

# **Avaliação de Políticas Públicas\***

## **Lista de Exercícios**

Alberson da Silva Miranda<sup>†</sup>

30 de dezembro de 2024

\*Código disponível em [https://github.com/albersonmiranda/politicas\\_publicas](https://github.com/albersonmiranda/politicas_publicas).

<sup>†</sup>Discente PPGEco/UFES.

# SUMÁRIO

<b>Q1</b>	<b>3</b>
a . . . . .	3
b . . . . .	5
c . . . . .	7
d . . . . .	12
<b>Q2</b>	<b>15</b>
a . . . . .	15
1 . . . . .	15
2 . . . . .	16
3 . . . . .	16
4 . . . . .	17
b . . . . .	17
c . . . . .	18
1 . . . . .	18
2 . . . . .	20
3 . . . . .	20
4 . . . . .	24
d . . . . .	24
1 . . . . .	24
2 . . . . .	24
3 . . . . .	25
e . . . . .	25
1 . . . . .	27
2 . . . . .	27
3 . . . . .	29
<b>Q3</b>	<b>30</b>
a . . . . .	30
i . . . . .	31
ii . . . . .	31
b . . . . .	31
c . . . . .	35
d . . . . .	37
e . . . . .	38

f . . . . .	44
g . . . . .	45
h . . . . .	47
<b>Q4</b>	<b>48</b>
a . . . . .	48
b . . . . .	49
c . . . . .	51
d . . . . .	51
e . . . . .	52
f . . . . .	54
g . . . . .	58
h . . . . .	58
i . . . . .	58
<b>Q5</b>	<b>61</b>
a . . . . .	61
i . . . . .	61
ii . . . . .	61
iii . . . . .	61
b . . . . .	62
c . . . . .	62
<b>Q6</b>	<b>63</b>
a . . . . .	64
b . . . . .	64
c . . . . .	64
d . . . . .	64
e . . . . .	64
f . . . . .	65

# 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.

```

1 # pacotes
2 library(tableone)
3
4 # carregar dados
5 dados <- readxl::read_excel("lista/data/PIAA_2017-2018.xlsx") >
6 janitor::clean_names()
7
8 # feature engineering
9 dados <- within(dados, {
10   tratamento <- ifelse(presenca > 4, 1, 0)
11   semestre <- as.factor(semestre)
12   periodo <- as.factor(periodo)
13   sexo <- as.factor(sexo)
14   primeiro_periodo <- as.factor(ifelse(periodo == 1, 1, 0))
15 })
16
17 # tabela
18 tableone <- CreateTableOne(
19   vars = c("faltas", "sexo", "semestre", "primeiro_periodo"),
20   strata = "tratamento",
21   data = dados
22 )
23 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.

```
1 # modelos
2 m1 <- lm(nota ~ tratamento, data = dados)
3 m2 <- lm(
4   nota ~ tratamento + faltas + sexo + semestre + primeiro_periodo,
5   data = dados
6 )
7
8 # matriz de covariância corrigida
9 cov_m1 <- sandwich::vcovHC(m1, type = "HC")
10 cov_m2 <- sandwich::vcovHC(m2, type = "HC")
11
12 # erros padrões robustos
13 robust_se_m1 <- sqrt(diag(cov_m1))
14 robust_se_m2 <- sqrt(diag(cov_m2))
15
16 # sumário
```

```

17 stargazer::stargazer(
18   m1,
19   m2,
20   single.row = TRUE,
21   se = list(robust_se_m1, robust_se_m2),
22   title = "Comparação de modelos",
23   header = FALSE
24 )

```

Tabela 2: Comparação de modelos

	<i>Dependent variable:</i>	
	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 <sup>2</sup>	0.139	0.520
Adjusted R <sup>2</sup>	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 claramente no balanceamento, com  $eCDF > 5\%$  para faltas, dois dos semestres e a distância (probabilidade predita de pertencer ao grupo das tratadas).

