# Avaliação de Políticas Públicas*

**Lista de Exercícios**

Alberson da Silva Miranda†

31 de dezembro de 2024

*Código disponível em https://github.com/albersonmiranda/politicas_publicas.
†Discente PPGEco/UFES.

# SUMÁRIO

# Q1

O banco de dados PIAA_2017-2018.xlsx contém informações desidentificadas sobre notas dos alunos e frequência na monitoria da disciplina Matemática I no departamento de economia nos anos 2017 e 2018. A tabela a seguir lista as variáveis contidas no banco de dados.

| Variável | Descrição |
|---|---|
| ID | Identificação do aluno |
| Semestre | Semestre no qual a disciplina foi cursada |
| Nota | Nota do aluno |
| Faltas | Número de faltas na disciplina |
| Situação | Situação final do aluno na disciplina |
| Presença | Número de seções de monitoria que o aluno foi |
| Período | Número de períodos matriculado |
| Sexo | Masculino ou feminino |

## a

> Usando o pacote "TableOne" do R, faça um teste de balanceamento da amostra de acordo com uma variável que indica se o aluno foi a mais do que 4 seções de monitoria (tratamento). Faça esse teste sobre as seguintes variáveis: Faltas, Sexo Masculino, variáveis dummies para os semestres e variável dummy para alunos do primeiro período. O que você conclui sobre o balanceamento dos alunos entre tratados e não tratados?

Em relação às faltas, a amostra está desbalanceada. A média de faltas dos alunos que frequentaram mais de 4 seções de monitoria é de 3,82, enquanto a média dos que frequentaram menos de 4 seções é de 17,11 faltas. Formalmente, no teste de diferença de médias, com um p-valor menor que 0.001, podemos rejeitar a hipótese nula de que as médias são iguais (sendo a diferença apenas por uma questão de aleatoriedade amostral).

Em relação ao sexo, sob a hipótese nula de proporções iguais, com p-valor de 0.511, o teste indica que a amostra está balanceada. A proporção de homens (sexo = Masculino) é de 68,9% entre os que não frequentaram mais de 4 seções de monitoria e de 62,2% entre os que frequentaram.

Em relação ao semestre no qual a disciplina foi cursada, a amostra está desbalanceada (provavelmente, o semestre 20172 está puxando esse resultado). Para alunos do primeiro período, também está desbalanceada, com 100% dos alunos dentre os tratados no primeiro período.

```r
# pacotes
library(tableone)

# carregar dados
dados <- readxl::read_excel("lista/data/PIAA_2017-2018.xlsx") |>
 janitor::clean_names()

# feature engineering
dados <- within(dados, {
 tratamento <- ifelse(presenca > 4, 1, 0)
 semestre <- as.factor(semestre)
 periodo <- as.factor(periodo)
 sexo <- as.factor(sexo)
 primeiro_periodo <- as.factor(ifelse(periodo == 1, 1, 0))
})

# tabela
tableone <- CreateTableOne(
 vars = c("faltas", "sexo", "semestre", "primeiro_periodo"),
 strata = "tratamento",
 data = dados
)
print(tableone)
```

```
                         Stratified by tratamento
                          0             1             p      test
  n                         148            45
  faltas (mean (SD))      17.11 (16.65) 3.82 (3.95)   <0.001
  sexo = M (%)              102 (68.9)    28 ( 62.2)   0.511
  semestre (%)                                         0.037
     20171                   30 (20.3)    14 ( 31.1)
     20172                   47 (31.8)     5 ( 11.1)
     20181                   36 (24.3)    11 ( 24.4)
     20182                   35 (23.6)    15 ( 33.3)
  primeiro_periodo = 1 (%)  131 (88.5)    45 (100.0)   0.037
```

## b

> Faça uma regressão da nota dos alunos contra uma variável indicativa de que o aluno foi a mais do que 4 seções de monitoria (tratamento). Interprete o resultado. O que seria necessário para que esse resultado possa ser interpretado como um efeito causal? Essas condições são válidas para esses dados? Por quê? Em seguida, repita essa regressão adicionando ao modelo as variáveis Faltas, variável dummy para Sexo Masculino, variáveis dummies para o semestre no qual a disciplina foi cursada e variável dummy para alunos do primeiro período. O que acontece com o valor do coeficiente da presença na monitoria em relação ao modelo do item "a"? Essa estimativa seria mais próxima de um efeito causal? Por quê? Para as regressões, use os desvios padrões robustos a heterocedasticidade.

Para o primeiro modelo, a regressão é conjuntamente significativa, com $R^2$ de 13,9% e coeficientes individualmente significativos. A nota média é de 2 para os alunos que não frequentaram mais de 4 seções de monitoria e de $2 + 2.6$ para os alunos que frequentaram. O coeficiente de 2.6 para a *dummy* de tratamento indica que, em média, os alunos que frequentaram mais de 4 seções de monitoria tiveram nota 2.6 pontos maior que os que não frequentaram.

Para que o resultado possa ser interpretado como um efeito causal, é necessário que o grupo de controle seja um contrafactual válido. Ou seja, que os alunos que frequentaram mais de 4 seções de monitoria sejam comparáveis aos que não frequentaram, exceto pelo tratamento. No entanto, como verificado em (a), a amostra está desbalanceada em relação às faltas, ao semestre e ao primeiro período. Portanto, o resultado não pode ser interpretado como um efeito causal. Além disso, isso resolveria apenas as questões em relação ao observáveis, não cobrindo possíveis variáveis não observadas, incorrendo em viés de seleção.

```r
# modelos
m1 <- lm(nota ~ tratamento, data = dados)
m2 <- lm(
 nota ~ tratamento + faltas + sexo + semestre + primeiro_periodo,
 data = dados
)

# matriz de covariância corrigida
cov_m1 <- sandwich::vcovHC(m1, type = "HC")
cov_m2 <- sandwich::vcovHC(m2, type = "HC")

# erros padrões robustos
robust_se_m1 <- sqrt(diag(cov_m1))
robust_se_m2 <- sqrt(diag(cov_m2))

# sumário
```

```r
stargazer::stargazer(
 m1,
 m2,
 single.row = TRUE,
 se = list(robust_se_m1, robust_se_m2),
 title = "Comparação de modelos",
 header = FALSE
)
```

Tabela 2: Comparação de modelos

| | *Dependent variable:* | |
|---|---|---|
| | nota | |
| | (1) | (2) |
| tratamento | 2.602*** (0.491) | 1.271*** (0.408) |
| faltas | | −0.081*** (0.010) |
| sexoM | | 0.851*** (0.293) |
| semestre20172 | | −2.135*** (0.459) |
| semestre20181 | | −3.318*** (0.393) |
| semestre20182 | | −1.743*** (0.465) |
| primeiro_periodo1 | | 0.425 (0.542) |
| Constant | 2.024*** (0.220) | 4.339*** (0.664) |
| Observations | 193 | 193 |
| $R^2$ | 0.139 | 0.520 |
| Adjusted $R^2$ | 0.134 | 0.502 |
| Residual Std. Error | 2.753 (df = 191) | 2.088 (df = 185) |
| F Statistic | 30.836*** (df = 1; 191) | 28.686*** (df = 7; 185) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

Com todas as variáveis, o segundo modelo resulta num coeficiente bem menor para a variável de tratamento, de 1.27. Isso indica que, controlando para as demais variáveis, o efeito da presença na monitoria é menor. Isso melhora o modelo em relação aos observáveis, porém não resolve o problema do desbalanceamento e nem em relação aos não observáveis e viés de seleção.

**c**

> Usando o pacote "MatchIt", faça o pareamento dos tratados e controles usando os seguintes critérios: pareamento exato, vizinho mais próximo por distância de Mahalanobis e vizinho mais próximo por escore de propensão usando o modelo logit.Para cada um deles, verifique o balanceamento entre tratados e controles e verifique se o pareamento foi bem-sucedido.

O pareamento exato obteve sucesso em balancear as variáveis entre tratados e controles, com apenas 7 observações dos tratados sem correspondência.

O pareamento por distância de Mahalanobis encontrou correspondência para toda a amostra tratamento (45 observações) e obteve sucesso no balanceamento para a maior parte das variáveis, exceto para faltas, que ainda apresenta diferença entre tratados e controles (eCDF Max de 11%, acima dos 5% considerados aceitáveis).

Já o pareamento por escore de propensão usando o modelo logit falhou claramento no balanceamento, com eCDF $> 5\%$ para faltas, dois dos semestres e a distância (probabilidade predita de pertencer ao grupo das tratadas).

```r
library(MatchIt)

m_exato <- matchit(
 tratamento ~ faltas + sexo + semestre + primeiro_periodo,
 data = dados,
 method = "exact"
)

m_mahalanobis <- matchit(
 tratamento ~ faltas + sexo + semestre + primeiro_periodo,
 data = dados,
 method = "nearest",
 distance = "mahalanobis"
)

m_logit <- matchit(
 tratamento ~ faltas + sexo + semestre + primeiro_periodo,
 data = dados,
 method = "nearest",
 distance = "logit"
)

# balanceamento
summary(m_exato)
```

```
Call:
matchit(formula = tratamento ~ faltas + sexo + semestre + primeiro_periodo,
    data = dados, method = "exact")

Summary of Balance for All Data:
                 Means Treated Means Control Std. Mean Diff. Var. Ratio
faltas                  3.8222       17.1149         -3.3651     0.0563
sexoF                   0.3778        0.3108          0.1381          .
sexoM                   0.6222        0.6892         -0.1381          .
semestre20171           0.3111        0.2027          0.2342          .
semestre20172           0.1111        0.3176         -0.6569          .
semestre20181           0.2444        0.2432          0.0028          .
semestre20182           0.3333        0.2365          0.2054          .
primeiro_periodo0       0.0000        0.1149         -0.4114          .
primeiro_periodo1       1.0000        0.8851          0.4114          .
                 eCDF Mean eCDF Max
faltas              0.2293   0.4874
sexoF               0.0670   0.0670
sexoM               0.0670   0.0670
semestre20171       0.1084   0.1084
semestre20172       0.2065   0.2065
semestre20181       0.0012   0.0012
semestre20182       0.0968   0.0968
primeiro_periodo0   0.1149   0.1149
primeiro_periodo1   0.1149   0.1149

Summary of Balance for Matched Data:
                 Means Treated Means Control Std. Mean Diff. Var. Ratio
faltas                  3.6842        3.6842              -0     0.9916
sexoF                   0.3421        0.3421               0          .
sexoM                   0.6579        0.6579               0          .
semestre20171           0.3684        0.3684              -0          .
semestre20172           0.1053        0.1053              -0          .
semestre20181           0.1842        0.1842              -0          .
semestre20182           0.3421        0.3421              -0          .
primeiro_periodo0       0.0000        0.0000               0          .
primeiro_periodo1       1.0000        1.0000               0          .
                 eCDF Mean eCDF Max Std. Pair Dist.
faltas                   0        0              0
sexoF                    0        0              0
sexoM                    0        0              0
semestre20171            0        0              0
```

```
semestre20172             0        0               0
semestre20181             0        0               0
semestre20182             0        0               0
primeiro_periodo0         0        0               0
primeiro_periodo1         0        0               0

Sample Sizes:
              Control Treated
All            148.        45
Matched (ESS)   29.03      38
Matched         44.        38
Unmatched      104.         7
Discarded        0.         0
```

```
1 summary(m_mahalanobis)
```

```
Call:
matchit(formula = tratamento ~ faltas + sexo + semestre + primeiro_periodo,
    data = dados, method = "nearest", distance = "mahalanobis")

Summary of Balance for All Data:
                  Means Treated Means Control Std. Mean Diff. Var. Ratio
faltas                   3.8222       17.1149         -3.3651     0.0563
sexoF                    0.3778        0.3108          0.1381          .
sexoM                    0.6222        0.6892         -0.1381          .
semestre20171            0.3111        0.2027          0.2342          .
semestre20172            0.1111        0.3176         -0.6569          .
semestre20181            0.2444        0.2432          0.0028          .
semestre20182            0.3333        0.2365          0.2054          .
primeiro_periodo0        0.0000        0.1149         -0.4114          .
primeiro_periodo1        1.0000        0.8851          0.4114          .
                  eCDF Mean eCDF Max
faltas               0.2293   0.4874
sexoF                0.0670   0.0670
sexoM                0.0670   0.0670
semestre20171        0.1084   0.1084
semestre20172        0.2065   0.2065
semestre20181        0.0012   0.0012
semestre20182        0.0968   0.0968
primeiro_periodo0    0.1149   0.1149
primeiro_periodo1    0.1149   0.1149
```

```
Summary of Balance for Matched Data:
                  Means Treated Means Control Std. Mean Diff. Var. Ratio
faltas                   3.8222        4.3556          -0.135     0.7643
sexoF                    0.3778        0.3778           0.000          .
sexoM                    0.6222        0.6222           0.000          .
semestre20171            0.3111        0.3111           0.000          .
semestre20172            0.1111        0.1111           0.000          .
semestre20181            0.2444        0.2444           0.000          .
semestre20182            0.3333        0.3333           0.000          .
primeiro_periodo0        0.0000        0.0000           0.000          .
primeiro_periodo1        1.0000        1.0000           0.000          .
                  eCDF Mean eCDF Max Std. Pair Dist.
faltas               0.0123   0.1111          0.2925
sexoF                0.0000   0.0000          0.0000
sexoM                0.0000   0.0000          0.0000
semestre20171        0.0000   0.0000          0.0000
semestre20172        0.0000   0.0000          0.0000
semestre20181        0.0000   0.0000          0.0000
semestre20182        0.0000   0.0000          0.0000
primeiro_periodo0    0.0000   0.0000          0.0000
primeiro_periodo1    0.0000   0.0000          0.0000

Sample Sizes:
          Control Treated
All           148      45
Matched        45      45
Unmatched     103       0
Discarded       0       0
```

```
1 summary(m_logit)
```

```
Call:
matchit(formula = tratamento ~ faltas + sexo + semestre + primeiro_periodo,
    data = dados, method = "nearest", distance = "logit")

Summary of Balance for All Data:
                  Means Treated Means Control Std. Mean Diff. Var. Ratio
distance                 0.4190        0.1767          1.7540     0.4937
faltas                   3.8222       17.1149         -3.3651     0.0563
sexoF                    0.3778        0.3108          0.1381          .
```

```
sexoM                    0.6222        0.6892         -0.1381          .
semestre20171            0.3111        0.2027          0.2342          .
semestre20172            0.1111        0.3176         -0.6569          .
semestre20181            0.2444        0.2432          0.0028          .
semestre20182            0.3333        0.2365          0.2054          .
primeiro_periodo0        0.0000        0.1149         -0.4114          .
primeiro_periodo1        1.0000        0.8851          0.4114          .
                  eCDF Mean eCDF Max
distance             0.2971   0.5820
faltas               0.2293   0.4874
sexoF                0.0670   0.0670
sexoM                0.0670   0.0670
semestre20171        0.1084   0.1084
semestre20172        0.2065   0.2065
semestre20181        0.0012   0.0012
semestre20182        0.0968   0.0968
primeiro_periodo0    0.1149   0.1149
primeiro_periodo1    0.1149   0.1149

Summary of Balance for Matched Data:
                  Means Treated Means Control Std. Mean Diff. Var. Ratio
distance                 0.4190        0.4067          0.0890     1.1192
faltas                   3.8222        3.6000          0.0563     0.9369
sexoF                    0.3778        0.3778          0.0000          .
sexoM                    0.6222        0.6222          0.0000          .
semestre20171            0.3111        0.4000         -0.1920          .
semestre20172            0.1111        0.1111          0.0000          .
semestre20181            0.2444        0.1778          0.1551          .
semestre20182            0.3333        0.3111          0.0471          .
primeiro_periodo0        0.0000        0.0000          0.0000          .
primeiro_periodo1        1.0000        1.0000          0.0000          .
                  eCDF Mean eCDF Max Std. Pair Dist.
distance             0.0096   0.1111          0.1039
faltas               0.0069   0.0889          0.3713
sexoF                0.0000   0.0000          0.3556
sexoM                0.0000   0.0000          0.3556
semestre20171        0.0889   0.0889          0.3840
semestre20172        0.0000   0.0000          0.0889
semestre20181        0.0667   0.0667          0.4654
semestre20182        0.0222   0.0222          0.8014
primeiro_periodo0    0.0000   0.0000          0.0000
primeiro_periodo1    0.0000   0.0000          0.0000
```

```
Sample Sizes:
          Control Treated
All           148      45
Matched        45      45
Unmatched     103       0
Discarded       0       0
```

