

Propensity Score Matching

Introdução

Na ciência de dados, um dos grandes desafios é entender a **relação causal** entre variáveis. Quando não podemos realizar experimentos aleatórios por razões éticas, financeiras ou logísticas, os pesquisadores frequentemente usam **estudos observacionais**.

No entanto, esses estudos nem sempre garantem que os grupos comparados sejam equivalentes, o que pode gerar vieses e dificultar a interpretação dos resultados.

Para superar esse desafio, usamos técnicas de inferência causal, que ajudam a identificar o impacto real de uma variável (tratamento) em meio a possíveis influências externas.

Neste projeto, aplicaremos uma técnica chamada *Propensity Score Matching*, que busca equilibrar as características dos grupos de tratamento e controle, tornando-os mais semelhantes e, assim, facilitando a comparação.

Contexto do Problema

Neste trabalho, analisaremos o efeito de frequentar uma escola católica, em comparação com escolas não católicas, no desempenho dos alunos.

Como os estudantes de escolas católicas geralmente têm perfis diferentes dos de escolas não católicas, utilizaremos o pareamento por escore de propensão para obter estimativas causais mais confiáveis.

- **A pergunta que surge é:** a diferença no desempenho é causada pelas escolas ou pelas trajetórias de vida dos alunos, que os colocam em situações de desigualdade?
- Em outras palavras, a escola é a responsável pela desigualdade ou é a história de vida dos alunos que gera essa disparidade?

Propensity Score Matching

Propensity Score Matching (PSM) é uma técnica estatística usada reduzir o viés de seleção em estudos observacionais, onde os grupos de tratamento e controle não são formados de maneira aleatória.

O PSM usa um modelo probabilístico (geralmente uma **regressão logística**) para estimar a **probabilidade de um indivíduo receber o tratamento** com base em várias covariáveis.

Esse escore de propensão é, então, utilizado para parear indivíduos do grupo de tratamento com indivíduos do grupo de controle que possuem escores semelhantes, formando pares de observações comparáveis.

O resultado é uma estimativa do efeito causal que é mais confiável e precisa do que uma simples comparação de médias entre os grupos.

Embora o PSM dependa de variáveis observáveis e possa reduzir o tamanho da amostra, é uma ferramenta eficaz para aumentar a validade interna dos estudos.

Variáveis utilizadas

catholic: Se a escola é católica (1) ou não (0);

c5r2mtsc_std: Notas padronizadas de matemática;

p5himage: idade da mãe;

race_white: se aluno é da raça branca (1) ou não (0);

w3income_1k: Renda familiar (em milhares);

p5numpla: Número de lugares onde o aluno viveu por pelo menos 4 meses;

w3momed_hsb: O nível de educação da mãe é ensino médio ou menos (1) ou algum nível universitário ou superior (0)?

Lendo dados

```
library(kableExtra)    # Formatação de tabelas
library(MatchIt)       # Matching de dados em análises de causalidade
library(tidyverse)     # manipulação de dados
library(readr)          # usado para ler arquivos CSV

# Lê o arquivo CSV e armazena no dataframe 'ecls'
ecls <- read_csv("ecls.csv")

# Seleciona colunas específicas do dataframe 'ecls'
ecls <- ecls %>%
  select('catholic', 'c5r2mtsc_std', 'p5hmage',
         'race_white','w3income_1k', 'p5numpla', 'w3momed_hsb')

# Exibe as duas primeiras linhas do dataframe em formato de tabela
ecls %>% head(2) %>% kable()
```

catholic	c5r2mtsc_std	p5hmage	race_white	w3income_1k	p5numpla	w3momed_hsb
0	0.9817533	47	1	62.5005	1	0
0	0.5943775	41	1	45.0005	1	0

Análise descritiva

Média de matemática

- As notas nos testes de **c5r2mtsc_std** (média padronizada de matemática, utilizando a padronização z) diferem, em média, entre alunos de escolas católicas e não católicas?

```
ecls %>%
  group_by(catholic) %>%
  summarise(n_students = n(),
            mean_math = mean(c5r2mtsc_std),
            std_error = sd(c5r2mtsc_std) / sqrt(n_students)) %>%
  kable()
```

catholic	n_students	mean_math	std_error
0	4499	0.1631279	0.0145166
1	930	0.2196851	0.0281317

Média de matemática

- A **média** das notas padronizadas de matemática é **maior** para alunos de **escolas católicas** (0.2197) do que para alunos de escolas não católicas (0.1631).
- A diferença de médias ($0.2197 - 0.1631 = 0.0566$) sugere que, em média, alunos de escolas católicas apresentam um desempenho levemente superior em matemática; no entanto, a significância dessa diferença ainda precisa ser verificada.