```
1 library(MatchIt)
2
3 m_exato <- matchit(
4   tratamento ~ faltas + sexo + semestre + primeiro_periodo,
5   data = dados,
6   method = "exact"
7 )
8
9 m_mahalanobis <- matchit(
10  tratamento ~ faltas + sexo + semestre + primeiro_periodo,
11  data = dados,
12  method = "nearest",
13  distance = "mahalanobis"
14 )
15
16 m_logit <- matchit(
17  tratamento ~ faltas + sexo + semestre + primeiro_periodo,
18  data = dados,
19  method = "nearest",
20  distance = "logit"
21 )
22
23 # balanceamento
24 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_perodo0	0	0	0
primeiro_perodo1	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_perodo,
        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_perodo0	0.0000	0.1149	-0.4114	.
primeiro_perodo1	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_perodo0	0.1149	0.1149
primeiro_perodo1	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_peri0do0	0.0000	0.0000	0.000	.
primeiro_peri0do1	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_peri0do0	0.0000	0.0000	0.0000
primeiro_peri0do1	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_peri0do,
  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_peri0do0	0.0000	0.1149	-0.4114	.
primeiro_peri0do1	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_peri0do0	0.1149	0.1149
primeiro_peri0do1	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_peri0do0	0.0000	0.0000	0.0000	.
primeiro_peri0do1	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_peri0do0	0.0000	0.0000	0.0000
primeiro_peri0do1	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).

```
1 # modelos
2 m_exato_bal <- lm(
3   nota ~ tratamento,
4   data = match.data(m_exato),
5   weights = weights
6 )
7 m_mahalanobis_bal <- lm(
8   nota ~ tratamento,
9   data = match.data(m_mahalanobis),
10  weights = weights
11 )
12 m_logit_bal <- lm(
13   nota ~ tratamento,
14   data = match.data(m_logit),
15   weights = weights
16 )
17
18 # matriz de covariância corrigida
19 cov_m_exato_bal <- sandwich::vcovHC(m_exato_bal, type = "HC")
20 cov_m_mahalanobis_bal <- sandwich::vcovHC(m_mahalanobis_bal, type = "HC")
```

```

21 cov_m_logit_bal <- sandwich::vcovHC(m_logit_bal, type = "HC")
22
23 # erros padrões robustos
24 robust_se_m_exato_bal <- sqrt(diag(cov_m_exato_bal))
25 robust_se_m_mahalanobis_bal <- sqrt(diag(cov_m_mahalanobis_bal))
26 robust_se_m_logit_bal <- sqrt(diag(cov_m_logit_bal))
27
28 # sumário
29 stargazer::stargazer(
30   m1,
31   m2,
32   m_exato_bal,
33   m_mahalanobis_bal,
34   m_logit_bal,
35   single.row = FALSE,
36   se = list(
37     robust_se_m1,
38     robust_se_m2,
39     robust_se_m_exato_bal,
40     robust_se_m_mahalanobis_bal,
41     robust_se_m_logit_bal
42   ),
43   title = "Comparação de modelos",
44   header = FALSE,
45   font.size = "scriptsize",
46   column.sep.width = "3pt"
47 )

```

Tabela 3: Comparação de modelos

	<i>Dependent variable:</i>				
	nota				
	(1)	(2)	(3)	(4)	(5)
tratamento	2.602*** (0.491)	1.271*** (0.408)	0.874 (0.709)	0.924 (0.610)	0.818 (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*** (0.220)	4.339*** (0.664)	4.084*** (0.521)	3.702*** (0.424)	3.809*** (0.444)
Observations	193	193	82	90	90
R <sup>2</sup>	0.139	0.520	0.022	0.025	0.019
Adjusted R <sup>2</sup>	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&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;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-mínimo 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.

```
1 # carregar dados
2 dados <- readxl::read_excel("lista/data/minwage.xlsx")
3
4 # filtrando observações
5 dados <- subset(dados, sample == 1)
```



```

6
7 # salário médio
8 salario_medio <- aggregate(
9   cbind(wage_st, wage_st2) ~ state,
10  data = dados,
11  FUN = mean
12 )
13
14 # diferença
15 salario_medio <- transform(salario_medio, diff = wage_st2 - wage_st)
16
17 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.

```

1 diff_in_diff <- diff(salario_medio$diff)
2
3 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 paralelas: 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.

```
1 # emprego médio
2 emprego_medio <- aggregate(
3   cbind(fte, fte2) ~ state,
4   data = dados,
5   FUN = mean
6 )
7
8 # diferença
9 emprego_medio <- transform(emprego_medio, diff = fte2 - fte)
10
11 # diferença em diferenças
12 diff_in_diff_emprego <- diff(emprego_medio$diff)
13
14 emprego_medio
```

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

```
1 diff_in_diff_emprego
```

[1] 2.301994

## c

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

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

onde  $Y_{ist}$  representa emprego no restaurante  $i$ , no estado  $s$  e período  $t$ ,  $\text{TREAT}_{is}$  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, copiando 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.

```
1 # pacotes
2 suppressPackageStartupMessages({
3   library(tidyr)
4   library(dplyr)
5 })
6
7 # alongando a tabela para tratar a mudança no tempo como
8 # linha ao invés de coluna
9 colunas_repetidas <- setdiff(
```

```

10 names(dados),
11 grep("2$", names(dados), value = TRUE)
12 )
13
14 dados_long <- pivot_longer(
15   data = dados,
16   cols = matches(".*2$|.*[^2]$"),
17   names_to = c(".value", "post"),
18   names_pattern = "(.*?)(2?)$"
19 )
20 mutate(
21   post = factor(ifelse(post == "2", "nov_dez", "mar"), levels = c("mar", "nov_dez")),
22   state = factor(ifelse(state == "1", "NJ", "PA"), levels = c("PA", "NJ"))
23 )
24 fill(all_of(colunas_repetidas), .direction = "down")
25
26 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.

```
1 # regressão
2 m_salario <- lm(wage_st ~ state * post, data = dados_long)
3 m_emplogo <- lm(fte ~ state * post, data = dados_long)
4
5 # sumário
6 stargazer::stargazer(
7   m_salario,
8   m_emplogo,
9   title = "Regressão de diferenças em diferenças",
10  header = FALSE
11 )
```

## 2

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

```
1 # regressão
2 m_salario_2 <- lm(wage_st ~ state * post + co_owned + chain, data = dados_long)
3 m_emplogo_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.

```
1 # sumário
2 stargazer::stargazer(
3   m_salario,
4   m_salario_2,
5   title = "Regressão de diferenças em diferenças",
6   header = FALSE
7 )
```