## d

> Para cada amostra pareada, faça uma regressão da nota dos alunos contra uma variável indicativa de que o aluno foi a mais do que 4 seções de monitoria(tratamento). Usando o pacote "Satargazer", organize os resultados das cinco regressões em uma tabela arrumada, interprete-os e compare os resultados da amostra pareada com o que você encontrou nas regressões sem pareamento. O que você conclui sobre o efeito da monitoria sobre as notas dos alunos?

Com as amostras pareadas, a regressão não é significativa (teste F) e o $R^2$ é reduzido drasticamente. O efeito do tratamento não é significativo em nenhuma das amostras pareadas. Isso significa que não há efeito causal da monitoria sobre as notas dos alunos (ou pelo menos que a amostra reduzida por conta do pareamento não é grande o suficiente para detectar o efeito, que é menos de 1 ponto em média).

```r
# modelos
m_exato_bal <- lm(
 nota ~ tratamento,
 data = match.data(m_exato),
 weights = weights
)
m_mahalanobis_bal <- lm(
 nota ~ tratamento,
 data = match.data(m_mahalanobis),
 weights = weights
)
m_logit_bal <- lm(
 nota ~ tratamento,
 data = match.data(m_logit),
 weights = weights
)

# matriz de covariância corrigida
cov_m_exato_bal <- sandwich::vcovHC(m_exato_bal, type = "HC")
cov_m_mahalanobis_bal <- sandwich::vcovHC(m_mahalanobis_bal, type = "HC")
```

```r
cov_m_logit_bal <- sandwich::vcovHC(m_logit_bal, type = "HC")

# erros padrões robustos
robust_se_m_exato_bal <- sqrt(diag(cov_m_exato_bal))
robust_se_m_mahalanobis_bal <- sqrt(diag(cov_m_mahalanobis_bal))
robust_se_m_logit_bal <- sqrt(diag(cov_m_logit_bal))

# sumário
stargazer::stargazer(
 m1,
 m2,
 m_exato_bal,
 m_mahalanobis_bal,
 m_logit_bal,
 single.row = FALSE,
 se = list(
 robust_se_m1,
 robust_se_m2,
 robust_se_m_exato_bal,
 robust_se_m_mahalanobis_bal,
 robust_se_m_logit_bal
 ),
 title = "Comparação de modelos",
 header = FALSE,
 font.size = "scriptsize",
 column.sep.width = "3pt"
)
```

Tabela 3: Comparação de modelos

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | nota | | | | |
| | (1) | (2) | (3) | (4) | (5) |
| tratamento | 2.602*** | 1.271*** | 0.874 | 0.924 | 0.818 |
| | (0.491) | (0.408) | (0.709) | (0.610) | (0.624) |
| faltas | | −0.081*** | | | |
| | | (0.010) | | | |
| sexoM | | 0.851*** | | | |
| | | (0.293) | | | |
| semestre20172 | | −2.135*** | | | |
| | | (0.459) | | | |
| semestre20181 | | −3.318*** | | | |
| | | (0.393) | | | |
| semestre20182 | | −1.743*** | | | |
| | | (0.465) | | | |
| primeiro_periodo1 | | 0.425 | | | |
| | | (0.542) | | | |
| Constant | 2.024*** | 4.339*** | 4.084*** | 3.702*** | 3.809*** |
| | (0.220) | (0.664) | (0.521) | (0.424) | (0.444) |
| Observations | 193 | 193 | 82 | 90 | 90 |
| $R^2$ | 0.139 | 0.520 | 0.022 | 0.025 | 0.019 |
| Adjusted $R^2$ | 0.134 | 0.502 | 0.010 | 0.014 | 0.008 |
| Residual Std. Error | 2.753 (df = 191) | 2.088 (df = 185) | 2.924 (df = 80) | 2.928 (df = 88) | 2.994 (df = 88) |
| F Statistic | 30.836*** (df = 1; 191) | 28.686*** (df = 7; 185) | 1.823 (df = 1; 80) | 2.243 (df = 1; 88) | 1.679 (df = 1; 88) |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

# Q2

Para essa questão, use o banco de dados `minwage.xlsx`. Ele contém informações coletadas por Card & Krueger (1994) para restaurantes de fast food nos estados de New Jersey (NJ) e Pennsylvania (PA) em duas rodadas de entrevistas: em Março e Novembro/Dezembro de 1992. Em Abril daquele ano, NJ aumentou seu salário mínimo de $4,25 para $5,05 por hora. Em um artigo bastante discutido, Card e Krueger usaram esse experimento natural para avaliar o efeito de um aumento do salário-mínimob sobre o emprego (um dos exemplos clássicos de controle de preços usados em livros texto de economia). Você vai usar esse banco de dados para replicar parte do estudo de Card e Krueger. No que segue, variáveis cujo nome que termina em "2" se referem a segunda rodada da pesquisa. As variáveis `fte` e `fte2` se referem a emprego equivalente em horário integral, ou seja, a soma do número de empregados em horário integral com metade do número de empregados que trabalham em meio expediente, excluindo gerentes, `dfte` se refere a mudança em `fte` entre a primeira e a segunda entrevista (`dfte` = `fte2` – `fte`); `dw` se refere a mudança no salário inicial dos funcionários entre a primeira e a segunda entrevistas; `state` é uma variável dummy para lojas localizadas em NJ e `sample` é uma variável dummy que assume o valor 1 se dados de salário e emprego estavam disponíveis na primeira e segunda entrevista. Na análise a seguir, você deve usar apenas as observações para as quais sample é igual a 1.

## a

Calcule o salário inicial médio (`wage_st`) separadamente para restaurantes em NJ e PA, em cada uma das rodadas de entrevistas.

## 1

Calcule a diferença nos salários médios para cada estado entre a primeira e a segunda entrevista.

```r
# carregar dados
dados <- readxl::read_excel("lista/data/minwage.xlsx")

# filtrando observações
dados <- subset(dados, sample == 1)
```

```r

# salário médio
salario_medio <- aggregate(
 cbind(wage_st, wage_st2) ~ state,
 data = dados,
 FUN = mean
)

# diferença
salario_medio <- transform(salario_medio, diff = wage_st2 - wage_st)

salario_medio
```

```
  state  wage_st wage_st2        diff
1     0 4.653636 4.618788 -0.03484848
2     1 4.612982 5.082141  0.46915807
```

**2**

Calcule a diferença entre as diferenças para NJ e PA que você calculou acima.

```r
diff_in_diff <- diff(salario_medio$diff)

diff_in_diff
```

```
[1] 0.5040066
```

**3**

> Qual a interpretação dessa estimativa de diferença em diferenças para o efeito sobre os salários? Sob que condições essa conta fornece uma estimativa válida do aumento do salário-mínimo sobre os salários nos restaurantes de fast food?

A estimativa de diferença em diferenças é de 0,5. Isso significa que, em média, o salário inicial dos funcionários dos restaurantes de fast food aumentou em 0.50 dólares por hora em NJ em relação a PA. Para que essa estimativa seja válida, é necessário que a diferença entre NJ e PA seja constante ao longo do tempo, exceto pelo aumento do salário mínimo em NJ. Isso é conhecido como a hipótese de tendências paralalelas: a diferença no resultado potencial sem tratamento $Y(0)$ para os indivíduos no grupo de tratamento $D = 1$ na mudança de $t = 0 \rightarrow t = 1$ deve ser igual à diferença no resultado potencial sem tratamento para os

indivíduos no grupo de controle $D = 0$ no mesmo intervalo de tempo. Se essa hipótese for válida, a estimativa de diferença em diferenças é um estimador consistente do efeito do aumento do salário mínimo sobre os salários.

### 4

> Interprete seu resultado.

Já interpretado acima.

## b

> Repita o mesmo exercício de (a) para a variável 'fte'. Qual o impacto do aumento do salário mínimo sobre o emprego nos restaurantes de NJ?

Referente à empregabilidade, a diferença em diferenças é de 2,30. Isso significa que, em média, o emprego nos restaurantes de fast food aumentou em 2,30 empregos equivalentes em horário integral em NJ em relação a PA. O mesmo pressuposto em relação à hipótese de tendências paralelas se aplica.

```r
# emprego médio
emprego_medio <- aggregate(
 cbind(fte, fte2) ~ state,
 data = dados,
 FUN = mean
)

# diferença
emprego_medio <- transform(emprego_medio, diff = fte2 - fte)

# diferença em diferenças
diff_in_diff_emprego <- diff(emprego_medio$diff)

emprego_medio
```

```
  state      fte     fte2       diff
1     0 20.11364 18.09848 -2.0151515
2     1 17.27544 17.56228  0.2868421
```

```
diff_in_diff_emprego
```

```
[1] 2.301994
```

## c

A metodologia de diferenças em diferenças (DD) também pode ser implementada por meio da seguinte regressão:

$$Y_{ist} = \alpha + \beta_1 \text{TREAT}_{is} + \gamma POST_t + \delta_{DD} \text{TREAT}_{is} \times \text{POST}_t + \varepsilon_{ist}$$

onde $Y_{ist}$ representa emprego no restaurante $i$, no estado $s$ e período $t$, $\text{TREAT}_i s$ é um indicador para a área de tratamento (NJ ou restaurantes de baixo salário em NJ), $\text{POST}_t$ é um indicador do período de tratamento (Novembro/Dezembro) e $\text{TREAT}_{is} \times \text{POST}_t$ é a interação entre essas duas *dummies*. Note que a regressão usa dados para restaurantes individuais, ao invés de dados para o estado, como em (a) e (b).

### 1

> Estime a regressão acima para salários e emprego. Como as estimativas diferem dos resultados que você encontrou em (a) e (b)?