Análise descritiva

Média das variáveis

```
ecls_cov <-c('race_white','p5hmage','w3income_1k','p5numpla','w3momed_hsb')  
ecls %>%  
  group_by(catholic) %>%  
  select(one_of(ecls_cov)) %>%  
  summarise_all(funs(mean(. ,na.rm = T))) %>% kable()
```

catholic	race_white	p5hmage	w3income_1k	p5numpla	w3momed_hsb
0	0.6537008	37.79462	65.39393	1.106246	0.3923094
1	0.7666667	39.77527	86.18063	1.073118	0.2053763

- **race_white:** A proporção de alunos brancos é maior em escolas católicas (76.67%) do que em não católicas (65.37%). Isso sugere uma maior concentração de alunos brancos em escolas católicas, o que pode estar relacionado com fatores históricos, sociais ou econômicos.

Média das variáveis

- *p5himage (Idade da mãe)*: A média da **idade das mães em escolas católicas** (39.78 anos) é ligeiramente **superior à das mães em escolas não católicas** (37.79 anos). Isso pode refletir um perfil socioeconômico mais elevado ou diferenças nos padrões familiares de planejamento em famílias de alunos que frequentam escolas católicas.
- *w3income_1k (Renda familiar em milhares)*: A **renda familiar** média é significativamente **maior entre alunos de escolas católicas** (86.180,63 /1000) em comparação com alunos de escolas não católicas (65.393,93 /1000). Esse é um dos fatores que podem influenciar o desempenho acadêmico, já que famílias com maior renda têm mais recursos para investir na educação e no desenvolvimento dos filhos.
- *p5numpla (Número de lugares onde o aluno viveu)*: Em média, os alunos de escolas não católicas (1.1062) tendem a ter vivido em um número ligeiramente maior de lugares do que os alunos de escolas católicas (1.0731). A diferença é pequena, mas pode sugerir uma leve maior estabilidade residencial em alunos de escolas católicas.
- *w3momed_hsb (Escolaridade da mãe até ensino médio)*: A proporção de mães com até ensino médio é maior em escolas não católicas (39.23%) do que em escolas católicas (20.54%). Isso sugere que, em média, as mães de alunos de escolas católicas têm níveis de escolaridade mais altos. Esse fator pode estar relacionado a uma maior conscientização sobre a importância da educação ou a melhores condições econômicas, o que também influencia o desempenho acadêmico dos alunos.

Média das variáveis

- Essas variáveis apontam para um perfil mais privilegiado dos alunos de escolas católicas em comparação com os de escolas não católicas.
- Alunos de escolas católicas, em geral, têm maior probabilidade de serem brancos, terem mães mais velhas e com maior escolaridade, além de virem de famílias com rendas mais altas.
- Esses fatores podem contribuir para o desempenho acadêmico superior observado nos testes de matemática, embora seja necessário um estudo mais aprofundado para confirmar essas relações e evitar conclusões simplistas.

Teste t

O teste t é uma ferramenta estatística usada para comparar a média de dois grupos e verificar se as diferenças observadas entre as médias são significativas ou podem ter ocorrido por acaso.

O que o teste t faz:

- **Hipótese nula (H_0)**: As médias dos dois grupos são iguais, ou seja, não há diferença significativa entre elas.
- **Hipótese alternativa (H_1)**: As médias dos dois grupos são diferentes, ou seja, há uma diferença significativa entre elas.
- O teste calcula uma estatística chamada valor de **t** e um **p-valor**, que nos ajuda a decidir se devemos rejeitar a hipótese nula ou não.
 - Se o **p-valor** for menor que um certo nível de significância (geralmente 0.05 ou 5%), rejeitamos a hipótese nula e concluímos que as médias são significativamente diferentes.
 - Caso contrário, não há evidências suficientes para rejeitar a hipótese nula.

Teste t - Média de matemática

```
with(ecls, t.test(c5r2mtsc_std ~ catholic))
```

```
##  
##      Welch Two Sample t-test  
##  
## data: c5r2mtsc_std by catholic  
## t = -1.7866, df = 1468.1, p-value = 0.07421  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.118653665 0.005539377  
## sample estimates:  
## mean in group 0 mean in group 1  
## 0.1631279      0.2196851
```

- **A diferença nas notas de matemática** entre alunos de escolas católicas (0,22) e não católicas (0,16) é pequena, e o teste t rejeita a hipótese de médias iguais apenas ao nível de significância de 10%, mas não ao de 5%.
- Isso sugere que a diferença **não é estatisticamente significativa** em um nível mais rigoroso, o que impede de afirmar com certeza que estudar em escola católica melhora o desempenho em matemática, já que a correlação pode ser casual mesmo considerando outros fatores.

Teste t - Raça

```
with(ecls, t.test(race_white ~ catholic)) # Raça branca
```

```
##  
##      Welch Two Sample t-test  
##  
## data: race_white by catholic  
## t = -7.2484, df = 1457.5, p-value = 6.814e-13  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.14353706 -0.08239463  
## sample estimates:  
## mean in group 0 mean in group 1  
## 0.6537008 0.7666667
```

- A análise mostra que a proporção de alunos brancos é significativamente maior em escolas católicas (76,67%) do que em não católicas (65,37%), com um p-valor extremamente baixo que permite rejeitar a hipótese de não haver diferença.
- Assim, ser branco está estatisticamente associado a estudar em uma escola católica.