Tabela 1: Regressão de diferenças em diferenças

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

Tabela 2: Regressão de diferenças em diferenças

	<i>Dependent variable:</i>	
	wage_st	
	(1)	(2)
stateNJ	−0.041 (0.038)	−0.039 (0.037)
postnov_dez	−0.035 (0.048)	−0.035 (0.047)
co_owned		0.064*** (0.022)
chain		0.034*** (0.010)
stateNJ:postnov_dez	0.504*** (0.054)	0.504*** (0.053)
Constant	4.654*** (0.034)	4.559*** (0.039)
Observations	702	702
R <sup>2</sup>	0.403	0.425
Adjusted R <sup>2</sup>	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)
<i>Note:</i>		*p<0.1; **p<0.05; ***p<0.01

```

1 stargazer::stargazer(
2   m_emprego,
3   m_emprego_2,
4   title = "Regressão de diferenças em diferenças",
5   header = FALSE
6 )

```

Tabela 3: Regressão de diferenças em diferenças

	<i>Dependent variable:</i>	
	fte	
	(1)	(2)
stateNJ	−2.838** (1.225)	−2.871** (1.216)
postnov_dez	−2.015 (1.561)	−2.015 (1.550)
co_owned		−2.243*** (0.728)
chain		−0.243 (0.322)
stateNJ:postnov_dez	2.302 (1.732)	2.302 (1.720)
Constant	20.114*** (1.104)	21.440*** (1.280)
Observations	702	702
R <sup>2</sup>	0.008	0.025
Adjusted R <sup>2</sup>	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.

```
1 # filtrando observações
2 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.

```
1 # variável de baixo salário
2 dados_nj <- by(dados_nj, dados_nj$sheet, function(conjunto) {
3   conjunto$low_wage <- ifelse(any(conjunto$wage_st < 5 & conjunto$post == "mar"), 1, 0)
4   conjunto
5 }) ▷
6 do.call(what = rbind)
7
8 # regressão
9 m_salario_nj <- lm(wage_st ~ low_wage * post, data = dados_nj)
10 m_emplo_nj <- lm(fte ~ low_wage * post, data = dados_nj)
11
12 # sumário
13 stargazer::stargazer(
14   m_salario_nj,
15   m_emplo_nj,
16   title = "Regressão de diferenças em diferenças",
17   header = FALSE
18 )
```

### 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.

```
1 # filtrando observações
2 dados_pa <- subset(dados_long, state == "PA")
3
4 # variável de baixo salário
5 dados_pa <- by(dados_pa, dados_pa$sheet, function(conjunto) {
```

Tabela 4: Regressão de diferenças em diferenças

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

```

6  conjunto$low_wage <- ifelse(
7  any(conjunto$wage_st < 5 & conjunto$post == "mar"),
8  1,
9  0
10 )
11 conjunto
12 }) ▷
13 do.call(what = rbind)
14
15 # regressão
16 m_salario_pa <- lm(wage_st ~ low_wage * post, data = dados_pa)
17 m_emplo_pa <- lm(fte ~ low_wage * post, data = dados_pa)
18
19 # sumário
20 stargazer::stargazer(
21   m_salario_pa,
22   m_emplo_pa,
23   title = "Regressão de diferenças em diferenças",
24   header = FALSE
25 )

```

## 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.

```

1  # coeficientes e erro padrão
2  beta_1 <- coef(m_salario_nj)["low_wage:postnov_dez"]
3  beta_2 <- coef(m_salario_pa)["low_wage:postnov_dez"]
4

```

Tabela 5: Regressão de diferenças em diferenças

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

```

5 se_1 <- summary(m_salario_nj)$coefficients["low_wage:postnov_dez", "Std. Error"]
6 se_2 <- summary(m_salario_pa)$coefficients["low_wage:postnov_dez", "Std. Error"]
7
8 # estatística de teste
9 z <- (beta_1 - beta_2) / sqrt(se_1^2 + se_2^2)
10
11 # p-valor
12 p_valor <- 2 * (1 - pnorm(abs(z)))
13
14 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.

```
1 dados <- readxl::read_excel("lista/data/Guns.xlsx") ▶
2   transform(
3     shall = as.factor(shall),
4     stateid = as.factor(stateid)
5   )
```

### a

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

```
1 # regressões
2 m1 <- lm(log(vio) ~ shall, data = dados)
3 m2 <- lm(
4   log(vio) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
5   data = dados
6 )
7
8 # sumário
9 stargazer::stargazer(
10   m1,
11   m2,
```

```
12 title = "Regressões",
13 header = FALSE
14 )
```

## 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  $e^{-0.443} = 0.6421$ , 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

	<i>Dependent variable:</i>	
	log(vio)	
	(1)	(2)
shall1	-0.443*** (0.042)	-0.368*** (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*** (0.021)	2.982*** (0.543)
Observations	1,173	1,173
R <sup>2</sup>	0.087	0.564
Adjusted R <sup>2</sup>	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&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;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.

$$y_{it} = x'_{it}\beta + \alpha_i + \theta_t + \varepsilon_{it} \quad (0.1)$$

```
1 # pacote para dados em painel
2 suppressPackageStartupMessages(library(plm))
3
4 # organização dos dados em painel
5 painel <- pdata.frame(dados, index = c("stateid", "year"))
6
7 # modelo
8 m3 <- plm(
9   log(vio) ~ shall,
10  data = painel,
11  model = "within",
12  effect = "twoways"
13 )
14
15 m4 <- plm(
16   log(vio) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,
17  data = painel,
18  model = "within",
19  effect = "twoways"
20 )
21
22 # sumário
23 stargazer::stargazer(
24   m3,
25   m4,
26   title = "Regressões em painel",
27   header = FALSE
28 )
```