Primeiramente, deve-se manipular o *dataset*, uma vez que os estados (salário e emprego) no tempo 1 e 2 estão sendo tratados em colunas ao invés de linhas. Com cada linha representando cada estado em cada tempo, é possível estimar a regressão proposta. Com isso, foram criadas uma linha adicional para cada observação, coopiando os dados fixos e alterando as informações que devem ser atualizadas no tempo 2. Além disso, foi criada uma variável `post` para indicar qual o tempo da observação. Também alterei os valores de `state` para facilitar a interpretação.

```
# pacotes
suppressPackageStartupMessages({
 library(tidyr)
 library(dplyr)
})

# alongando a tabela para tratar a mudança no tempo como
# linha ao invés de coluna
colunas_repetidas <- setdiff(
```

```
 names(dados),
 grep("2$", names(dados), value = TRUE)
)

dados_long <- pivot_longer(
 data = dados,
 cols = matches(".*2$|.*[^2]$"),
 names_to = c(".value", "post"),
 names_pattern = "(.*?)(2?)$"
) ▷
 mutate(
 post = factor(ifelse(post == "2", "nov_dez", "mar"), levels = c("mar", "nov_dez")),
 state = factor(ifelse(state == "1", "NJ", "PA"), levels = c("PA", "NJ"))
 ) ▷
 fill(all_of(colunas_repetidas), .direction = "down")

dados_long
```

```
# A tibble: 702 x 13
   post   sheet chain co_owned state empft emppt wage_st   fte  dfte   gap    dw
   <fct>  <dbl> <dbl>    <dbl> <fct> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
 1 mar       56     4        1 PA       20    20    5     30   -12       0  0.25
 2 nov_d~    56     4        1 PA        0    36    5.25  18   -12       0  0.25
 3 mar       61     4        1 PA        6    26    5.5   19    10.5     0 -0.75
 4 nov_d~    61     4        1 PA       28     3    4.75  29.5  10.5     0 -0.75
 5 mar      445     1        0 PA       50    35    5     67.5 -43.5     0 -0.25
 6 nov_d~   445     1        0 PA       15    18    4.75  24   -43.5     0 -0.25
 7 mar      451     1        0 PA       10    17    5     18.5  12       0  0
 8 nov_d~   451     1        0 PA       26     9    5     30.5  12       0  0
 9 mar      455     2        1 PA        2     8    5.25   6     3       0 -0.25
10 nov_d~   455     2        1 PA        3    12    5      9     3       0 -0.25
# i 692 more rows
# i 1 more variable: sample <dbl>
```

Agora, é possível estimar a regressão proposta. Na primeira, em relação aos salários iniciais, o único coeficiente significativo foi o da interação `state:post`. Para interpretá-la, deve-se ter atenção aos valores de linha de base. Como a linha de base é PA e março, `state:post` representa o efeito quando o estado é NJ e o tempo é novembro/dezembro simultaneamente. Isso quer dizer que a diferença entre os estados em março não é significativa, assim como a diferença entre os tempos em PA. No entanto, a diferença entre os estados em novembro/dezembro é significativa, indicando que o salário em NJ aumentou em relação a PA. O coeficiente de 0.504

é consistente (na verdade, exato!) com o resultado de diferença em diferenças encontrado em (a).

Para o emprego, o coeficiente da interação `state:post` também é consistente com o resultado de diferença em diferenças encontrado em (b), porém não foi significativo, indicando que não podemos apartar essa diferença do puro acaso amostral. Essa é uma vantagem de se usar a regressão, pois ela permite testar a significância estatística do efeito.

```
# regressão
m_salario <- lm(wage_st ~ state * post, data = dados_long)
m_emprego <- lm(fte ~ state * post, data = dados_long)

# sumário
stargazer::stargazer(
 m_salario,
 m_emprego,
 title = "Regressão de diferenças em diferenças",
 header = FALSE
)
```

### 2

A implementação por regressão permite que você inclua controles adicionais. Estime as regressões para salários e emprego incluindo uma variável idicativa de que a loja é própria ou franquia (co_owned) e dummies para as cadeias(chain) de restaurantes.

```
# regressão
m_salario_2 <- lm(wage_st ~ state * post + co_owned + chain, data = dados_long)
m_emprego_2 <- lm(fte ~ state * post + co_owned + chain, data = dados_long)
```

### 3

Coloque os resultados que você obteve em 1 e 2 lado a lado na mesma tabela. Faça uma tabela para salário e outra para emprego.

```
# sumário
stargazer::stargazer(
 m_salario,
 m_salario_2,
 title = "Regressão de diferenças em diferenças",
 header = FALSE
)
```

Tabela 1: Regressão de diferenças em diferenças

| | *Dependent variable:* | |
|---|---|---|
| | wage_st | fte |
| | (1) | (2) |
| stateNJ | −0.041 | −2.838** |
| | (0.038) | (1.225) |
| postnov_dez | −0.035 | −2.015 |
| | (0.048) | (1.561) |
| stateNJ:postnov_dez | 0.504*** | 2.302 |
| | (0.054) | (1.732) |
| Constant | 4.654*** | 20.114*** |
| | (0.034) | (1.104) |
| Observations | 702 | 702 |
| $R^2$ | 0.403 | 0.008 |
| Adjusted $R^2$ | 0.400 | 0.004 |
| Residual Std. Error (df = 698) | 0.277 | 8.966 |
| F Statistic (df = 3; 698) | 156.907*** | 1.869 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

Tabela 2: Regressão de diferenças em diferenças

| | *Dependent variable:* | |
|---|---|---|
| | wage_st | |
| | (1) | (2) |
| stateNJ | −0.041 | −0.039 |
| | (0.038) | (0.037) |
| postnov_dez | −0.035 | −0.035 |
| | (0.048) | (0.047) |
| co_owned | | 0.064*** |
| | | (0.022) |
| chain | | 0.034*** |
| | | (0.010) |
| stateNJ:postnov_dez | 0.504*** | 0.504*** |
| | (0.054) | (0.053) |
| Constant | 4.654*** | 4.559*** |
| | (0.034) | (0.039) |
| Observations | 702 | 702 |
| $R^2$ | 0.403 | 0.425 |
| Adjusted $R^2$ | 0.400 | 0.421 |
| Residual Std. Error | 0.277 (df = 698) | 0.272 (df = 696) |
| F Statistic | 156.907*** (df = 3; 698) | 102.945*** (df = 5; 696) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

```r
stargazer::stargazer(
 m_emprego,
 m_emprego_2,
 title = "Regressão de diferenças em diferenças",
 header = FALSE
)
```

Tabela 3: Regressão de diferenças em diferenças

| | *Dependent variable:* | |
|---|---|---|
| | fte | |
| | (1) | (2) |
| stateNJ | −2.838** | −2.871** |
| | (1.225) | (1.216) |
| postnov_dez | −2.015 | −2.015 |
| | (1.561) | (1.550) |
| co_owned | | −2.243*** |
| | | (0.728) |
| chain | | −0.243 |
| | | (0.322) |
| stateNJ:postnov_dez | 2.302 | 2.302 |
| | (1.732) | (1.720) |
| Constant | 20.114*** | 21.440*** |
| | (1.104) | (1.280) |
| Observations | 702 | 702 |
| $R^2$ | 0.008 | 0.025 |
| Adjusted $R^2$ | 0.004 | 0.018 |
| Residual Std. Error | 8.966 (df = 698) | 8.902 (df = 696) |
| F Statistic | 1.869 (df = 3; 698) | 3.552*** (df = 5; 696) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

### 4

> Os seus resultados mudaram quando incluiu as dummies para restaurantes? Essa mudança era esperada? Explique por quê.

Os resultados não mudaram (o coeficiente de interação). O que mudou foi a qualidade do ajuste. Isso pode ser importante na hora de detectar um efeito, dado que o desvio-padrão dos coeficientes é menor e pode ser o suficiente para detectar um efeito que não seria detectado sem os controles adicionais.

## d

Uma alternativa a comparação entre restaurantes em NJ e PA seria comparar restaurantes em NJ que pagam salários altos vs. restaurantes que pagam salários mais baixos antes do aumento do salário-mínimo. Restrinja sua amostra para os restaurantes de NJ apenas.

```
# filtrando observações
dados_nj <- subset(dados_long, state == "NJ")
```

### 1

> Você esperaria que as suposições para a metodologia DD sejam mais fáceis de serem defendidas na comparação de restaurantes em NJ do que na comparação de restaurantes em NJ vs. restaurantes em PA?

Não, pois haveria apenas o grupo de tratamento (NJ) sem um grupo de controle que servisse como um contrafactual válido para a comparação. A hipótese de tendências paralelas não poderia ser testada, pois não haveria um grupo de controle para comparar a evolução do salário e do emprego. Aliás, tendo apenas um ponto no tempo antes da intervenção, a hipótese de tendências paralelas já não pode ser testada!

### 2

> Construa uma variável que indique os restaurantes que pagam menos que $5 antes do aumento do salário-mínimo. Use uma regressão para calcular a estimativa DD do efeito do aumento do salário-mínimo sobre emprego e salários. Qual impacto você encontra para cada uma dessas variáveis usando os restaurantes de NJ apenas?

A variável de interação `low_wage:post` é significativa para o salário, indicando que o salário aumentou em 0.62 dólares por hora para os restaurantes de baixo salário em NJ. Para o emprego, a interação não é significativa a 95% de confiança.

```r
# variável de baixo salário
dados_nj <- by(dados_nj, dados_nj$sheet, function(conjunto) {
 conjunto$low_wage <- ifelse(any(conjunto$wage_st < 5 & conjunto$post == "mar"), 1, 0)
 conjunto
}) ▷
 do.call(what = rbind)

# regressão
m_salario_nj <- lm(wage_st ~ low_wage * post, data = dados_nj)
m_emprego_nj <- lm(fte ~ low_wage * post, data = dados_nj)

# sumário
stargazer::stargazer(
 m_salario_nj,
 m_emprego_nj,
 title = "Regressão de diferenças em diferenças",
 header = FALSE
)
```

### 3

> Compare as estimativas obtidas com o que você obteve anteriormente na parte (c). Os resultados são muito diferentes?

Os resultados para `low_wage` mostram um impacto maior no salário (aproximadamente 10 cents a mais). Para a empregabilidade, continua não significativo a 95% de confiança.

### e

Repita a regressão em (d) usando agora os restaurantes em PA.

```r
# filtrando observações
dados_pa <- subset(dados_long, state == "PA")

# variável de baixo salário
dados_pa <- by(dados_pa, dados_pa$sheet, function(conjunto) {
```

Tabela 4: Regressão de diferenças em diferenças

| | *Dependent variable:* | |
|---|---|---|
| | wage_st | fte |
| | (1) | (2) |
| low_wage | −0.651*** | −2.230* |
| | (0.023) | (1.211) |
| postnov_dez | −0.004 | −2.250 |
| | (0.029) | (1.501) |
| low_wage:postnov_dez | 0.616*** | 3.301* |
| | (0.033) | (1.713) |
| Constant | 5.113*** | 18.989*** |
| | (0.020) | (1.062) |
| Observations | 570 | 570 |
| $R^2$ | 0.776 | 0.008 |
| Adjusted $R^2$ | 0.775 | 0.002 |
| Residual Std. Error (df = 566) | 0.164 | 8.625 |
| F Statistic (df = 3; 566) | 653.183*** | 1.443 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

```r
 conjunto$low_wage <- ifelse(
 any(conjunto$wage_st < 5 & conjunto$post == "mar"),
 1,
 0
 )
 conjunto
}) ▷
 do.call(what = rbind)

# regressão
m_salario_pa <- lm(wage_st ~ low_wage * post, data = dados_pa)
m_emprego_pa <- lm(fte ~ low_wage * post, data = dados_pa)

# sumário
stargazer::stargazer(
 m_salario_pa,
 m_emprego_pa,
 title = "Regressão de diferenças em diferenças",
 header = FALSE
)
```

**1**

> Compare os resultados que você encontrou para PA com os resultados que você encontrou para NJ.

Obtemos resultados semelhantes, com `low_wage:postnov_dez` significativo para salário, numa magnitude de 0.35 cents, quase metade de NJ, e não significativo para emprego.

**2**

> Faça um teste estatístico para a hipótese que o coeficiente para a variável de baixo salário tenha o mesmo valor em NJ e PA.

Com p-valor de 0.013, o teste de diferença de médias nos permite rejeitar a hipótese nula de que os coeficientes são iguais. Os coeficientes não têm o mesmo valor em NJ e PA.

```r
# coeficientes e erro padrão
beta_1 <- coef(m_salario_nj)["low_wage:postnov_dez"]
beta_2 <- coef(m_salario_pa)["low_wage:postnov_dez"]

```

Tabela 5: Regressão de diferenças em diferenças

| | *Dependent variable:* | |
|---|---|---|
| | wage_st | fte |
| | (1) | (2) |
| low_wage | −0.632*** | −0.893 |
| | (0.071) | (2.666) |
| postnov_dez | −0.265*** | −3.848 |
| | (0.081) | (3.043) |
| low_wage:postnov_dez | 0.354*** | 2.813 |
| | (0.101) | (3.770) |
| Constant | 5.065*** | 20.696*** |
| | (0.057) | (2.152) |
| Observations | 132 | 132 |
| $R^2$ | 0.426 | 0.015 |
| Adjusted $R^2$ | 0.412 | −0.009 |
| Residual Std. Error (df = 128) | 0.275 | 10.318 |
| F Statistic (df = 3; 128) | 31.605*** | 0.630 |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

```r
se_1 <- summary(m_salario_nj)$coefficients["low_wage:postnov_dez", "Std. Error"]
se_2 <- summary(m_salario_pa)$coefficients["low_wage:postnov_dez", "Std. Error"]

# estatística de teste
z <- (beta_1 - beta_2) / sqrt(se_1^2 + se_2^2)

# p-valor
p_valor <- 2 * (1 - pnorm(abs(z)))

p_valor
```

```
low_wage:postnov_dez
          0.01308211
```

### 3

Por que verificar se o aumento do salário-mínimo em NJ teve impacto em PA pode ser uma maneira de confirmar que a metodologia produz resultados sensatos? O que você pode concluir com essa comparação?

