Exercício 2 - Solução

Microeconometria - IBM0288

Prof. Raphael Gouvea)

2026-04-10

1 Resultados Potenciais e Efeitos Tratamento

1. Considere um cenário de avaliação de impacto em que temos interesse em estudar os efeitos de uma intervenção sobre uma variável de resultado.

a. Defina a variável de resultado Y_i como função da variável indicadora de tratamento $D_i \in \{0, 1\}$, utilizando a notação de resultados potenciais.

Resposta: Seja $D_i \in \{0, 1\}$ uma variável indicadora de tratamento para o indivíduo i, onde $D_i = 1$ indica que o indivíduo recebeu o tratamento e $D_i = 0$ indica que pertence ao grupo de controle.

Definimos dois **resultados potenciais**:

- $Y_i(1)$: resultado que o indivíduo i teria *caso recebesse* o tratamento.
- $Y_i(0)$: resultado que o indivíduo i teria *caso não recebesse* o tratamento.

O resultado **observado** do indivíduo i pode então ser escrito como:

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

Esta equação captura o **problema fundamental da inferência causal**: para cada indivíduo i, apenas um dos dois resultados potenciais é observado. O outro constitui o chamado *contrafactual*.

b. Formalize o efeito de tratamento individual, isto é, a diferença entre o resultado potencial sob tratamento e o resultado potencial sob controle para um indivíduo i.

O **efeito de tratamento individual** do indivíduo i é definido como a diferença entre seus dois resultados potenciais:

$$\tau_i = Y_i(1) - Y_i(0)$$

Como $Y_i(1)$ e $Y_i(0)$ nunca são simultaneamente observados para o mesmo indivíduo, τ_i não é identificável.

2. Defina e apresente a formulação matemática dos seguintes conceitos em termos de esperança condicional:

a. O efeito médio do tratamento (ATE).

Resposta:

O **Efeito Médio do Tratamento** (*Average Treatment Effect*, ATE) é definido como o valor esperado do efeito de tratamento individual na população:

$$\text{ATE} = E[Y_i(1) - Y_i(0)]$$

b. O efeito médio do tratamento sobre os tratados (ATT).

Resposta:

O **Efeito Médio do Tratamento sobre os Tratados** (*Average Treatment Effect on the Treated*, ATT) é o valor esperado do efeito de tratamento, condicional à participação no tratamento:

$$\text{ATT} = E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

c. O efeito médio do tratamento sobre os não tratados (ATU).

Resposta:

O **Efeito Médio do Tratamento sobre os Não Tratados** (*Average Treatment Effect on the Untreated*, ATU) é o valor esperado do efeito de tratamento, condicional à não participação:

$$\text{ATU} = E[Y_i(1) - Y_i(0) \mid D_i = 0]$$

3. Mostre o que ocorre ao utilizar a diferença entre as médias observadas dos grupos de tratamento ($D_i = 1$) e de controle ($D_i = 0$) como estimador do efeito do tratamento.

a. Derive $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$ e mostre que pode ser decomposto em dois termos: ATT e viés de seleção.

Resposta:

Considere o estimador da diferença de médias observadas entre os grupos de tratamento e controle:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

Substituindo $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$$

Somando e subtraindo $E[Y_i(0) \mid D_i = 1]$:

$$= \underbrace{E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]}_{\text{ATT}} + \underbrace{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]}_{\text{Viés de Seleção}}$$

Interpretação: O viés de seleção captura a diferença no resultado potencial sob controle $Y_i(0)$ entre tratados e não tratados. Ou seja, ele captura a diferença em resultados potenciais que já existia antes do tratamento! Se os indivíduos se selecionarem para o tratamento de acordo com seus resultados potenciais, a simples comparação de médias não identifica o efeito causal do tratamento.