Tabela 2: Regressões em painel

	<i>Dependent variable:</i>	
	log(vio)	
	(1)	(2)
shall1	0.002 (0.017)	−0.028 (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 <sup>2</sup>	0.00001	0.056
Adjusted R <sup>2</sup>	−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.

```
1 # ano de início do tratamento
2 anos_tratamento <- aggregate(
3   year ~ stateid,
4   data = subset(dados, shall == 1),
5   FUN = min
6 )
7
8 # merge com dataset
9 dados <- merge(dados, anos_tratamento, by = "stateid", all.x = TRUE) >
10   \(dadinho) sort_by(dadinho, dadinho$stateid, dadinho$year.x))()
11
12 # renomear variáveis
13 names(dados)[names(dados) %in% c("year.x", "year.y")] <- c("year", "year_tratamento")
14
15 # mais data engineering!
16 dados <- within(dados, {
17   # anos relativos ao tratamento
18   year_rel <- year - year_tratamento
19   # indicador de pré ou pós tratamento
20   tratamento <- ifelse(year_rel ≥ 0, "post", "pre")
21   # indicador de controle
22   tratamento <- ifelse(is.na(tratamento), "controle", tratamento)
23 })
24
25 # plotar quantidade de `vio` por ano relativo ao tratamento para ambos grupos
26 library(ggplot2)
27 dados >
```

```

28 subset(tratamento != "controle") >
29 ggplot(aes(x = year_rel, y = vio, color = tratamento)) +
30 geom_point() +
31 geom_smooth(method = "lm", se = FALSE, formula = y ~ poly(x, 3)) +
32 facet_wrap(~ stateid, scales = "free") +
33 labs(
34   x = "Ano relativo ao tratamento",
35   y = "Violência"
36 ) +
37 # removendo elementos para melhor visualização
38 theme(
39   # legenda
40   legend.position = "none",
41   # removendo números de eixos
42   axis.text.x = element_blank(),
43   axis.text.y = element_blank(),
44   # reduzindo strip text e largura
45   strip.text = element_text(size = 6),
46   strip.background = element_blank()
47 )

```

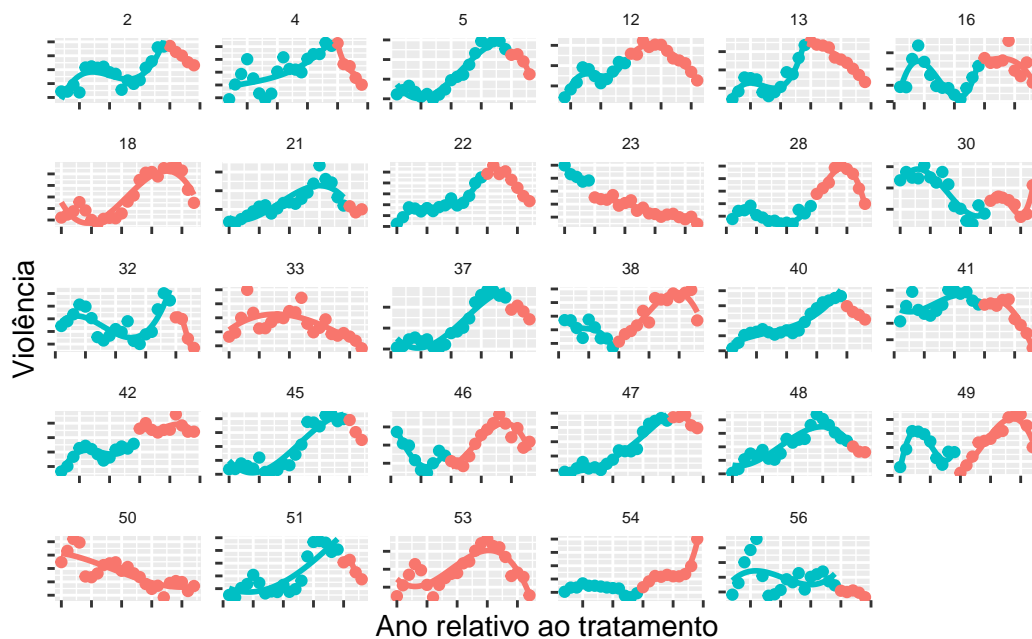


Figura 1: Violência por ano relativo ao tratamento

```

1 # plotar quantidade de `vio` por ano, considerando apenas período sem tratamento
2 dados >
3 #subset(tratamento != "post") >
4 ggplot(aes(x = year, y = vio, color = tratamento)) +
5 geom_point() +
6 geom_smooth(method = "lm", se = FALSE) +
7 labs(
8   title = "Violência por ano",
9   x = "Ano",
10  y = "Violência"
11 )