Há dois caminhos: ou é esperado que o aumento do salário-mínimo em NJ tenha um efeito de *spillover* em PA — causado, por exemplo, pela migração de trabalhadores de PA para NJ e consequente redução da oferta de trabalho em PA, aumentando o salário médio —, de forma que o resultado confirma a metodologia, ou é esperado que o aumento do salário-mínimo em NJ não tenha efeito em PA, e o resultado indica que, ao menos, parte do resultado em NJ seja devido a fatores não observáveis (que também afetaram PA, mas não foram capturados na análise).

Sem o conhecimento de economia do trabalho e da região, eu suponho que o segundo caso seja mais provável, ou seja, que o aumento do salário-mínimo em NJ não tenha efeito em PA e que a metodologia tenha sucesso em capturar o efeito causal do aumento do salário-mínimo em NJ mas que parte desse efeito seja devido a fatores não observáveis, registrados no coeficiente de `low_wage:postnov_dez` em PA.

# Q3

Para essa questão, use o banco de dados `Guns.xlsx`. Uma descrição detalhada dos dados está contida no arquivo `Guns_Description.pdf`. Alguns estados dos EUA promulgaram leis que permitem que os cidadãos carreguem armas escondidas. Essas leis, conhecidas como "shall-issue laws", instruem as autoridades locais a emitirem uma permissão de armas ocultas a todos os requerentes que sejam cidadãos, sejam mentalmente competentes e não tenham sido condenados por crime doloso (alguns estados têm algumas restrições adicionais). Os proponentes argumentam que, se mais pessoas portarem armas ocultas, o crime diminuirá, porque os criminosos são dissuadidos de atacar outras pessoas. Oponentes argumentam que o crime aumentará por causa do uso acidental ou espontâneo da arma. Nessa questão, você usará os dados de Ayres & Donohue (2003) para estimar o efeito das leis de armas ocultas em crimes violentos.

```
dados <- readxl::read_excel("lista/data/Guns.xlsx") ▷
  transform(
    shall = as.factor(shall),
    stateid = as.factor(stateid)
  )
```

## a

Estime uma regressão de `ln(vio)` contra `shall` e uma regressão de `ln(vio)` contra `shall`, `incarc_rate`, `density`, `avginc`, `pop`, `pb1064`, `pw1064` e `pm1029`.

```
# regressões
m1 <- lm(log(vio) ~ shall, data = dados)
m2 <- lm(
  log(vio) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
  data = dados
)

# sumário
stargazer::stargazer(
 m1,
 m2,
```

```
  title = "Regressões",
  header = FALSE
)
```

### i

> Interprete o coeficiente de shall na segunda regressão. Essa estimativa pode ser considerada um efeito causal? Por que?

O coeficiente de `shall` em ambas regressões são significativos e negativos, o que quer dizer que a presença de leis de armas ocultas está associada a uma redução no crime violento. No entanto, a estimativa não pode ser considerada um efeito causal, pois não foi estabelecido um contrafactual válido: Não há controle de indivíduos ou de tempo, de forma que não é possível erificar a hipótese de tendências paralelas. O efeito pode ser devido a fatores não observados associados à passagem do tempo ou não presente em todos os estados que implementaram as leis de armas ocultas.

### ii

> As variáveis de controle adicionadas na segunda regressão mudam muito a magnitude do efeito das leis de armas ocultas na primeira regressão? Mudam a significância estatística do coeficiente estimado? Você espera que essa estimativa se aproxime mais de um efeito causal do que a anterior?

Na primeira regressão, o efeito de `shall` é de $e^{-0443} = 0.6421$, ou seja, uma redução de 35,8% no crime violento. Na segunda regressão, o efeito é de $e^{-0.368} = 0.692$, ou seja, uma redução de 30,8%. A diferença não é muito grande e também não muda a significância estatística do coeficiente, que continua significativo a 95% de confiança. A inclusão das variáveis de controle não torna a estimativa mais próxima de um efeito causal, pois os problemas citados anteriormente não foram resolvidos.

## b

> Os resultados se alteram se você adicionar efeitos fixos para estados e períodos de tempo (‘TWFE‘)? Qual dos resultados é mais crível para estimar um efeito causal e por quê?

Tabela 1: Regressões

| | *Dependent variable:* | |
|---|---|---|
| | log(vio) | |
| | (1) | (2) |
| shall1 | $-0.443^{***}$ | $-0.368^{***}$ |
| | (0.042) | (0.033) |
| incarc_rate | | $0.002^{***}$ |
| | | (0.0001) |
| density | | $0.027^{**}$ |
| | | (0.013) |
| avginc | | 0.001 |
| | | (0.008) |
| pop | | $0.043^{***}$ |
| | | (0.003) |
| pb1064 | | $0.081^{***}$ |
| | | (0.017) |
| pw1064 | | $0.031^{***}$ |
| | | (0.008) |
| pm1029 | | 0.009 |
| | | (0.011) |
| Constant | $6.135^{***}$ | $2.982^{***}$ |
| | (0.021) | (0.543) |
| Observations | 1,173 | 1,173 |
| $R^2$ | 0.087 | 0.564 |
| Adjusted $R^2$ | 0.086 | 0.561 |
| Residual Std. Error | 0.617 (df = 1171) | 0.428 (df = 1164) |
| F Statistic | $111.079^{***}$ (df = 1; 1171) | $188.411^{***}$ (df = 8; 1164) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Partindo para um modelo do tipo *within* com efeitos fixos individuais e temporais (Equação 0.1), representados no código pelo argumento `effect = "twoways"`, temos que o coeficiente de `shall` deixa de ser significativo, ou seja, a implementação do porte de armas escondidas não tem efeito significativo no crime violento.

O uso de efeitos fixos para estados e períodos de tempo é mais crível para estimar um efeito causal, uma vez que foram adicionados os controle para estados e tempo, permitindo a comparação de cada estado consigo mesmo ao longo do tempo. Isso permite que a hipótese de tendências paralelas seja testada, o que não era possível nos modelos anteriores.

$$y_{it} = x'it\beta + \alpha_i + \theta_t + \varepsilon_{it} \tag{0.1}$$

```r
# pacote para dados em painel
suppressPackageStartupMessages(library(plm))

# organização dos dados em painel
painel <- pdata.frame(dados, index = c("stateid", "year"))

# modelo
m3 <- plm(
  log(vio) ~ shall,
  data = painel,
  model = "within",
  effect = "twoways"
)

m4 <- plm(
  log(vio) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
  data = painel,
  model = "within",
  effect = "twoways"
)

# sumário
stargazer::stargazer(
  m3,
  m4,
  title = "Regressões em painel",
  header = FALSE
)
```

Tabela 2: Regressões em painel

| | *Dependent variable:* | |
|---|---|---|
| | log(vio) | |
| | (1) | (2) |
| shall1 | 0.002 | −0.028 |
| | (0.017) | (0.017) |
| incarc_rate | | 0.0001 |
| | | (0.0001) |
| density | | −0.092 |
| | | (0.076) |
| avginc | | 0.001 |
| | | (0.006) |
| pop | | −0.005 |
| | | (0.008) |
| pb1064 | | 0.029 |
| | | (0.023) |
| pw1064 | | 0.009 |
| | | (0.008) |
| pm1029 | | 0.073*** |
| | | (0.016) |
| Observations | 1,173 | 1,173 |
| $R^2$ | 0.00001 | 0.056 |
| Adjusted $R^2$ | −0.066 | −0.013 |
| F Statistic | 0.013 (df = 1; 1099) | 8.151*** (df = 8; 1092) |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

## c

> Usando o modelo de estudos de evento, tente fornecer alguma evidência da validade da hipótese de tendências paralelas entre tratados e controles.

Para testar a hipótese de tendências paralelas, é necessário que a diferença entre os grupos tratados e de controle seja constante ao longo do tempo, exceto pelo tratamento. Entretanto, vemos na Figura 2 que a inclinação da reta para o grupo tratado pré-tratamento é diferente da inclinação do grupo controle, indicando que a hipótese de tendências paralelas não é válida. Além disso, a inclinação do grupo controle é a mesma do grupo tratado pós-tratamento, o que pode ser um indício da ausência de efeito causal do tratamento.

A Figura 1 mostra que, em muitos dos estados, a quantidade de crimes violentos aumenta após o tratamento, o que é contrário à hipótese de que o porte de armas escondidas reduziria o crime violento.

```r
# ano de início do tratamento
anos_tratamento <- aggregate(
  year ~ stateid,
  data = subset(dados, shall == 1),
  FUN = min
)

# merge com dataset
dados <- merge(dados, anos_tratamento, by = "stateid", all.x = TRUE) |>
  (\(dadinho) sort_by(dadinho, dadinho$stateid, dadinho$year.x))()

# renomear variáveis
names(dados)[names(dados) %in% c("year.x", "year.y")] <- c("year", "year_tratamento")

# mais data engineering!
dados <- within(dados, {
  # anos relativos ao tratamento
  year_rel <- year - year_tratamento
  # indicador de pré ou pós tratamento
  tratamento <- ifelse(year_rel >= 0, "post", "pre")
  # indicador de controle
  tratamento <- ifelse(is.na(tratamento), "controle", tratamento)
})

# plotar quantidade de `vio` por ano relativo ao tratamento para ambos grupos
library(ggplot2)
dados |>
```

```r
  subset(tratamento != "controle") |>
  ggplot(aes(x = year_rel, y = vio, color = tratamento)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 3)) +
  facet_wrap(~ stateid, scales = "free") +
  labs(
    x = "Ano relativo ao tratamento",
    y = "Violência"
  ) +
  # removendo elementos para melhor visualização
  theme(
    # legenda
    legend.position = "none",
    # removendo números de eixos
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    # reduzindo strip text e largura
    strip.text = element_text(size = 6),
    strip.background = element_blank()
  )
```

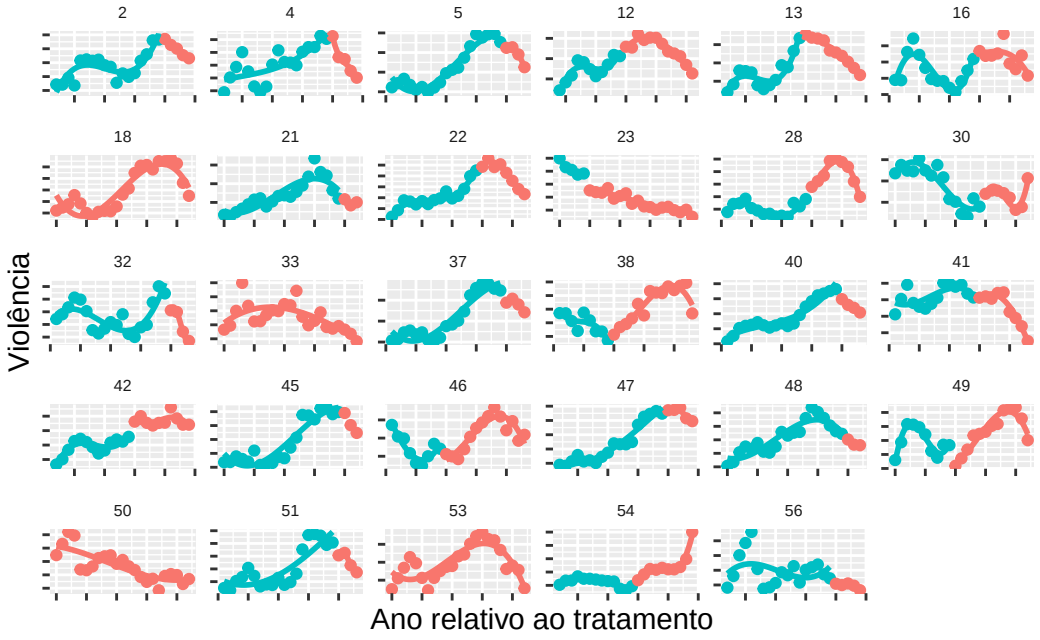


Figura 1: Violência por ano relativo ao tratamento

```r
# plotar quantidade de `vio` por ano, considerando apenas período sem tratamento
dados |>
  #subset(tratamento != "post") |>
  ggplot(aes(x = year, y = vio, color = tratamento)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Violência por ano",
    x = "Ano",
    y = "Violência"
  )
```

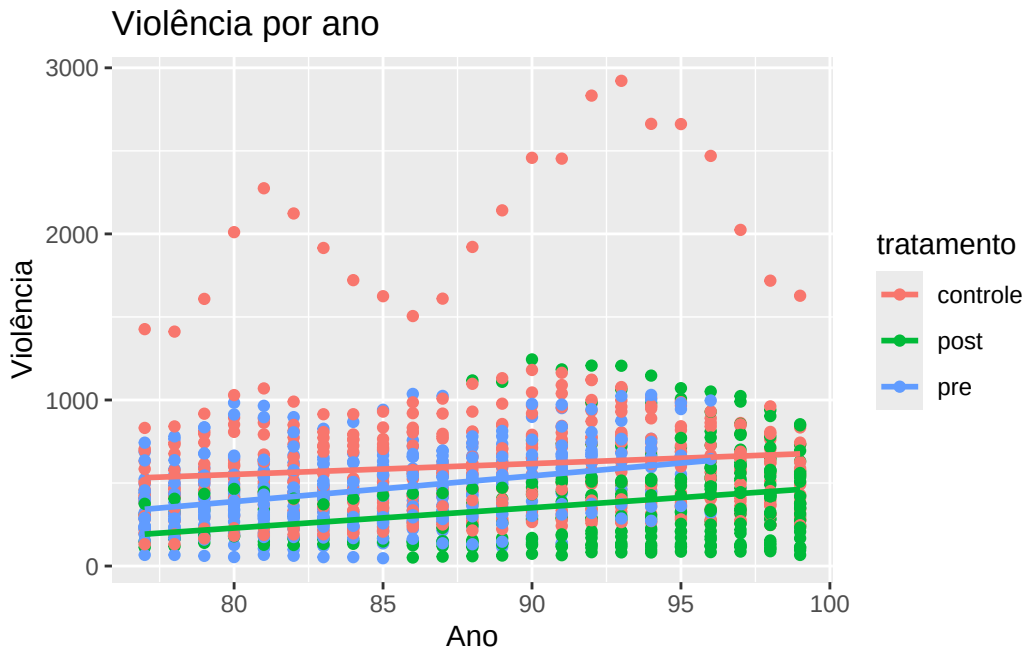


Figura 2: Violência por ano, grupos tratado e controle

## d

> Usando a nomenclatura de Callaway e Sant'Anna (2021), quantos grupos de tratados existem no banco de dados?