4. A *Lei da Expectativa Total* afirma que, para uma variável de interesse X e uma partição dada por $D_i \in \{0, 1\}$, vale:

$$E[X] \;=\; \Pr(D_i = 1)\, E[X \mid D_i = 1] \;+\; \Pr(D_i = 0)\, E[X \mid D_i = 0].$$

a. Utilize este resultado para mostrar que o efeito médio do tratamento (ATE) pode ser escrito como:

$$ATE = E[Y_i(1) - Y_i(0)] = \Pr(D_i = 1)ATT + \Pr(D_i = 0)ATU.$$

Resposta:

Pela **Lei da Expectativa Total**, para qualquer variável aleatória X e $D_i \in \{0, 1\}$:

$$E[X] = \Pr(D_i = 1)\, E[X \mid D_i = 1] + \Pr(D_i = 0)\, E[X \mid D_i = 0]$$

Aplicando este resultado à variável $X = Y_i(1) - Y_i(0)$ e lembrando que ATE $= E[Y_i(1) - Y_i(0)]$, temos:

$$= \Pr(D_i = 1)\, E[Y_i(1) - Y_i(0) \mid D_i = 1] + \Pr(D_i = 0)\, E[Y_i(1) - Y_i(0) \mid D_i = 0]$$

Reconhecendo as definições de ATT e ATU:

$$\text{ATE} = \Pr(D_i = 1)\text{ATT} + \Pr(D_i = 0)\text{ATU}$$

b. Interprete a equação obtida.

Resposta: ATE é média ponderada de ATT e ATU, com pesos dados pelas proporções dos grupos de tratamento e controle na amostra.

5. Com base em (3) e (4), responda:

a. Em que condição a diferença de médias entre grupos de tratamento e controle coincide exatamente com o efeito médio do tratamento (ATE)?

Resposta: São necessárias duas condições para que a diferença de médias entre os grupos identifique o ATE:

i) Os grupos de controle e tramento tenham sido definidos por atribuição aleatória, de modo que $E[Y_i \mid D_i = 1] = E[Y_i \mid D_i = 0]$.

ii) Os efeitos tratamentos sejam homogêneos, ou seja, $ATT = ATU$. Isso implica que, quando o tratamento é heterogêneo (i.e., ATT $\neq$ ATU), o ATE não coincide com nenhum dos dois efeitos específicos.

2 Experimentos Aleatórios — Exercício E13.1 (Stock & Watson)

Antes de iniciar a atividade, vamos chamar os pacotes que serão utilizados. Nesta resolução são utilizados os seguintes pacotes do R. [1]

```
library(tidyverse) # Conjunto de pacotes e funções para realizar análise de dados no
library(readxl) # Para fazer leitura e também exportar arquivos em formato Excel.
library(gt) # Para elaboração de tabela com qualidade publicação
library(here) # Para facilitar trabalhar com caminhos de pastas
```

Conforme descrito na atividade, o primeiro passo é ler os dados no R. Vamos salvar os dados no objeto `names_data`.

```
# Carregar os dados
# Ajuste o caminho conforme a localização do arquivo na sua máquina
names_data <- read_excel(here::here("labs","Names","Names.xlsx"))
```

a. *Callback rate*

Defina a taxa de retorno (callback rate) como a fração de currículos que geram uma ligação telefônica do empregador em potencial. Qual foi a taxa de retorno de chamada para brancos? E para afro-americanos? Construa um intervalo de confiança de 95% para a diferença entre as taxas de retorno. A diferença é estatisticamente significativa? Ela é grande em termos práticos no mundo real?

Do dicionário de variáveis, temos que a variável `call_back = 1` se o candidato recebeu uma ligação de retorno do empregador em potencial e `call_back = 0` caso contrário. Além disso, a variável `black`identifica se o currículo possui um nome associado a afro-americanos.

Dado essas variáveis, o primeiro passo é construir as taxas de retorno. Como a variável `call_back` é uma variável *dummy*, sua média pode ser interpretada como proporção ou taxa. Assim, podemos calcular a média de `call_back`para cada grupo racial:

```
# Calcular taxas de retorno por grupo racial
names_data %>%
  group_by(black) %>%
  summarise(
```