```

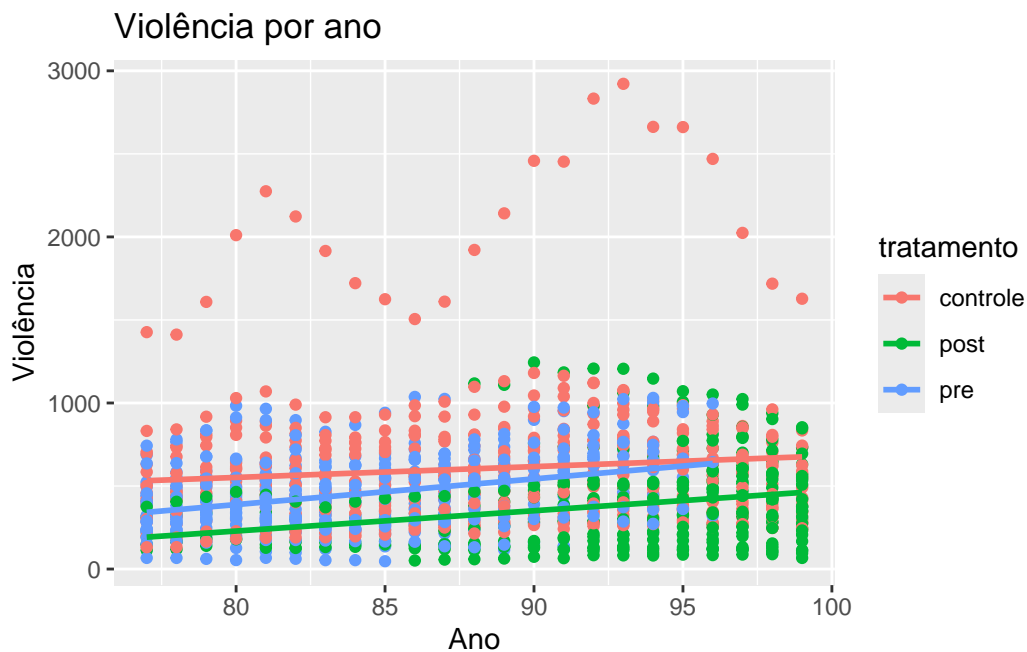


Figura 2: Violência por ano, grupos tratado e controle

**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.

```

1 # quantidade de grupos de tratados
2 grupos_unicos <- unique(dados$year[dados$shall == 1])
3 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
```

```
1 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.

```

1 # mais features
2 dados <- within(dados, {
3   G <- ifelse(is.na(year_tratamento), 0, year_tratamento)
4   D <- ifelse(year >= year_tratamento & !is.na(year_tratamento), 1, 0)
5   stateid_n <- as.numeric(stateid)
6 })
7
8 # Estimar os efeitos médios do programa para grupo-período
9 efeitos_medios <- did::att_gt(
10   yname = "vio",
11   tname = "year",
12   idname = "stateid_n",
13   gname = "G",
14   xformula = ~ incarc_rate + pb1064 + pw1064 + avginc + density,
15   data = dados,
16   panel = TRUE
17 )
18
19 # sumário
20 summary(efeitos_medios)

```

Call:

```
did::att_gt(yname = "vio", tname = "year", idname = "stateid_n",  
  gname = "G", xformula = ~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." *Journal of Econometrics* 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
82	79	NA	NA	NA
82	80	NA	NA	NA
82	81	NA	NA	NA
82	82	NA	NA	NA
82	83	NA	NA	NA
82	84	NA	NA	NA
82	85	NA	NA	NA
82	86	NA	NA	NA
82	87	NA	NA	NA
82	88	NA	NA	NA
82	89	NA	NA	NA
82	90	NA	NA	NA
82	91	NA	NA	NA
82	92	NA	NA	NA
82	93	NA	NA	NA
82	94	NA	NA	NA
82	95	NA	NA	NA
82	96	NA	NA	NA
82	97	NA	NA	NA
82	98	NA	NA	NA
82	99	NA	NA	NA
86	78	NA	NA	NA
86	79	NA	NA	NA
86	80	NA	NA	NA
86	81	NA	NA	NA
86	82	NA	NA	NA
86	83	NA	NA	NA
86	84	NA	NA	NA
86	85	NA	NA	NA
86	86	NA	NA	NA
86	87	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	14.7117	-18.8115	52.4342
90	79	2.2446	14.6607	-33.2549	37.7441
90	80	-16.0294	10.8760	-42.3646	10.3058
90	81	-29.3239	17.7960	-72.4152	13.7674
90	82	-33.4837	28.4356	-102.3378	35.3703
90	83	17.5104	9.9770	-6.6479	41.6687
90	84	-4.3933	9.5550	-27.5298	18.7431
90	85	-3.1396	7.7610	-21.9321	15.6529
90	86	-9.0373	10.4932	-34.4457	16.3711
90	87	30.9261	13.8815	-2.6867	64.5389
90	88	9.9978	31.2613	-65.6984	85.6941
90	89	11.1119	9.3056	-11.4207	33.6444
90	90	-30.9216	20.6303	-80.8758	19.0327
90	91	-64.4779	48.9812	-183.0811	54.1253
90	92	-106.4677	66.8201	-268.2661	55.3307
90	93	-108.3028	56.1489	-244.2618	27.6562
90	94	-88.5237	54.2067	-219.7798	42.7325
90	95	-59.2638	30.6966	-133.5926	15.0650
90	96	-32.8639	66.8169	-194.6544	128.9266
90	97	-56.1566	59.1070	-199.2786	86.9653
90	98	-88.0526	92.3991	-311.7879	135.6828
90	99	-3.0983	67.3418	-166.1598	159.9633
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	20.7273	-86.2904	14.0877
91	90	-46.5126	13.5887	-79.4162	-13.6089 *
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	21.4496	-23.7904	80.0857
92	79	-0.3729	48.4951	-117.7990	117.0532
92	80	-40.9712	10.7737	-67.0586	-14.8839 *
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	21.4594	-47.6879	56.2357
95	79	-18.2749	21.1901	-69.5847	33.0348
95	80	-10.0527	33.7807	-91.8495	71.7441
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	24.4202	-77.7087	40.5537
95	87	11.9033	28.2681	-56.5452	80.3517
95	88	5.6073	19.0268	-40.4644	51.6790
95	89	-60.1374	17.4061	-102.2847	-17.9901 *
95	90	-20.3368	26.3532	-84.1486	43.4750
95	91	-30.9435	36.4031	-119.0900	57.2031
95	92	3.9848	18.0607	-39.7474	47.7170
95	93	63.0494	34.7919	-21.1957	147.2945
95	94	52.7887	13.3217	20.5314	85.0460 *
95	95	61.8424	15.8928	23.3595	100.3253 *
95	96	75.6614	17.1373	34.1650	117.1578 *
95	97	94.9832	24.2411	36.2858	153.6806 *
95	98	54.3264	32.7793	-25.0454	133.6983
95	99	71.2528	30.3937	-2.3426	144.8482
96	78	NA	NA	NA	NA
96	79	NA	NA	NA	NA
96	80	-14.0876	24.5798	-73.6051	45.4300
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	15.3009	-17.6442	56.4550
96	89	-36.3295	55.7194	-171.2487	98.5896
96	90	-60.0790	28.7767	-129.7590	9.6009
96	91	-67.2576	22.1866	-120.9803	-13.5349 *
96	92	-5.5084	18.0293	-49.1647	38.1479
96	93	104.8199	43.0225	0.6450	208.9947 *
96	94	105.6578	32.7037	26.4691	184.8465 *
96	95	33.8210	15.5563	-3.8472	71.4891
96	96	-2.5534	15.8725	-40.9872	35.8803
96	97	2.9448	20.8384	-47.5133	53.4028
96	98	-12.8283	38.8334	-106.8595	81.2029
96	99	-32.2859	49.5374	-152.2359	87.6640
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	36.3187	-139.2682	36.6162
97	92	NA	NA	NA	NA
97	93	27.9079	64.4760	-128.2144	184.0302
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.