Neste caso, a quantidade de grupos de tratados corresponde aos anos distintos de implementação do tratamento, que são 23.

```r
# quantidade de grupos de tratados
grupos_unicos <- unique(dados$year[dados$shall == 1])
grupos_unicos
```

```
 [1] 95 96 97 98 99 88 89 90 91 92 93 94 77 78 79 80 81 82 83 84 85 86 87
```

```r
length(grupos_unicos)
```

```
[1] 23
```

## e

> Usando o método de Callaway e Sant'Anna (2021) estime os diferentes Efeitos Médios do Programa para Grupo-Período. Existe evidência de efeitos heterogêneos por grupo-período?

Sim. Na maior parte dos grupos, ou não há observações suficientes para estimação do modelo, ou não há efeito significativo, ou quando há é apenas pontual, em um ano. Apenas no grupo 95 há efeito significativo em vários anos consecutivos, sendo negativo para o ano de 1989 mas positivo de 1994 em diante.

```r
# mais features
dados <- within(dados, {
  G <- ifelse(is.na(year_tratamento), 0, year_tratamento)
  D <- ifelse(year >= year_tratamento & !is.na(year_tratamento), 1, 0)
  stateid_n <- as.numeric(stateid)
})

# Estimar os efeitos médios do programa para grupo-período
efeitos_medios <- did::att_gt(
  yname = "vio",
  tname = "year",
  idname = "stateid_n",
  gname = "G",
  xformla = ~ incarc_rate + pb1064 + pw1064 + avginc + density,
  data = dados,
  panel = TRUE
)

# sumário
summary(efeitos_medios)
```

```
Call:
did::att_gt(yname = "vio", tname = "year", idname = "stateid_n",
    gname = "G", xformla = ~incarc_rate + pb1064 + pw1064 + avginc +
        density, data = dados, panel = TRUE)

Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time Periods." Journ
230, 2021. <https://doi.org/10.1016/j.jeconom.2020.12.001>, <https://arxiv.org/abs/1803.09015>

Group-Time Average Treatment Effects:
 Group Time  ATT(g,t) Std. Error [95% Simult.  Conf. Band]
    82   78        NA         NA            NA          NA
    82   79        NA         NA            NA          NA
    82   80        NA         NA            NA          NA
    82   81        NA         NA            NA          NA
    82   82        NA         NA            NA          NA
    82   83        NA         NA            NA          NA
    82   84        NA         NA            NA          NA
    82   85        NA         NA            NA          NA
    82   86        NA         NA            NA          NA
    82   87        NA         NA            NA          NA
    82   88        NA         NA            NA          NA
    82   89        NA         NA            NA          NA
    82   90        NA         NA            NA          NA
    82   91        NA         NA            NA          NA
    82   92        NA         NA            NA          NA
    82   93        NA         NA            NA          NA
    82   94        NA         NA            NA          NA
    82   95        NA         NA            NA          NA
    82   96        NA         NA            NA          NA
    82   97        NA         NA            NA          NA
    82   98        NA         NA            NA          NA
    82   99        NA         NA            NA          NA
    86   78        NA         NA            NA          NA
    86   79        NA         NA            NA          NA
    86   80        NA         NA            NA          NA
    86   81        NA         NA            NA          NA
    86   82        NA         NA            NA          NA
    86   83        NA         NA            NA          NA
    86   84        NA         NA            NA          NA
    86   85        NA         NA            NA          NA
    86   86        NA         NA            NA          NA
    86   87        NA         NA            NA          NA
```

| | | | | | |
|---|---|---|---|---|---|
| 86 | 88 | NA | NA | NA | NA |
| 86 | 89 | NA | NA | NA | NA |
| 86 | 90 | NA | NA | NA | NA |
| 86 | 91 | NA | NA | NA | NA |
| 86 | 92 | NA | NA | NA | NA |
| 86 | 93 | NA | NA | NA | NA |
| 86 | 94 | NA | NA | NA | NA |
| 86 | 95 | NA | NA | NA | NA |
| 86 | 96 | NA | NA | NA | NA |
| 86 | 97 | NA | NA | NA | NA |
| 86 | 98 | NA | NA | NA | NA |
| 86 | 99 | NA | NA | NA | NA |
| 87 | 78 | NA | NA | NA | NA |
| 87 | 79 | NA | NA | NA | NA |
| 87 | 80 | NA | NA | NA | NA |
| 87 | 81 | NA | NA | NA | NA |
| 87 | 82 | NA | NA | NA | NA |
| 87 | 83 | NA | NA | NA | NA |
| 87 | 84 | NA | NA | NA | NA |
| 87 | 85 | NA | NA | NA | NA |
| 87 | 86 | NA | NA | NA | NA |
| 87 | 87 | NA | NA | NA | NA |
| 87 | 88 | NA | NA | NA | NA |
| 87 | 89 | NA | NA | NA | NA |
| 87 | 90 | NA | NA | NA | NA |
| 87 | 91 | NA | NA | NA | NA |
| 87 | 92 | NA | NA | NA | NA |
| 87 | 93 | NA | NA | NA | NA |
| 87 | 94 | NA | NA | NA | NA |
| 87 | 95 | NA | NA | NA | NA |
| 87 | 96 | NA | NA | NA | NA |
| 87 | 97 | NA | NA | NA | NA |
| 87 | 98 | NA | NA | NA | NA |
| 87 | 99 | NA | NA | NA | NA |
| 88 | 78 | NA | NA | NA | NA |
| 88 | 79 | NA | NA | NA | NA |
| 88 | 80 | NA | NA | NA | NA |
| 88 | 81 | NA | NA | NA | NA |
| 88 | 82 | NA | NA | NA | NA |
| 88 | 83 | NA | NA | NA | NA |
| 88 | 84 | NA | NA | NA | NA |
| 88 | 85 | NA | NA | NA | NA |
| 88 | 86 | NA | NA | NA | NA |

```
88   87        NA         NA           NA          NA
88   88        NA         NA           NA          NA
88   89        NA         NA           NA          NA
88   90        NA         NA           NA          NA
88   91        NA         NA           NA          NA
88   92        NA         NA           NA          NA
88   93        NA         NA           NA          NA
88   94        NA         NA           NA          NA
88   95        NA         NA           NA          NA
88   96        NA         NA           NA          NA
88   97        NA         NA           NA          NA
88   98        NA         NA           NA          NA
88   99        NA         NA           NA          NA
90   78   16.8113    14.2500      -35.8785     69.5011
90   79    2.2446    16.1696      -57.5430     62.0322
90   80  -16.0294    10.4382      -54.6248     22.5660
90   81  -29.3239    15.6927      -87.3481     28.7002
90   82  -33.4837    25.4814     -127.7021     60.7347
90   83   17.5104     9.7476      -18.5315     53.5523
90   84   -4.3933     8.1127      -34.3901     25.6034
90   85   -3.1396     8.1436      -33.2506     26.9714
90   86   -9.0373    11.0189      -49.7801     31.7056
90   87   30.9261    13.2425      -18.0383     79.8905
90   88    9.9978    31.5071     -106.5007    126.4964
90   89   11.1119     9.0010      -22.1696     44.3934
90   90  -30.9216     8.1300      -60.9824     -0.8608 *
90   91  -64.4779    38.5958     -207.1871     78.2313
90   92 -106.4677    35.9431     -239.3685     26.4331
90   93 -108.3028    46.3288     -279.6050     62.9994
90   94  -88.5237    41.4425     -241.7584     64.7111
90   95  -59.2638    28.5275     -164.7451     46.2175
90   96  -32.8639    62.9775     -265.7252    199.9974
90   97  -56.1566    56.1643     -263.8258    151.5126
90   98  -88.0526    85.3405     -403.6017    227.4965
90   99   -3.0983    66.9395     -250.6092    244.4126
91   78        NA         NA           NA          NA
91   79        NA         NA           NA          NA
91   80        NA         NA           NA          NA
91   81        NA         NA           NA          NA
91   82        NA         NA           NA          NA
91   83        NA         NA           NA          NA
91   84        NA         NA           NA          NA
91   85        NA         NA           NA          NA
```

| | | | | | |
|---|---|---|---|---|---|
| 91 | 86 | NA | NA | NA | NA |
| 91 | 87 | NA | NA | NA | NA |
| 91 | 88 | NA | NA | NA | NA |
| 91 | 89 | -36.1013 | 20.1173 | -110.4857 | 38.2830 |
| 91 | 90 | -46.5126 | 14.9563 | -101.8141 | 8.7890 |
| 91 | 91 | NA | NA | NA | NA |
| 91 | 92 | NA | NA | NA | NA |
| 91 | 93 | NA | NA | NA | NA |
| 91 | 94 | NA | NA | NA | NA |
| 91 | 95 | NA | NA | NA | NA |
| 91 | 96 | NA | NA | NA | NA |
| 91 | 97 | NA | NA | NA | NA |
| 91 | 98 | NA | NA | NA | NA |
| 91 | 99 | NA | NA | NA | NA |
| 92 | 78 | 28.1476 | 21.3252 | -50.7031 | 106.9983 |
| 92 | 79 | -0.3729 | 49.7066 | -184.1644 | 183.4186 |
| 92 | 80 | -40.9712 | 11.6313 | -83.9784 | 2.0359 |
| 92 | 81 | NA | NA | NA | NA |
| 92 | 82 | NA | NA | NA | NA |
| 92 | 83 | NA | NA | NA | NA |
| 92 | 84 | NA | NA | NA | NA |
| 92 | 85 | NA | NA | NA | NA |
| 92 | 86 | NA | NA | NA | NA |
| 92 | 87 | NA | NA | NA | NA |
| 92 | 88 | NA | NA | NA | NA |
| 92 | 89 | NA | NA | NA | NA |
| 92 | 90 | NA | NA | NA | NA |
| 92 | 91 | NA | NA | NA | NA |
| 92 | 92 | NA | NA | NA | NA |
| 92 | 93 | NA | NA | NA | NA |
| 92 | 94 | NA | NA | NA | NA |
| 92 | 95 | NA | NA | NA | NA |
| 92 | 96 | NA | NA | NA | NA |
| 92 | 97 | NA | NA | NA | NA |
| 92 | 98 | NA | NA | NA | NA |
| 92 | 99 | NA | NA | NA | NA |
| 95 | 78 | 4.2739 | 19.8934 | -69.2824 | 77.8302 |
| 95 | 79 | -18.2749 | 18.7835 | -87.7274 | 51.1775 |
| 95 | 80 | -10.0527 | 31.4857 | -126.4721 | 106.3667 |
| 95 | 81 | NA | NA | NA | NA |
| 95 | 82 | NA | NA | NA | NA |
| 95 | 83 | NA | NA | NA | NA |
| 95 | 84 | NA | NA | NA | NA |