[1] Como não quero reportar no arquivo pdf as mensagens que o R reproduz após a execução do comando `library`, inseri opções específicas para esse bloco de código. Para maiores detalhes, veja a página de referência sobre opções de execução.

```
    n = n(), # incluir o total de observações por grupo
    n_callbacks = sum(call_back), # incluir o total de ligações de retorno por grupo
    call_back_rate = mean(call_back)) # taxa como média da dummy
```

```
# A tibble: 2 x 4
  black     n n_callbacks call_back_rate
  <dbl> <int>       <dbl>          <dbl>
1     0  2435         235         0.0965
2     1  2435         157         0.0645
```

Para saber a significância, podemos realizar um teste t de diferença de médias:

```
# Diferença de médias e intervalo de confiança via teste t
t.test(call_back ~ black, data = names_data)
```

```
    Welch Two Sample t-test

data:  call_back by black
t = 4.1147, df = 4711.6, p-value = 3.943e-05
alternative hypothesis: true difference in means between group 0 and group 1 is not e
95 percent confidence interval:
 0.01677067 0.04729503
sample estimates:
mean in group 0 mean in group 1
     0.09650924      0.06447639
```

Interpretação:

A taxa de retorno para brancos é igual a 0,0965, o que significa que 9,7% dos currículos com nomes que soam de pessoas brancas receberam uma ligação de retorno. Para os currículos associados a afrodescentes, a taxa de retorno foi de apenas 6,5%. A diferença entre essas taxas foi de 3,2%. Como o intervalo de confiança reportado acima não inclui o zero, concluímos que essa diferença é significativa. [2]

b. Diferença por gênero

Podemos abordar a questão do diferencial racial por gênero fazendo a análise separada por grupos. No dicionário de variáveis, vemos que a variável `female = 1` indica se o currículo é identificado com um nome feminino. Assim, temos:

[2] É possível chegar a mesma conclusão analisando a estatística t ou o p-valor.

```r
# Calcular taxas de retorno por grupo racial
names_data %>%
  group_by(black, female) %>%
  summarise(
    n = n(), # incluir o total de observações por grupo
    n_callbacks = sum(call_back), # incluir o total de ligações de retorno por grupo
    call_back_rate = mean(call_back)) # taxa como média da dummy
```

```
# A tibble: 4 x 5
# Groups:   black [2]
  black female     n n_callbacks call_back_rate
  <dbl>  <dbl> <int>       <dbl>          <dbl>
1     0      0   575          51         0.0887
2     0      1  1860         184         0.0989
3     1      0   549          32         0.0583
4     1      1  1886         125         0.0663
```

Assim, os homens negros tem uma taxa de ligação de retorno de 5,8%, enquanto para mulheres negras essa taxa é de 6,6%. Ou seja, há uma diferença de 0,8 pontos percentuais.

Para saber a significância, podemos realizar um teste t de diferença de médias:

```r
# Diferença de médias e intervalo de confiança via teste t
names_data %>%
  filter(black == 1) %>%
  t.test(call_back ~ female, data = .)
```

```
    Welch Two Sample t-test

data:  call_back by female
t = -0.69284, df = 936.84, p-value = 0.4886
alternative hypothesis: true difference in means between group 0 and group 1 is not e
95 percent confidence interval:
 -0.03062227  0.01464219
sample estimates:
mean in group 0 mean in group 1
     0.05828780      0.06627784
```

Como vemos do resultado acima, o p-valor do teste t de diferenças de médias é maior do que os níveis de significância usuais (1%, 5% ou 10%) de modo que o resultado não é significativo. Ou seja, a taxa de retorno entre as pessoas negras não é distinta por genêro.

c. Qualidade dos currículos

Assim como nos casos anteriorres, podemos abordar a questão do diferencial da qualidade dos currículos fazendo a análise separada por grupos. No dicionário de variáveis, vemos que a variável `high = 1` indica se o currículo é de alta qualidade. Assim, temos:

```
# Diferença de médias e intervalo de confiança via teste t
names_data %>%
  t.test(call_back ~ high, data = .)
```

```
    Welch Two Sample t-test

data:  call_back by high
t = -1.8038, df = 4843.8, p-value = 0.07132
alternative hypothesis: true difference in means between group 0 and group 1 is not e
95 percent confidence interval:
 -0.02933557  0.00122070
sample estimates:
mean in group 0 mean in group 1
     0.07343234      0.08748978
```

