = c("Non-smokers", "Cigar/pipe smokers", "Cigarette smokers"),
  Canada = c(20.2, 35.5, 19.8),
  UK = c(11.3, 14.8, 11.0),
  US = c(13.5, 21.2, 13.7)
)

# R Markdown code for the table

kable(data, caption = "Age adjusted death rates per 1,000 person-years Cochran 1968, apud Cunningham 2022")
```

Table 4: Age adjusted death rates per 1,000 person-years Cochran 1968, apud Cunningham 2022

Group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigar/pipe smokers	35.5	14.8	21.2
Cigarette smokers	19.8	11.0	13.7

Cochran, W. G. 1968. “The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies.” *Biometrics* 24 (2): 295–313.