| | | | | | | |
|---|---|---|---|---|---|---|
| 95 | 85 | NA | NA | NA | NA | |
| 95 | 86 | -18.5775 | 23.3120 | -104.7741 | 67.6192 | |
| 95 | 87 | 11.9033 | 26.2937 | -85.3184 | 109.1249 | |
| 95 | 88 | 5.6073 | 18.1068 | -61.3430 | 72.5576 | |
| 95 | 89 | -60.1374 | 17.9908 | -126.6588 | 6.3840 | |
| 95 | 90 | -20.3368 | 25.3727 | -114.1531 | 73.4796 | |
| 95 | 91 | -30.9435 | 35.6845 | -162.8880 | 101.0011 | |
| 95 | 92 | 3.9848 | 16.0350 | -55.3051 | 63.2746 | |
| 95 | 93 | 63.0494 | 32.3292 | -56.4888 | 182.5876 | |
| 95 | 94 | 52.7887 | 13.6058 | 2.4810 | 103.0964 | * |
| 95 | 95 | 61.8424 | 15.7665 | 3.5453 | 120.1395 | * |
| 95 | 96 | 75.6614 | 16.5871 | 14.3302 | 136.9926 | * |
| 95 | 97 | 94.9832 | 23.0666 | 9.6937 | 180.2727 | * |
| 95 | 98 | 54.3264 | 31.6748 | -62.7921 | 171.4449 | |
| 95 | 99 | 71.2528 | 28.9393 | -35.7512 | 178.2568 | |
| 96 | 78 | NA | NA | NA | NA | |
| 96 | 79 | NA | NA | NA | NA | |
| 96 | 80 | -14.0876 | 26.7296 | -112.9210 | 84.7458 | |
| 96 | 81 | NA | NA | NA | NA | |
| 96 | 82 | NA | NA | NA | NA | |
| 96 | 83 | NA | NA | NA | NA | |
| 96 | 84 | NA | NA | NA | NA | |
| 96 | 85 | NA | NA | NA | NA | |
| 96 | 86 | NA | NA | NA | NA | |
| 96 | 87 | NA | NA | NA | NA | |
| 96 | 88 | 19.4054 | 16.2034 | -40.5073 | 79.3181 | |
| 96 | 89 | -36.3295 | 54.8688 | -239.2086 | 166.5495 | |
| 96 | 90 | -60.0790 | 29.4724 | -169.0539 | 48.8959 | |
| 96 | 91 | -67.2576 | 23.4836 | -154.0890 | 19.5738 | |
| 96 | 92 | -5.5084 | 18.2845 | -73.1157 | 62.0990 | |
| 96 | 93 | 104.8199 | 42.7937 | -53.4112 | 263.0509 | |
| 96 | 94 | 105.6578 | 34.1807 | -20.7265 | 232.0420 | |
| 96 | 95 | 33.8210 | 16.5716 | -27.4530 | 95.0949 | |
| 96 | 96 | -2.5534 | 16.3600 | -63.0449 | 57.9380 | |
| 96 | 97 | 2.9448 | 20.0314 | -71.1220 | 77.0116 | |
| 96 | 98 | -12.8283 | 41.2687 | -165.4206 | 139.7639 | |
| 96 | 99 | -32.2859 | 50.8514 | -220.3104 | 155.7385 | |
| 97 | 78 | NA | NA | NA | NA | |
| 97 | 79 | NA | NA | NA | NA | |
| 97 | 80 | NA | NA | NA | NA | |
| 97 | 81 | NA | NA | NA | NA | |
| 97 | 82 | NA | NA | NA | NA | |
| 97 | 83 | NA | NA | NA | NA | |

```
    97   84        NA         NA            NA          NA
    97   85        NA         NA            NA          NA
    97   86        NA         NA            NA          NA
    97   87        NA         NA            NA          NA
    97   88        NA         NA            NA          NA
    97   89        NA         NA            NA          NA
    97   90        NA         NA            NA          NA
    97   91  -51.3260    35.4410     -182.3701     79.7181
    97   92        NA         NA            NA          NA
    97   93   27.9079    57.8016     -185.8152    241.6310
    97   94        NA         NA            NA          NA
    97   95        NA         NA            NA          NA
    97   96        NA         NA            NA          NA
    97   97        NA         NA            NA          NA
    97   98        NA         NA            NA          NA
    97   99        NA         NA            NA          NA
---
Signif. codes: `*' confidence band does not cover 0

Control Group:  Never Treated,  Anticipation Periods:  0
Estimation Method:  Doubly Robust
```

## f

> A partir da resposta anterior, obtenha o efeito médio agregado total e compare esse resultado como o que você obteve na letra "a".

O efeito médio agregado total é de -6,56, bem menor do que o obtido na letra "a", onde os coeficientes foram de -0.44 e -0.37.

```r
mean(efeitos_medios$att, na.rm = TRUE)
```

```
[1] -6.560892
```

```r
coef(m1)["shall1"]
```

```
    shall1
-0.4429646
```

```r
coef(m2)["shall1"]
```

```
    shall1
-0.3683869
```

## g

> Repita a análise usando ‘ln(rob)‘ e ‘ln(mur)‘ no lugar de ‘ln(vio)‘. Coloque seus resultados em tabelas arrumadas.

O resultado da regressão com efeitos fixos para estados e períodos de tempo mostra que a implementação do porte de armas escondidas também não tem efeito significativo no número de roubos e assassinatos.

```r
# modelo
m5 <- plm(
  log(rob) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
  data = painel,
  model = "within",
  effect = "twoways"
)

m6 <- plm(
  log(mur) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
  data = painel,
  model = "within",
  effect = "twoways"
)

# sumário
stargazer::stargazer(
  m5,
  m6,
  title = "Regressões em painel",
  header = FALSE
)
```

Tabela 3: Regressões em painel

| | *Dependent variable:* | |
|---|---|---|
| | log(rob) | log(mur) |
| | (1) | (2) |
| shall1 | 0.027 | −0.015 |
| | (0.024) | (0.025) |
| incarc_rate | 0.00003 | −0.0001 |
| | (0.0001) | (0.0001) |
| density | −0.045 | −0.544*** |
| | (0.105) | (0.110) |
| avginc | 0.014 | 0.057*** |
| | (0.009) | (0.009) |
| pop | 0.00002 | −0.032*** |
| | (0.011) | (0.011) |
| pb1064 | 0.014 | 0.022 |
| | (0.031) | (0.033) |
| pw1064 | −0.013 | −0.0005 |
| | (0.011) | (0.011) |
| pm1029 | 0.105*** | 0.069*** |
| | (0.022) | (0.023) |
| Observations | 1,173 | 1,173 |
| $R^2$ | 0.049 | 0.116 |
| Adjusted $R^2$ | −0.021 | 0.051 |
| F Statistic (df = 8; 1092) | 7.048*** | 17.845*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

## h

> Baseado na sua análise, que conclusões você tiraria a respeito dos efeitos das leis de armas ocultas sobre as taxas de criminalidade? Use uma linguagem clara e acessível, de maneira que até o deputado Marcel Van Hatten (Novo/RS) entenda.

Eu tenho alguma experiência adestrando cachorros, mas asnos não são minha especialidade. Farei meu melhor.

Não existem evidências para se afirmar que as "shall issue law" têm qualquer efeito sobre crimes violentos, roubos e assassinatos. A análise estatística não encontrou efeitos causais entre a promulgação dessas leis e a criminalidade. Como podemos ver na Figura 1, em alguns estados a incidência de crimes violentos diminuiu e em outros aumentou. Isso quer dizer que existe uma grande variabilidade nos resultados, e a verdadeira causa dessas variações não foi capturada na análise. Se houvesse efeito causal das leis de armas ocultas, esperaríamos que a criminalidade diminuísse de forma consistente após a implementação dessas leis em todos os estados, o que não foi observado. Portanto, não podemos afirmar que as "shall issue laws" têm qualquer efeito sobre a criminalidade.

Além disso, a Figura 2 mostra que a tendência de crimes violentos entre os estados que nunca implementaram as leis de armas ocultas é a mesma que a dos estados que implementaram após a implementação. Isso também reforça a ideia de que a implementação das leis de armas ocultas não teve efeito sobre a criminalidade.

# Q4

Para essa questão, use o banco de dados CARD.xlsx. Uma descrição detalhada dos dados está contida no arquivo CARD_Description.pdf. Card (1995) estimou o retorno da educação para homens jovens no ano de 1976 usando como instrumento para educação uma variável dummy indicando se a pessoa havia crescido próximo a uma faculdade com cursos de graduação de 4 anos. Nessa questão você vai repetir alguns passos da análise e realizar algumas extensões.

```
dados <- readxl::read_excel("lista/data/CARD.xlsx")
```

## a

> Quais as condições que a proximidade de uma faculdade com cursos de graduação de 4 anos deve satisfazer para ser um instrumento válido para educação? Qual dessas condições pode ser testada empiricamente e qual o procedimento para isso?

A proximidade de uma faculdade com cursos de graduação de 4 anos deve satisfazer as seguintes condições para ser um instrumento válido para educação:

1. Relevância: A proximidade da faculdade deve estar associada à educação do indivíduo.
2. Exogeneidade: A proximidade da faculdade não deve estar associada a características não observadas que afetam o rendimento do indivíduo (não pode estar correlacionada com o termo do erro).
3. Exclusão: A proximidade da faculdade não deve afetar o rendimento do indivíduo diretamente, mas apenas por meio da educação.

A condição de relevância pode ser testada empiricamente verificando se a proximidade da faculdade está associada à educação do indivíduo. A condição de exclusão pode ser testada empiricamente verificando se a proximidade da faculdade afeta o rendimento do indivíduo diretamente.

O procedimento para testar a relevância é estimar a equação de educação com a variável de proximidade da faculdade como variável explicativa. Se o coeficiente for significativo, a variável é relevante. Para testar a exclusão, deve-se estimar a equação de rendimento com a variável de proximidade da faculdade como variável explicativa. Se o coeficiente não for significativo, a variável é válida.

## b

Estime equações para o efeito da educação sobre o logaritmo dos rendimentos usando MQO e incluindo as variáveis: educ, exper, expersq, black, south e smsa. A seguir, adicione ao primeiro modelo as variáveis indicadoras das regiões (reg661 – reg668)e smsa66. Em terceiro lugar, adicione ao segundo modelo as variáveis fatheduce motheduc. Por fim, adicione ao terceiro modelo as variáveis momdad14 e sinmon14. Organize os quatro modelos em uma tabela exibindo apenas os coeficientes das variáveis do primeiro modelo e indicando as variáveis incluídas nos demais modelos. Interprete os resultados obtidos para a variável educ, mostrando o que significa essa magnitude e o quão diferente são as estimativas nas diferentes especificações.

O coeficiente de `educ` é significativo em todos os modelos (0.074, 0.075, 0.074 e 0.073), indicando que um ano a mais de educação aumenta o salário em média 7,7% ($e^0$.074). Os coeficientes são muitos próximos entre si, indicando que a inclusão de variáveis adicionais não afetou a estimativa de `educ`.

```r
# modelo 1
m1 <- lm(lwage ~ educ + exper + expersq + black + south + smsa, data = dados)

# modelo 2
m2 <- lm(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 +reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
  data = dados
)

# modelo 3
m3 <- lm(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc,
  data = dados
)

# modelo 4
m4 <- lm(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc + momdad14 + sinmom14,
  data = dados
)
```

```r

# sumário
stargazer::stargazer(
  m1,
  m2,
  m3,
  m4,
  title = "Regressões",
  header = FALSE,
  font.size = "scriptsize",
  single.row = TRUE,
  column.sep.width = "1pt"
)
```

Tabela 1: Regressões

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | lwage | | | |
| | (1) | (2) | (3) | (4) |
| educ | 0.074*** (0.004) | 0.075*** (0.003) | 0.074*** (0.004) | 0.073*** (0.004) |
| exper | 0.084*** (0.007) | 0.085*** (0.007) | 0.090*** (0.008) | 0.090*** (0.008) |
| expersq | −0.002*** (0.0003) | −0.002*** (0.0003) | −0.002*** (0.0004) | −0.002*** (0.0004) |
| black | −0.190*** (0.018) | −0.199*** (0.018) | −0.165*** (0.025) | −0.163*** (0.025) |
| south | −0.125*** (0.015) | −0.148*** (0.026) | −0.124*** (0.031) | −0.123*** (0.031) |
| smsa | 0.161*** (0.016) | 0.136*** (0.020) | 0.138*** (0.024) | 0.137*** (0.024) |
| reg661 | | −0.119*** (0.039) | −0.092** (0.046) | −0.094** (0.046) |
| reg662 | | −0.022 (0.028) | −0.003 (0.032) | −0.002 (0.032) |
| reg663 | | 0.026 (0.027) | 0.030 (0.031) | 0.029 (0.031) |
| reg664 | | −0.063* (0.036) | −0.053 (0.041) | −0.055 (0.041) |
| reg665 | | 0.009 (0.036) | 0.004 (0.042) | 0.001 (0.042) |
| reg666 | | 0.022 (0.040) | 0.016 (0.049) | 0.014 (0.049) |
| reg667 | | −0.001 (0.039) | 0.009 (0.046) | 0.008 (0.046) |
| reg668 | | −0.175*** (0.046) | −0.159*** (0.052) | −0.158*** (0.052) |
| smsa66 | | 0.026 (0.019) | 0.024 (0.023) | 0.025 (0.023) |
| fatheduc | | | −0.001 (0.003) | −0.0005 (0.003) |
| motheduc | | | 0.008** (0.003) | 0.008** (0.003) |
| momdad14 | | | | 0.071* (0.040) |
| sinmom14 | | | | 0.126 (0.100) |
| Constant | 4.734*** (0.068) | 4.739*** (0.072) | 4.615*** (0.085) | 4.557*** (0.091) |
| Observations | 3,010 | 3,010 | 2,220 | 2,220 |
| $R^2$ | 0.291 | 0.300 | 0.274 | 0.276 |
| Adjusted $R^2$ | 0.289 | 0.296 | 0.269 | 0.269 |
| Residual Std. Error | 0.374 (df = 3003) | 0.372 (df = 2994) | 0.376 (df = 2202) | 0.376 (df = 2200) |
| F Statistic | 204.932*** (df = 6; 3003) | 85.476*** (df = 15; 2994) | 48.967*** (df = 17; 2202) | 44.031*** (df = 19; 2200) |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

## c

> Qual crítica pode ser feita em relação às estimativas para o efeito da educação estimado na equação b? Essa estimativa será viesada? Se for, em que direção deve ocorrer esse viés?