A diferença de taxa de retorno entre currículos de alta (8,7%) vs baixa (7,3%) qualidade é de 1,4 pontos percentuais. Essa diferença não é significativa a 5%, mas é significativa a 10%. Os resultados abaixo mostram que não existe diferença significativa entre por tipo de currículo no caso dos brancos, enquanto a diferença de taxa de retorno por tipo de currículo entre os negros é significativa a 10%.

```
# Diferença de médias e intervalo de confiança via teste t
names_data %>%
  filter(black == 1) %>%
  t.test(call_back ~ high, data = .)
```

```
    Welch Two Sample t-test

data:  call_back by high
t = -0.51898, df = 2431.1, p-value = 0.6038
alternative hypothesis: true difference in means between group 0 and group 1 is not e
95 percent confidence interval:
 -0.02469058  0.01435647
sample estimates:
mean in group 0 mean in group 1
     0.06188119      0.06704824
```

```
# Diferença de médias e intervalo de confiança via teste t
names_data %>%
  filter(black == 0) %>%
  t.test(call_back ~ high, data = .)
```

```
    Welch Two Sample t-test

data:  call_back by high
t = -1.919, df = 2410.1, p-value = 0.05511
alternative hypothesis: true difference in means between group 0 and group 1 is not e
95 percent confidence interval:
 -0.0463976089  0.0005019727
sample estimates:
mean in group 0 mean in group 1
      0.0849835       0.1079313
```

d. Balanceamento de covariáveis pré-determinadas

Quando a atribuição aleatória é realizada, espera-se que caracterísiticas dos indivíduos determinadas antes do tratamento não sejam distintas entre os dois grupos. Como os autores aleatorizaram apenas os nomes dos currículos, podemos montar uma tabela de balanceamento de covariáveis pré-determinadas utilizando as demais informações da base de dados. Caso não existam diferenças significativas entre grupo de controle e tratamento para diversas dessas variáveis, teremos evidências que mostram que a atribuição de tratamento foi de fato aleatória.

Como mostram os dados da Tabela 1, há evidências de que a atribuição dos nomes foi, de fato, aleatória. Isto porque apenas uma das variáveis **pré-determinadas** mostrou diferença significativa, o que pode ter ocorrido apenas por acaso. Assim, não há evidência de diferenças sistemáticas entre os grupos antes do tratamento. [3]

```r
#  Variáveis de interesse
vars <- c(
  "ofjobs", "yearsexp", "honors", "volunteer", "military",
  "empholes", "workinschool", "email", "computerskills", "specialskills",
  "eoe", "manager", "supervisor",
  "secretary", "offsupport", "salesrep", "retailsales", "req",
  "expreq", "comreq", "educreq", "compreq", "orgreq",
  "manuf", "transcom", "bankreal", "trade", "busservice",
  "othservice", "missind", "chicago", "high",
  "female", "college", "call_back"
)

#  Labels legíveis (opcional - ajuste conforme necessário)
var_labels <- c(
  ofjobs        = "Número de empregos anteriores",
  yearsexp      = "Anos de experiência",
  honors        = "Menção honrosa no currículo",
  volunteer     = "Trabalho voluntário",
  military      = "Serviço militar",
  empholes      = "Lacunas de emprego",
```