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

```
[1] -6.560892
```

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

```
shall1
-0.4429646
```

```
1 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.

```
1 # modelo  
2 m5 <- plm(  
3   log(rob) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,  
4   data = painel,  
5   model = "within",  
6   effect = "twoways"  
7 )  
8  
9 m6 <- plm(  
10  log(mur) ~ shall + incarc_rate + density + avginc + pop + pb1064 + pw1064 + pm1029,  
11  data = painel,  
12  model = "within",  
13  effect = "twoways"  
14 )  
15  
16 # sumário  
17 stargazer::stargazer(  
18   m5,  
19   m6,  
20   title = "Regressões em painel",  
21   header = FALSE  
22 )
```

Tabela 3: Regressões em painel

	<i>Dependent variable:</i>	
	log(rob)	log(mur)
	(1)	(2)
shall1	0.027 (0.024)	-0.015 (0.025)
incarc_rate	0.00003 (0.0001)	-0.0001 (0.0001)
density	-0.045 (0.105)	-0.544*** (0.110)
avginc	0.014 (0.009)	0.057*** (0.009)
pop	0.00002 (0.011)	-0.032*** (0.011)
pb1064	0.014 (0.031)	0.022 (0.033)
pw1064	-0.013 (0.011)	-0.0005 (0.011)
pm1029	0.105*** (0.022)	0.069*** (0.023)
Observations	1,173	1,173
R <sup>2</sup>	0.049	0.116
Adjusted R <sup>2</sup>	-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.

```
1 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 fatheduc e motheduc. Por fim, adicione ao terceiro modelo as variáveis momdad14 e sinmom14. 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 muito próximos entre si, indicando que a inclusão de variáveis adicionais não afetou a estimativa de educ.

```
1 # modelo 1
2 m1 <- lm(lwage ~ educ + exper + expersq + black + south + smsa, data = dados)
3
4 # modelo 2
5 m2 <- lm(
6   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
7   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
8   data = dados
9 )
10
11 # modelo 3
12 m3 <- lm(
13   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
14   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
15   motheduc,
16   data = dados
17 )
18
19 # modelo 4
20 m4 <- lm(
21   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
22   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
23   motheduc + momdad14 + sinmom14,
24   data = dados
25 )
```

```

26
27 # sumário
28 stargazer::stargazer(
29   m1,
30   m2,
31   m3,
32   m4,
33   title = "Regressões",
34   header = FALSE,
35   font.size = "scriptsize",
36   single.row = TRUE,
37   column.sep.width = "1pt"
38 )

```

Tabela 1: Regressões

	<i>Dependent variable:</i>			
	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 <sup>2</sup>	0.291	0.300	0.274	0.276
Adjusted R <sup>2</sup>	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.

```
1 # modelo 1
2 m1_reduzido <- lm(lwage ~ nearc4, data = dados)
3
4 # modelo 2
5 m2_reduzido <- lm(
6   lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
7   reg667 + reg668 + smsa66,
```

```

8   data = dados
9   )
10
11  # modelo 3
12  m3_reduzido <- lm(
13    lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
14    reg667 + reg668 + smsa66 + fatheduc + motheduc,
15    data = dados
16  )
17
18  # modelo 4
19  m4_reduzido <- lm(
20    lwage ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 + reg666 +
21    reg667 + reg668 + smsa66 + fatheduc + motheduc + momdad14 + sinmom14,
22    data = dados
23  )
24
25  # sumário
26  stargazer::stargazer(
27    m1_reduzido,
28    m2_reduzido,
29    m3_reduzido,
30    m4_reduzido,
31    header = FALSE,
32    font.size = "scriptsize",
33    single.row = TRUE,
34    column.sep.width = "1pt"
35  )