A crítica que pode ser feita é que a variável `educ` pode ser endógena, ou seja, correlacionada com o termo do erro. Isso ocorre porque a educação pode ser afetada por características não observadas que afetam o rendimento do indivíduo. Por exemplo, habilidade. Indivíduos mais talentosos ou motivados podem investir mais em educação e, ao mesmo tempo, obter maiores salários. Além disso, podem existir variáveis confundidoras que afetam tanto a educação quanto o salário (como estrutura familiar).

Outro problema é direção da causalidade: salários mais altos podem permitir que indivíduos invistam mais em educação.

Se a educação for positivamente correlacionada com o termo do erro, a estimativa de MQO será viesada para cima, ou seja, o efeito da educação sobre o salário será superestimado.

## d

> Para cada modelo estimado na letra b, estime o modelo correspondente à forma reduzida. Organize os quatro modelos em uma tabela exibindo apenas os coeficientes das variáveis do primeiro modelo e indicando as variáveis incluídas nos demais modelos. Qual variável dos modelos da letra b foi substituída? Interprete os resultados obtidos para essa variável incluída, mostrando o que significa essa magnitude e o quão diferente são as estimativas nas diferentes especificações.

A variável `educ` foi substituída pela variável `nearc4` nos modelos da forma reduzida. O coeficiente de `nearc4` é significativo em dois dos modelos, indicando que a proximidade de uma faculdade de 4 anos pode aumentar o salário. A inclusão de variáveis adicionais afeta a estimativa de `nearc4`, sendo significativa nos modelos com menos covariáveis.

Obs.: a quantidade reduzida de observações nos modelos 3 e 4 pode ter afetado a detecção do efeito.

```r
# modelo 1
m1_reduzido <- lm(lwage ~ nearc4, data = dados)

# modelo 2
m2_reduzido <- lm(
  lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
  reg667 + reg668 + smsa66,
```

```r
  data = dados
)

# modelo 3
m3_reduzido <- lm(
  lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
  reg667 + reg668 + smsa66 + fatheduc + motheduc,
  data = dados
)

# modelo 4
m4_reduzido <- lm(
  lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
  reg667 + reg668 + smsa66 + fatheduc + motheduc + momdad14 + sinmom14,
  data = dados
)

# sumário
stargazer::stargazer(
  m1_reduzido,
  m2_reduzido,
  m3_reduzido,
  m4_reduzido,
  header = FALSE,
  font.size = "scriptsize",
  single.row = TRUE,
  column.sep.width = "1pt",
  title = "Regressões da forma reduzida"
)
```

### e

> Para cada modelo estimado na letra b, estime o modelo correspondente ao primeiro estágio. Organize os quatro modelos em uma tabela exibindo apenas os coeficientes das variáveis do primeiro modelo e indicando as variáveis incluídas nos demais modelos. Quais variáveis dos modelos da letra b foram substituídas? Interprete os resultados obtidos para essa variável incluída, mostrando o que significa essa magnitude e o quão diferente são as estimativas nas diferentes especificações.

A variável `lwage` foi substituída pela variável `educ` como variável dependente e esta foi removida como covariável. O coeficiente de `nearc4` é significativo em todos os modelos, indicando que a

Tabela 2: Regressões da forma reduzida

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | lwage | | | |
| | (1) | (2) | (3) | (4) |
| nearc4 | 0.156*** (0.017) | 0.043** (0.019) | 0.027 (0.022) | 0.024 (0.022) |
| reg661 | | −0.132*** (0.044) | −0.114** (0.051) | −0.120** (0.051) |
| reg662 | | −0.052 (0.032) | −0.040 (0.035) | −0.039 (0.035) |
| reg663 | | −0.014 (0.031) | −0.016 (0.035) | −0.018 (0.034) |
| reg664 | | −0.081** (0.040) | −0.082* (0.045) | −0.086* (0.045) |
| reg665 | | −0.258*** (0.031) | −0.184*** (0.036) | −0.185*** (0.036) |
| reg666 | | −0.252*** (0.037) | −0.192*** (0.044) | −0.193*** (0.044) |
| reg667 | | −0.247*** (0.035) | −0.171*** (0.040) | −0.171*** (0.040) |
| reg668 | | −0.146*** (0.052) | −0.164*** (0.058) | −0.162*** (0.058) |
| smsa66 | | 0.118*** (0.019) | 0.100*** (0.022) | 0.102*** (0.022) |
| fatheduc | | | 0.007** (0.003) | 0.007** (0.003) |
| motheduc | | | 0.017*** (0.004) | 0.017*** (0.004) |
| momdad14 | | | | 0.171*** (0.044) |
| sinmom14 | | | | 0.237** (0.111) |
| Constant | 6.155*** (0.014) | 6.287*** (0.031) | 6.051*** (0.047) | 5.887*** (0.063) |
| Observations | 3,010 | 3,010 | 2,220 | 2,220 |
| $R^2$ | 0.027 | 0.105 | 0.099 | 0.105 |
| Adjusted $R^2$ | 0.026 | 0.102 | 0.094 | 0.100 |
| Residual Std. Error | 0.438 (df = 3008) | 0.421 (df = 2999) | 0.418 (df = 2207) | 0.417 (df = 2205) |
| F Statistic | 82.745*** (df = 1; 3008) | 35.212*** (df = 10; 2999) | 20.212*** (df = 12; 2207) | 18.542*** (df = 14; 2205) |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

proximidade de uma faculdade de 4 anos aumenta a educação. Portanto, `nearc4` é relevante para `educ` e pode ser um instrumento adequado se a hipótese de exclusão não for violada. A inclusão de variáveis adicionais afeta a estimativa de `nearc4`, reduzindo seu coeficiente e aumentando seu desvio-padrão.

```
# modelo 1
m1_estagio1 <- lm(
  educ ~ nearc4 + exper + expersq + black + south + smsa,
  data = dados
)

# modelo 2
m2_estagio1 <- lm(
  educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
  data = dados
)

# modelo 3
m3_estagio1 <- lm(
  educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
```

```
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc,
  data = dados
)

# modelo 4
m4_estagio1 <- lm(
  educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc + momdad14 + sinmom14,
  data = dados
)

# sumário
stargazer::stargazer(
  m1_estagio1,
  m2_estagio1,
  m3_estagio1,
  m4_estagio1,
  header = FALSE,
  font.size = "scriptsize",
  single.row = TRUE,
  column.sep.width = "1pt",
  title = "Estágio 1"
)
```

## f

> Para cada modelo estimado na letra b, estime agora usando a variável nearc4 como instrumento para educação. Organize os quatro modelos em uma tabela exibindo apenas os coeficientes das variáveis do primeiro modelo e indicando as variáveis incluídas nos demais modelos. Interprete os resultados obtidos para a variável educ, mostrando o que significa essa magnitude e o quão diferente são as estimativas nas diferentes especificações. Compare essas estimativas às obtidas na letra b, comentando se a diferença obtida é consistente com a sua resposta na letra c.

A variável `educ` é significativa apenas nos dois primeiros modelos, diferentemente dos modelos da letra b em que `educ` foi significativo em todos os modelos. Considerando que há centenas de linhas de dados faltantes em relação à educação dos pais, os modelos 3 e 4 contam com

Tabela 3: Estágio 1

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | educ | | | |
| | (1) | (2) | (3) | (4) |
| nearc4 | 0.337*** (0.083) | 0.320*** (0.088) | 0.261*** (0.098) | 0.250** (0.098) |
| exper | −0.410*** (0.034) | −0.413*** (0.034) | −0.380*** (0.038) | −0.379*** (0.038) |
| expersq | 0.001 (0.002) | 0.001 (0.002) | 0.003 (0.002) | 0.002 (0.002) |
| black | −1.006*** (0.090) | −0.936*** (0.094) | −0.344*** (0.122) | −0.310** (0.122) |
| south | −0.291*** (0.079) | −0.052 (0.135) | −0.054 (0.155) | −0.047 (0.154) |
| smsa | 0.404*** (0.085) | 0.402*** (0.105) | 0.422*** (0.117) | 0.411*** (0.116) |
| reg661 | | −0.210 (0.202) | −0.386* (0.226) | −0.416* (0.226) |
| reg662 | | −0.289** (0.147) | −0.320** (0.158) | −0.318** (0.157) |
| reg663 | | −0.238* (0.143) | −0.361** (0.154) | −0.374** (0.154) |
| reg664 | | −0.093 (0.186) | −0.089 (0.202) | −0.114 (0.202) |
| reg665 | | −0.483** (0.188) | −0.286 (0.209) | −0.307 (0.208) |
| reg666 | | −0.513** (0.210) | −0.401* (0.243) | −0.427* (0.242) |
| reg667 | | −0.427** (0.206) | −0.238 (0.228) | −0.250 (0.227) |
| reg668 | | 0.314 (0.242) | 0.072 (0.258) | 0.076 (0.257) |
| smsa66 | | 0.025 (0.106) | −0.217* (0.117) | −0.202* (0.116) |
| fatheduc | | | 0.111*** (0.015) | 0.112*** (0.015) |
| motheduc | | | 0.133*** (0.017) | 0.133*** (0.017) |
| momdad14 | | | | 0.797*** (0.197) |
| sinmom14 | | | | 0.850* (0.493) |
| Constant | 16.659*** (0.176) | 16.849*** (0.211) | 14.030*** (0.298) | 13.281*** (0.349) |
| Observations | 3,010 | 3,010 | 2,220 | 2,220 |
| $R^2$ | 0.474 | 0.477 | 0.486 | 0.490 |
| Adjusted $R^2$ | 0.473 | 0.474 | 0.482 | 0.485 |
| Residual Std. Error | 1.943 (df = 3003) | 1.941 (df = 2994) | 1.862 (df = 2202) | 1.856 (df = 2200) |
| F Statistic | 451.866*** (df = 6; 3003) | 182.129*** (df = 15; 2994) | 122.485*** (df = 17; 2202) | 111.175*** (df = 19; 2200) |
| *Note:* | | | | *p<0.1; **p<0.05; ***p<0.01 |

quase 800 observações a menos, sendo mais difícil detectar um efeito. Além disso, a maioria das variáveis incluídas nos modelos 3 e 4 não são significativas, resultando na adição de ruído trazido por essas covariáveis.

## g

> Aponte e discuta brevemente ao menos duas razões pelas quais a hipótese de exclusão da proximidade de uma faculdade de 4 anos da equação de salários pode ser violada

A faculdade pode estar localizada em região com melhores empregos e maior renda, de forma que a localidade seja uma variável confundidora, afetando tanto `educ` quanto `lwage`. Além isso, região com uma universidade presente pode estar associado a uma cultura educacional mais forte, com uma população que valoriza mais a educação.

## h

> Faça uma regressão da variável 'IQ' contra a variável 'nearc4' para verificar se existe correlação entre o coeficiente de 'QI' da pessoa com a proximidade de uma faculdade. O que você verifica? O que isso significa para a hipótese de exclusão do instrumento 'nearc4'?

O coeficiente de `nearc4` é significativo, indicando que a proximidade de uma faculdade de 4 anos está associada ao QI da pessoa. Isso sugere que a hipótese de exclusão do instrumento `nearc4` é violada, uma vez que o instrumento `nearc4` está afetando `lwage` não só através apenas de `educ`, mas também a partir de `IQ`.

```r
# modelo
m_iq <- lm(IQ ~ nearc4, data = dados)

# sumário
stargazer::stargazer(
  m_iq,
  header = FALSE,
  title = "Regressão de IQ contra nearc4"
)
```

Tabela 4

```r
# pacote
library(AER)

# modelo 1
m1_iv <- ivreg(
  lwage ~ educ + exper + expersq + black + south + smsa |
  nearc4 + exper + expersq + black + south + smsa,
  data = dados
)

# modelo 2
m2_iv <- ivreg(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 |
  nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
  data = dados
)

# modelo 3
m3_iv <- ivreg(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc |
  nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc,
  data = dados
)

# modelo 4
m4_iv <- ivreg(
  lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc + momdad14 + sinmom14 |
  nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
  motheduc + momdad14 + sinmom14,
  data = dados
)

# sumário
stargazer::stargazer(
  m1_iv,
  m2_iv,
  m3_iv,
  m4_iv,
  header = FALSE,
  font.size = "scriptsize",
  single.row = TRUE,
  column.sep.width = "1pt",
```

Tabela 6: Regressão de IQ contra nearc4

| | *Dependent variable:* |
|---|---|
| | IQ |
| nearc4 | 2.596*** |
| | (0.745) |
| Constant | 100.611*** |
| | (0.627) |
| Observations | 2,061 |
| $R^2$ | 0.006 |
| Adjusted $R^2$ | 0.005 |
| Residual Std. Error | 15.382 (df = 2059) |
| F Statistic | 12.128*** (df = 1; 2059) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

## i

> Inclua no modelo da letra h as variáveis regionais (smas66 e reg661 – reg668). QI e nearc4 apresentam correlação nessa especificação? Tendo em vista esse achado, o que você conclui sobre a importância de incluir as variáveis regionais de 1966 na equação para o logaritmo do salário?