[3] Para essa tabela, eu pedi auxílio do Claude. Veja o prompt inicial no anexo. Depois, revisei o código e fiz alguns ajustes para corrigir erros e detalhes de formatação.

```
  workinschool  = "Trabalhou durante os estudos",
  email         = "E-mail no currículo",
  computerskills= "Habilidades em informática",
  specialskills = "Habilidades especiais",
  eoe           = "Empregador equal opportunity",
  manager       = "Vaga de gerente",
  supervisor    = "Vaga de supervisor",
  secretary     = "Vaga de secretário(a)",
  offsupport    = "Vaga de apoio administrativo",
  salesrep      = "Vaga de representante de vendas",
  retailsales   = "Vaga de vendas no varejo",
  req           = "Requisitos listados",
  expreq        = "Requisito de experiência",
  comreq        = "Requisito de comunicação",
  educreq       = "Requisito de escolaridade",
  compreq       = "Requisito de informática",
  orgreq        = "Requisito de organização",
  manuf         = "Setor: manufatura",
  transcom      = "Setor: transporte/comunicação",
  bankreal      = "Setor: banco/imóveis",
  trade         = "Setor: comércio",
  busservice    = "Setor: serviços empresariais",
  othservice    = "Setor: outros serviços",
  missind       = "Indicador de dado ausente",
  chicago       = "Chicago (vs. Boston)",
  high          = "Qualidade alta do currículo",
  female        = "Feminino",
  college       = "Ensino superior",
  call_back     = "Recebeu retorno (callback)"
)

#  Calcular estatísticas por grupo
balance_tbl <- vars |>
  map_dfr(function(var) {
    formula <- as.formula(paste(var, "~ black"))

    grupo0 <- names_data |> filter(black == 0) |> pull(!!sym(var))
    grupo1 <- names_data |> filter(black == 1) |> pull(!!sym(var))

    media0 <- mean(grupo0, na.rm = TRUE)
    media1 <- mean(grupo1, na.rm = TRUE)
    diff   <- media1 - media0

    tt     <- t.test(formula, data = names_data)
    pval   <- tt$p.value

    tibble(
      variable = var,
```

```
      label    = var_labels[var],
      mean0    = media0,
      mean1    = media1,
      diff     = diff,
      pval     = pval
    )
  })

#  Montar tabela gt
tabela_gt <- balance_tbl |>
  gt(rowname_col = "label") |>

  # ----- Cabeçalho ------------------------------------------------------
  tab_header(
    title    = md("**Tabela de Balanceamento de Covariáveis**"),
    subtitle = md("Comparação entre currículos com nomes *negros* e *brancos*")
  ) |>

  # ----- Rótulos das colunas --------------------------------------------
  cols_label(
    label  = "Variável",
    mean0  = md("Brancos"),
    mean1  = md("Negros"),
    diff   = md("Diferença"),
    pval   = md("*p*-valor")
  ) |>

  # ----- Ocultar coluna auxiliar ----------------------------------------
  cols_hide(variable) |>

  # ----- Formato numérico -----------------------------------------------
  fmt_number(columns = c(mean0, mean1, diff), decimals = 3) |>
  fmt(
    columns = pval,
    fns = function(x) {
      case_when(
        x < 0.001 ~ "<0,001",
        TRUE      ~ formatC(x, digits = 3, format = "f")
      )
    }
  ) |>

  # ----- Destaque: p-valor significativo --------------------------------
  tab_style(
    style = list(
      cell_fill(color = "#FFF3CD"),
      cell_text(weight = "bold")
    ),
```

```
    locations = cells_body(
      columns = pval,
      rows    = pval < 0.05
    )
  ) |>

  # ----- Linha de destaque para callback -----------------------------------
  tab_style(
    style = cell_fill(color = "#D6EAF8"),
    locations = cells_body(rows = variable == "call_back")
  ) |>

  # ----- Nota de rodapé ----------------------------------------------------
  tab_footnote(
    footnote = "Diferença de médias testada via t-test bicaudal. Células em amarelo i
    locations = cells_column_labels(columns = pval)
  ) |>
  tab_footnote(
    footnote = "A variável 'black' foi atribuída aleatoriamente aos currículos.",
    locations = cells_title(groups = "title")
  ) |>

  # ----- Fonte -------------------------------------------------------------
  tab_source_note(
    source_note = md("Fonte: Bertrand & Mullainathan (2004). *Are Emily and Greg More
  ) |>

  # ----- Tema visual -------------------------------------------------------
  opt_stylize(style = 6, color = "gray") |>
  tab_options(
    heading.title.font.size    = px(18),
    heading.subtitle.font.size = px(13),
    column_labels.font.weight  = "bold",
    table.width                = pct(90),
    data_row.padding           = px(5)
  )

tabela_gt  # exibe no RStudio / Quarto / R Markdown
```

Tabela 1

Tabela de Balanceamento de Covariáveis[1]