```

**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: Coeficientes da forma reduzida

	<i>Dependent variable:</i>			
	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 <sup>2</sup>	0.027	0.105	0.099	0.105
Adjusted R <sup>2</sup>	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.

```

1 # modelo 1
2 m1_estagio1 <- lm(
3   educ ~ nearc4 + exper + expersq + black + south + smsa,
4   data = dados
5 )
6
7 # modelo 2
8 m2_estagio1 <- lm(
9   educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
10  reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
11  data = dados
12 )
13
14 # modelo 3
15 m3_estagio1 <- lm(
16   educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +

```

```

17   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
18   motheduc,
19   data = dados
20 )
21
22 # modelo 4
23 m4_estagio1 <- lm(
24   educ ~ nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
25   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
26   motheduc + momdad14 + sinmom14,
27   data = dados
28 )
29
30 # sumário
31 stargazer::stargazer(
32   m1_estagio1,
33   m2_estagio1,
34   m3_estagio1,
35   m4_estagio1,
36   header = FALSE,
37   font.size = "scriptsize",
38   single.row = TRUE,
39   column.sep.width = "1pt"
40 )

```

## 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 quase 800 observações a menos, sendo mais difícil detectar um efeito. Além disso, a maioria

Tabela 3: Coeficientes do primeiro estágio

	<i>Dependent variable:</i>			
	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 <sup>2</sup>	0.474	0.477	0.486	0.490
Adjusted R <sup>2</sup>	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&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01



das variáveis incluídas nos modelos 3 e 4 não são significativas, resultando na adição de ruído trazido por essas covariáveis.

```
1 # pacote
2 library(AER)
3
4 # modelo 1
5 m1_iv <- ivreg(
6   lwage ~ educ + exper + expersq + black + south + smsa |
7   nearc4 + exper + expersq + black + south + smsa,
8   data = dados
9 )
10
11 # modelo 2
12 m2_iv <- ivreg(
13   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
14   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 |
15   nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
16   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66,
17   data = dados
18 )
19
20 # modelo 3
21 m3_iv <- ivreg(
22   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
23   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
24   motheduc |
25   nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
26   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
27   motheduc,
28   data = dados
29 )
30
31 # modelo 4
32 m4_iv <- ivreg(
33   lwage ~ educ + exper + expersq + black + south + smsa + reg661 + reg662 +
34   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
35   motheduc + momdad14 + sinmom14 |
36   nearc4 + exper + expersq + black + south + smsa + reg661 + reg662 +
37   reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + smsa66 + fatheduc +
38   motheduc + momdad14 + sinmom14,
39   data = dados
40 )
```

```

41
42 # sumário
43 stargazer::stargazer(
44   m1_iv,
45   m2_iv,
46   m3_iv,
47   m4_iv,
48   header = FALSE,
49   font.size = "scriptsize",
50   single.row = TRUE,
51   column.sep.width = "1pt"
52 )

```

Tabela 4: Coeficientes com nearc4 como instrumento

	<i>Dependent variable:</i>			
	lwage			
	(1)	(2)	(3)	(4)
educ	0.132*** (0.049)	0.132** (0.055)	0.084 (0.076)	0.078 (0.080)
exper	0.107*** (0.021)	0.108*** (0.024)	0.093*** (0.030)	0.091*** (0.031)
expersq	−0.002*** (0.0003)	−0.002*** (0.0003)	−0.002*** (0.0004)	−0.002*** (0.0004)
black	−0.131** (0.053)	−0.147*** (0.054)	−0.162*** (0.035)	−0.161*** (0.034)
south	−0.105*** (0.023)	−0.145*** (0.027)	−0.123*** (0.032)	−0.123*** (0.031)
smsa	0.131*** (0.030)	0.112*** (0.032)	0.134*** (0.042)	0.135*** (0.042)
reg661		−0.108*** (0.042)	−0.088 (0.054)	−0.092* (0.056)
reg662		−0.007 (0.033)	0.0001 (0.039)	−0.001 (0.040)
reg663		0.040 (0.032)	0.033 (0.042)	0.031 (0.044)
reg664		−0.058 (0.038)	−0.052 (0.042)	−0.055 (0.042)
reg665		0.038 (0.047)	0.007 (0.048)	0.003 (0.049)
reg666		0.055 (0.053)	0.021 (0.061)	0.017 (0.063)
reg667		0.027 (0.049)	0.012 (0.051)	0.010 (0.051)
reg668		−0.191*** (0.051)	−0.159*** (0.052)	−0.158*** (0.052)
smsa66		0.019 (0.022)	0.026 (0.025)	0.026 (0.025)
fatheduc			−0.002 (0.009)	−0.001 (0.009)
motheduc			0.007 (0.011)	0.008 (0.011)
momdad14				0.068 (0.076)
sinmom14				0.121 (0.123)
Constant	3.753*** (0.829)	3.774*** (0.935)	4.481*** (1.077)	4.495*** (1.067)
Observations	3,010	3,010	2,220	2,220
R <sup>2</sup>	0.225	0.238	0.273	0.275
Adjusted R <sup>2</sup>	0.224	0.234	0.267	0.269
Residual Std. Error	0.391 (df = 3003)	0.388 (df = 2994)	0.376 (df = 2202)	0.376 (df = 2200)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 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 disso, 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`.

```
1 # modelo
2 m_iq <- lm(IQ ~ nearc4, data = dados)
3
4 # sumário
5 stargazer::stargazer(
6   m_iq,
7   header = FALSE
8 )
```