A inclusão das variáveis regionais torna `nearc4` não significativo, indicando que a correlação entre `IQ` e `nearc4` pode ser explicada por essas variáveis. Isso sugere que a inclusão de variáveis regionais é importante para controlar o efeito da região sobre o QI e o salário, e suportando `nearc4` como um instrumento válido.

```r
# modelo
m_iq_regional <- lm(IQ ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 +
  reg666 + reg667 + reg668 + smsa66, data = dados)

# sumário
stargazer::stargazer(
  m_iq_regional,
  header = FALSE,
  single.row = TRUE,
  title = "Regressão de IQ contra nearc4 e variáveis regionais"
)
```

Tabela 7: Regressão de IQ contra nearc4 e variáveis regionais

| | *Dependent variable:* |
|---|---|
| | IQ |
| nearc4 | 0.348 (0.814) |
| reg661 | 2.892 (1.797) |
| reg662 | $3.991^{***}$ (1.294) |
| reg663 | 1.333 (1.259) |
| reg664 | 2.349 (1.635) |
| reg665 | $-5.584^{***}$ (1.322) |
| reg666 | $-4.529^{***}$ (1.698) |
| reg667 | $-5.502^{***}$ (1.520) |
| reg668 | $-0.033$ (2.111) |
| smsa66 | 1.089 (0.809) |
| Constant | $101.882^{***}$ (1.292) |
| Observations | 2,061 |
| $R^2$ | 0.063 |
| Adjusted $R^2$ | 0.058 |
| Residual Std. Error | 14.969 (df = 2050) |
| F Statistic | $13.700^{***}$ (df = 10; 2050) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

# Q5

Esta questão é motivada pelo artigo "The Persistent Effects of Peru's Mining Mita" de Melissa Dell (2010). A autora busca compreender os efeitos persistentes da Mita, um sistema extrativista de trabalho forçado implementado pelo império espanhol no Peru e na Bolívia. Dell avalia se essa instituição, que foi encerrada há duzentos anos, ainda tem efeitos sobre o consumo e a renda das famílias nos dias de hoje. O artigo está disponível no link; familiarize-se com ele antes de começar. Você precisará usaro arquivo mitaData.csv para realizar o exercício a seguir.

```
dados <- read.csv("lista/data/mitaData.csv") ▷
  janitor::clean_names()
```

## a

Dell utiliza um design de regressão descontínua para comparar os resultados entre pessoas que vivem em distritos que participaram da antiga Mita e aquelas que vivem em distritos que não participaram. Discuta brevemente a estratégia de identificação do efeito causal ressaltando:

### i

> Por que a comparação dos distritos que tiveram o sistema de Mita com os que não tiveram é inadequada?

Para que seja um contrafactual válido, todas as características dos elementos dos grupos de controle devem ser iguais a de tratamento, exceto pela intervenção. A comparação dos distritos que tiveram o sistema de Mita com os que não tiveram é inadequada porque os distritos que tiveram a Mita podem ser diferentes dos que não tiveram em outras características que afetam o consumo e a renda das famílias. Por exemplo, os distritos que tiveram a Mita podem ter sido escolhidos por serem mais ricos ou mais pobres, ou por terem uma população com características diferentes. Portanto, a comparação direta entre os distritos que tiveram a Mita e os que não tiveram pode levar a um viés de seleção.

### ii

Qual a hipótese de identificação usada pela autora?

Na página 1.871 ela cita dois pressupostos: de que os distritos próximos à fronteira da Mita são semelhantes em todas as características, exceto pela participação na Mita. Portanto, a comparação entre os distritos próximos à fronteira da Mita que participaram e os que não participaram pode fornecer uma estimativa causal do efeito da Mita sobre o consumo e a renda das famílias. E que os resultados potenciais sobre tratamento e controle são contínuos ao longo das latitudes e longitudes da fronteira.

### iii

Qual a equação estimada e qual variável corresponde a estratégia usada pela autora?

A equação estimada é:

$$c_{idb} = \alpha + \gamma mita_d + X'_{id}\beta + f(\text{geographic location}_d) + \phi_b + \varepsilon_{idb}$$

Para brevidade, a descrição de cada uma das variáveis está no último parágrafo da página 1.870.

## b

Agora vamos reproduzir o resultado principal de Dell. Dadas as variáveis de longitude e latitude $x$ e $y$, construa $x^2$, $y^2$, $xy$, $x^3$, $y^3$, $x^2y$ e $xy^2$. Faça uma regressão do logaritmo do consumo equivalente dos domicílios (2001) (`lhhequiv`) em relação à variável `Mita` (pothuan_mita), todos os termos polinomiais, elevação (elv_sh), inclinação média (mean_slope), número de bebês (infants), crianças (children), adultos (adults) e efeitos fixos dos segmentos de fronteira (fe4_1, bfe4_2, bfe4_3). Agrupe os erros padrão por distrito. Execute a regressão de 3 maneiras: primeiro, para observações em que a distância até a fronteira da Mita (d_bnd) seja menor que 100 km; depois, quando for menor que 75 km; e, por último, quando for menor que 50 km. Reporte os resultados em uma única tabela e comente sobre os efeitos encontrados, considerando um nível de significância de 5%.

Obs.: Não há variável chamada `mean_slope` no banco de dados. Vou assumir que a variável correta é `slope`. Também não há `fe4_1`, vou supor que é `bfe4_1`.

Na Tabela 1, vemos que os coeficientes de `pothuan_mita` são significativos (5% significância) para os modelos 1 e 3, indicando que a participação na Mita reduz o consumo das famílias, resultado esse consistente com Dell (2010).

```r
# criar termos polinomiais
dados <- dados |>
  transform(
    x2 = lon^2,
    y2 = lat^2,
    xy = lon * lat,
    x3 = lon^3,
    y3 = lat^3,
    x2y = lon^2 * lat,
    xy2 = lon * lat^2
  )

# função para regressão
regressao <- function(distancia) {
  lm(
    lhhequiv ~ pothuan_mita + lon + lat + x2 + y2 + xy + x3 + y3 + x2y + xy2 +
    elv_sh + slope + infants + children + adults + bfe4_1 + bfe4_2 + bfe4_3,
      data = subset(dados, d_bnd < distancia)
  )
}

# regressões
m100 <- regressao(100)
m75 <- regressao(75)
m50 <- regressao(50)

# erros padrão robustos
robust_se_m100 <- sqrt(diag(sandwich::vcovHC(m100, type = "HC1", cluster = ~district)))
robust_se_m75 <- sqrt(diag(sandwich::vcovHC(m75, type = "HC1", cluster = ~district)))
robust_se_m50 <- sqrt(diag(sandwich::vcovHC(m50, type = "HC1", cluster = ~district)))

# sumário
stargazer::stargazer(
  m100, m75, m50,
  se = list(robust_se_m100, robust_se_m75, robust_se_m50),
  header = FALSE,
  single.row = TRUE,
  font.size = "scriptsize",
  title = "Regressões com termos polinomiais",
  label = "tbl-q5.b"
)
```

Tabela 1: Regressões com termos polinomiais

[!htbp]

| | *Dependent variable:* | | |
|---|---|---|---|
| | lhhequiv | | |
| | (1) | (2) | (3) |
| pothuan_mita | −0.284** (0.115) | −0.216* (0.117) | −0.331*** (0.124) |
| lon | 503.348 (998.065) | −4,020.989* (2,241.039) | −2,743.130 (3,244.512) |
| lat | 1,713.701** (835.462) | 202.027 (932.106) | 277.133 (958.834) |
| x2 | −4.155 (12.582) | 53.093* (29.201) | 36.099 (43.509) |
| y2 | −46.972*** (14.425) | −77.970*** (16.196) | −59.514*** (18.963) |
| xy | −28.783 (17.950) | 24.933 (22.767) | 15.725 (23.977) |
| x3 | 0.010 (0.054) | −0.226* (0.127) | −0.152 (0.193) |
| y3 | 0.598*** (0.155) | 0.906*** (0.188) | 0.633*** (0.195) |
| x2y | 0.141 (0.101) | −0.278* (0.152) | −0.197 (0.168) |
| xy2 | 0.293** (0.131) | 0.543*** (0.151) | 0.448** (0.198) |
| elv_sh | 0.068 (0.104) | −0.016 (0.114) | 0.040 (0.112) |
| slope | −0.022** (0.011) | −0.033*** (0.012) | −0.014 (0.013) |
| infants | −0.004 (0.028) | −0.009 (0.030) | −0.036 (0.030) |
| children | 0.014 (0.016) | −0.005 (0.017) | −0.008 (0.018) |
| adults | 0.015 (0.021) | 0.025 (0.024) | 0.011 (0.024) |
| bfe4_1 | 0.878*** (0.289) | 0.933*** (0.317) | 0.840* (0.447) |
| bfe4_2 | 0.602*** (0.228) | 0.902*** (0.233) | 0.621** (0.259) |
| bfe4_3 | 0.091 (0.119) | 0.173 (0.126) | 0.228* (0.138) |
| Constant | −20,226.220 (26,982.300) | 96,107.710* (57,095.820) | 65,328.340 (80,608.450) |
| Observations | 1,478 | 1,161 | 1,013 |
| $R^2$ | 0.059 | 0.060 | 0.069 |
| Adjusted $R^2$ | 0.048 | 0.045 | 0.052 |
| Residual Std. Error | 0.986 (df = 1459) | 0.894 (df = 1142) | 0.832 (df = 994) |
| F Statistic | 5.124*** (df = 18; 1459) | 4.054*** (df = 18; 1142) | 4.115*** (df = 18; 994) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

### c

Execute as mesmas regressões de antes, mas, em vez de termos polinomiais em longitude e latitude, utilize um polinômio cúbico na distância até Potosí (dpot). Ou seja, inclua a primeira, a segunda e a terceira potência dessa variável nas regressões. Novamente, agrupe os erros padrão por distrito e execute a regressão de 3 maneiras: primeiro, para observações em que a distância até a fronteira da Mita (d_bnd) seja menor que 100 km; depois, quando for menor que 75 km; e, por último, quando for menor que 50 km. Reporte os resultados em uma única tabela e comente sobre os efeitos encontrados, considerando um nível de significância de 5%.

Na Tabela 2, vemos que controlando por distância até Potosí, os coeficientes de `pothuan_mita` são significativos (5% significância) para todos os modelos, também consistente com os resultados de Dell (2010).

```r
# criar termos polinomiais de dpot
dados <- dados |>
  transform(
    dpot2 = dpot^2,
    dpot3 = dpot^3
  )

# função para regressão
regressao_dpot <- function(distancia) {
  lm(
    lhhequiv ~ pothuan_mita + dpot + dpot2 + dpot3 +
    elv_sh + slope + infants + children + adults + bfe4_1 + bfe4_2 + bfe4_3,
    data = subset(dados, d_bnd < distancia)
  )
}

# regressões
m100_dpot <- regressao_dpot(100)
m75_dpot <- regressao_dpot(75)
m50_dpot <- regressao_dpot(50)

# erros padrão agrupados por distrito
robust_se_m100_dpot <- sqrt(diag(sandwich::vcovHC(m100_dpot, type = "HC1", cluster = ~district)))
robust_se_m75_dpot <- sqrt(diag(sandwich::vcovHC(m75_dpot, type = "HC1", cluster = ~district)))
robust_se_m50_dpot <- sqrt(diag(sandwich::vcovHC(m50_dpot, type = "HC1", cluster = ~district)))

# sumário
stargazer::stargazer(
```

Tabela 2: Regressões com polinômio cúbico de dpot

[!htbp]

| | *Dependent variable:* | | |
|---|---|---|---|
| | lhhequiv | | |
| | (1) | (2) | (3) |
| pothuan_mita | −0.337*** (0.053) | −0.307*** (0.060) | −0.329*** (0.061) |
| dpot | −2.838 (3.494) | 9.210* (5.458) | 17.330 (13.842) |
| dpot2 | 0.270 (0.414) | −1.023 (0.623) | −2.077 (1.453) |
| dpot3 | −0.008 (0.016) | 0.038 (0.023) | 0.081 (0.051) |
| elv_sh | −0.176** (0.085) | −0.163 (0.108) | −0.173* (0.104) |
| slope | −0.028*** (0.011) | −0.023** (0.011) | −0.011 (0.010) |
| infants | −0.011 (0.029) | −0.019 (0.030) | −0.046 (0.031) |
| children | 0.010 (0.016) | −0.005 (0.017) | −0.012 (0.018) |
| adults | 0.017 (0.021) | 0.020 (0.024) | 0.006 (0.024) |
| bfe4_1 | 0.515*** (0.083) | 0.439*** (0.107) | 0.452*** (0.109) |
| bfe4_2 | −0.071 (0.156) | 0.069 (0.174) | −0.179 (0.231) |
| bfe4_3 | 0.084 (0.092) | 0.115 (0.104) | 0.097 (0.106) |
| Constant | 16.494* (9.616) | −20.764 (15.631) | −40.235 (43.724) |
| Observations | 1,478 | 1,161 | 1,013 |
| $R^2$ | 0.046 | 0.036 | 0.047 |
| Adjusted $R^2$ | 0.039 | 0.026 | 0.036 |
| Residual Std. Error | 0.990 (df = 1465) | 0.903 (df = 1148) | 0.839 (df = 1000) |
| F Statistic | 5.930*** (df = 12; 1465) | 3.565*** (df = 12; 1148) | 4.142*** (df = 12; 1000) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

```
  m100_dpot, m75_dpot, m50_dpot,
  se = list(robust_se_m100_dpot, robust_se_m75_dpot, robust_se_m50_dpot),
  header = FALSE,
  single.row = TRUE,
  font.size = "scriptsize",
  title = "Regressões com polinômio cúbico de dpot",
  label = "tbl-q5.c"
)
```