Comparação entre currículos com nomes *negros* e *brancos*

	Brancos	Negros	Diferença	p-valor[2]
Número de empregos anteriores	3.664	3.658	-0.006	0.860
Anos de experiência	7.856	7.830	-0.027	0.854
Menção honrosa no currículo	0.054	0.051	-0.003	0.654
Trabalho voluntário	0.409	0.414	0.006	0.684
Serviço militar	0.092	0.102	0.009	0.266
Lacunas de emprego	0.450	0.446	-0.004	0.773
Trabalhou durante os estudos	0.558	0.561	0.003	0.840
E-mail no currículo	0.479	0.480	0.001	0.954
Habilidades em informática	0.809	0.832	0.024	**0.030**
Habilidades especiais	0.330	0.327	-0.003	0.831
Empregador equal opportunity	0.291	0.291	0.000	1.000
Vaga de gerente	0.152	0.152	0.000	0.968
Vaga de supervisor	0.077	0.077	0.000	1.000
Vaga de secretário(a)	0.333	0.333	0.000	0.976
Vaga de apoio administrativo	0.119	0.119	0.000	1.000
Vaga de representante de vendas	0.151	0.151	0.000	1.000
Vaga de vendas no varejo	0.168	0.168	0.000	1.000
Requisitos listados	0.787	0.787	0.000	1.000
Requisito de experiência	0.435	0.435	0.000	1.000
Requisito de comunicação	0.125	0.125	0.000	1.000
Requisito de escolaridade	0.107	0.107	0.000	1.000
Requisito de informática	0.437	0.437	0.000	0.977
Requisito de organização	0.073	0.073	0.000	1.000
Setor: manufatura	0.083	0.083	0.000	1.000
Setor: transporte/comunicação	0.030	0.030	0.000	1.000
Setor: banco/imóveis	0.085	0.085	0.000	1.000
Setor: comércio	0.214	0.214	0.000	1.000
Setor: serviços empresariais	0.268	0.268	0.000	1.000
Setor: outros serviços	0.155	0.155	0.000	1.000
Indicador de dado ausente	0.165	0.165	0.000	1.000
Chicago (vs. Boston)	0.555	0.555	0.000	1.000
Qualidade alta do currículo	0.502	0.502	0.000	1.000
Feminino	0.764	0.775	0.011	0.377
Ensino superior	0.716	0.723	0.007	0.610
Recebeu retorno (callback)	0.097	0.064	-0.032	**<0,001**

[1]A variável 'black' foi atribuída aleatoriamente aos currículos.

[2]Diferença de médias testada via t-test bicaudal. Células em amarelo indicam p < 0,05.

Fonte: Bertrand & Mullainathan (2004). *Are Emily and Greg More Employable Than Lakisha and Jamal?* AER.

3 Anexo - Prompt

Claude, eu tenho uma base de dados com as seguintes variáveis:

```
[1] "ofjobs"         "yearsexp"       "honors"         "volunteer"      "military"
 [6] "empholes"       "workinschool"   "email"          "computerskills" "specialskil
[11] "firstname"      "expminreq"      "eoe"            "manager"        "supervisor"
[16] "secretary"      "offsupport"     "salesrep"       "retailsales"    "req"
[21] "expreq"         "comreq"         "educreq"        "compreq"        "orgreq"
[26] "manuf"          "transcom"       "bankreal"       "trade"          "busservice"
[31] "othservice"     "missind"        "black"          "chicago"        "high"
[36] "female"         "college"        "call_back"
```

A variável black foi atribuída de modo aleatório. Eu quero que você produza uma tabela de balanceamento de covariáveis utilizando todas as variáveis dessa base de dados. Para isso, vamos montar uma tabela que reportará nas colunas a média de cada variável para o grupo de controle (black == 0) e grupo de controle (black ==1). Uma terceira coluna deverá reportar a diferença de médias entre os dois grupos e o p-valor dessa diferença. Quero que o código use tudo no estilo tidyverse. o pacote para gerar a tabela com qualidade de publicação deve ser o gt. Utilize o comando t.test para realizar o teste de diferença de médias.“‘