## 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?

Tabela 5: Regressão de IQ contra nearc4

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

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.

```

1 # modelo
2 m_iq_regional <- lm(IQ ~ nearc4 + reg661 + reg662 + reg663 + reg664 + reg665 +
3   reg666 + reg667 + reg668 + smsa66, data = dados)
4
5 # sumário
6 stargazer::stargazer(
7   m_iq_regional,
8   header = FALSE,
9   single.row = TRUE
10 )

```

Tabela 6: Regressão de IQ contra nearc4 com variáveis regionais

	<i>Dependent variable:</i>
	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 <sup>2</sup>	0.063
Adjusted R <sup>2</sup>	0.058
Residual Std. Error	14.969 (df = 2050)
F Statistic	13.700*** (df = 10; 2050)
<i>Note:</i>	*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 usar o arquivo `mitaData.csv` para realizar o exercício a seguir.

**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?

**ii**

Qual a hipótese de identificação usada pela autora?

**iii**

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

## **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%.

## **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%.

## Q6

Essa questão replica a análise contida no artigo “Evaluating the Effect of Homicide Prevention Strategies in São Paulo, Brazil: A Synthetic Control Approach” de Danilo Freire (2010). O autor busca mensurar o efeito causal agregado de políticas públicas sobre as taxas de homicídios em São Paulo. O artigo está disponível no link; familiarize-se com ele antes de começar. Você precisará usar o arquivo `df.Rdata` para realizar o exercício a seguir. Descrição da base: O painel de dados contém informações de 1990 a 2009, para os 26 estados e o Distrito Federal. Os seguintes estados constam na base, mas não são utilizados na análise: Bahia e Sergipe.

- Formato: dataframe - painel.
- 1 unidade tratada: nome do estado: São Paulo.
- 24 regiões de controle.
- 20 anos (1990:2009).
- Identificadores:
  - abbreviation: sigla do estado
  - code: código do estado
  - state: nome do estado
  - year: ano
- Variável dependente: `homicide.rates` (número de homicídios por 100.000 habitantes)
- Preditores: 1990:1998:
  - `homicide.rates`: taxa de homicídios (número de homicídios por 100.000 habitantes);
  - `state.gdp.capita`: PIB per capita estadual (em reais de 2010);
  - `state.gdp.growth.percent`: crescimento do PIB estadual (percentual, em reais de 2010);
  - `population.projection.ln`: logaritmo natural da população (projeções anuais da população do estado);
  - `proportion.extreme.poverty`: população vivendo em extrema pobreza (percentual da população do estado que não atinge a ingestão mínima de 2.000 calorias por dia);
  - `years.schooling.imp`: anos de escolaridade (número médio de anos de instrução formal em estabelecimentos de ensino-homens e mulheres com 25 anos ou mais). As observações ausentes foram imputadas com interpolação linear;
  - `gini.imp`: índice de Gini (medida de desigualdade que varia de 0 a 1, onde 0 é o mais igual e 1 o mais desigual). As observações ausentes foram imputadas com interpolação linear.



- Para organizar os dados, use os seguintes parâmetros na função `dataprep()`:
  - `predictors.op = "mean";`
  - `time.predictors.prior = c(1990:1998);`
  - `time.optimize.ssr = c(1990:1998);`
  - `time.plot = c(1990:2009);`

## **a**

Discuta brevemente a estratégia de identificação empregada pelo autor ressaltando suas vantagens em relação a estudos de caso comparativos qualitativos. Indique também uma crítica a abordagem seguida pelo autor.

## **b**

Obtenha os pesos para as covariadas (preditores) e para os estados doadores para construção do São Paulo sintético. Organize os resultados como na tabela 1 do artigo.

## **c**

Elabore uma tabela que compara os valores dos preditores pré-tratamento para a unidade tratada, a unidade sintética de controle e todas as unidades na amostra (Tabela 2, no artigo). Comente sobre a semelhança entre SP real e seu controle em cada uma das dimensões.

## **d**

Elabore dois gráficos ilustrativos dos resultados: um mostrando a evolução da trajetória da taxa de homicídios em SP e em seu controle (figura 2 do artigo) e outro mostrando o gap entre a taxa de homicídios em SP e seu controle (figura 3 do artigo) e comente brevemente os resultados.

## **e**

Escolha o estado não tratado com maior peso no controle sintético e faça um teste de placebo tentando encontrar algum efeito da política sobre esse estado. Comente seu resultado ressaltando sua implicação para a análise.

## **f**

Aumente essa análise para todos os demais estados e coloque os resultados em um gráfico ilustrativo (figura 6 do artigo) fazendo o teste de permutação sugerido por Abadie Diamond e Hainmueller (2010). Comente brevemente os resultados.