Teste t - Idade da mãe

```
with(ecls, t.test(p5himage ~ catholic)) # Idade da mae
```

```
##  
##      Welch Two Sample t-test  
##  
## data: p5himage by catholic  
## t = -11.207, df = 1562.1, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -2.327294 -1.634002  
## sample estimates:  
## mean in group 0 mean in group 1  
##            37.79462            39.77527
```

- A análise revela que a idade média das mães de alunos em escolas católicas (39,78 anos) é significativamente maior do que a das mães de alunos de escolas não católicas (37,79 anos), com um p-valor extremamente baixo (2.2e-16).
- Isso permite rejeitar a hipótese de que não há diferença, concluindo que as mães de alunos em escolas católicas são, em média, mais velhas de forma estatisticamente significativa.

Teste t - Renda familiar (em milhares)

```
with(ecls, t.test(w3income_1k ~ catholic)) # Renda familiar
```

```
##  
##      Welch Two Sample t-test  
##  
## data: w3income_1k by catholic  
## t = -13.238, df = 1314.5, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -23.86719 -17.70620  
## sample estimates:  
## mean in group 0 mean in group 1  
##       65.39393      86.18063
```

- A análise mostra que a renda familiar média dos alunos em escolas católicas (86.180,63) é significativamente maior que a dos alunos em escolas não católicas (65.393,93). Com uma estatística t de -13,24, o teste t indica que essa diferença é estatisticamente significativa, tornando improvável que seja devida ao acaso.
- Assim, frequentar uma escola católica está associado a uma renda familiar significativamente mais alta.

Teste t - Quantidade de lugares em que alunos moraram

```
with(ecls, t.test(p5numpla ~ catholic)) # Quantidade de lugares em que alunos moraram
```

```
##  
##      Welch Two Sample t-test  
##  
## data: p5numpla by catholic  
## t = 3.128, df = 1600.4, p-value = 0.001792  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
##  0.01235472 0.05390038  
## sample estimates:  
## mean in group 0 mean in group 1  
##           1.106246           1.073118
```

- A análise mostra que os alunos de escolas não católicas viveram, em média, em mais lugares (1,11) do que os alunos de escolas católicas (1,07). Com uma estatística t de 3,13, essa diferença é estatisticamente significativa, sugerindo que os alunos de escolas católicas tendem a ter vivido em menos lugares.
- Portanto, há uma diferença significativa no número de lugares onde os alunos viveram, com os de escolas não católicas tendo maior mobilidade.

Teste t - Escolaridade da mãe até ensino médio (w3momed_hsb)

```
with(ecls, t.test(w3momed_hsb ~ catholic)) # Escolaridade da mae
```

```
##  
##      Welch Two Sample t-test  
##  
## data: w3momed_hsb by catholic  
## t = 12.362, df = 1545.1, p-value < 2.2e-16  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
##  0.1572715 0.2165946  
## sample estimates:  
## mean in group 0 mean in group 1  
## 0.3923094      0.2053763
```

- A análise mostra que a proporção de mães com escolaridade até o ensino médio é maior entre alunos de escolas não católicas (39,23%) em comparação com escolas católicas (20,54%). Com uma estatística t de 12,36, a diferença é estatisticamente significativa, sugerindo que não é devida ao acaso.
- Portanto, a escolaridade das mães de alunos de escolas católicas é significativamente mais alta, indicando uma menor proporção de mães com escolaridade até o ensino médio nesse grupo.

O que isso significa?

Significa que simplesmente comparar a média de notas de alunos de escolas católicas com alunos de escolas não católicas não é uma abordagem adequada, pois **o contexto e o background desses alunos são diferentes**.

Para abordar essa questão, utilizaremos a estimativa de *propensity score*. Essa técnica nos permitirá criar um contrafactual, buscando entre todos os alunos que não frequentam escolas católicas aqueles que mais se assemelham ao perfil dos alunos de escolas católicas.

Dessa forma, poderemos realizar comparações mais justas e precisas.

Estimativa do Propensity Score

Na prática, o processo de Propensity Score Matching geralmente envolve os seguintes passos:

Modelagem do Propensity Score: Utilizar regressão logística ou outro modelo para calcular a probabilidade de receber o tratamento com base em variáveis observáveis.

Matching: Parear indivíduos tratados e não tratados com base em seus propensity scores, utilizando métodos como matching por pares, matching com reposição ou matching mais próximo.

Análise de Resultados: Comparar os resultados entre os grupos pareados para avaliar o efeito do tratamento.

Para que o matching forneça uma estimativa causal confiável, é necessário incluir covariáveis que estejam relacionadas tanto à variável de tratamento quanto aos resultados potenciais.

Estimativa do Propensity Score - Regressão logística

Queremos calcular um propensity score individual que nos forneça a probabilidade de um aluno ser ou não de escola católica, levando em consideração variáveis como raça, renda, idade da mãe e o número de lugares onde a pessoa morou.

```
m_ps <- glm(catholic ~ race_white + w3income_1k + p5hmage + p5numpla + w3momod_hsb,  
            family = binomial(),  
            data=ecls) # modelo de regressão logística  
  
kable(summary(m_ps)$coefficients) # tabela com os coeficientes do modelo
```

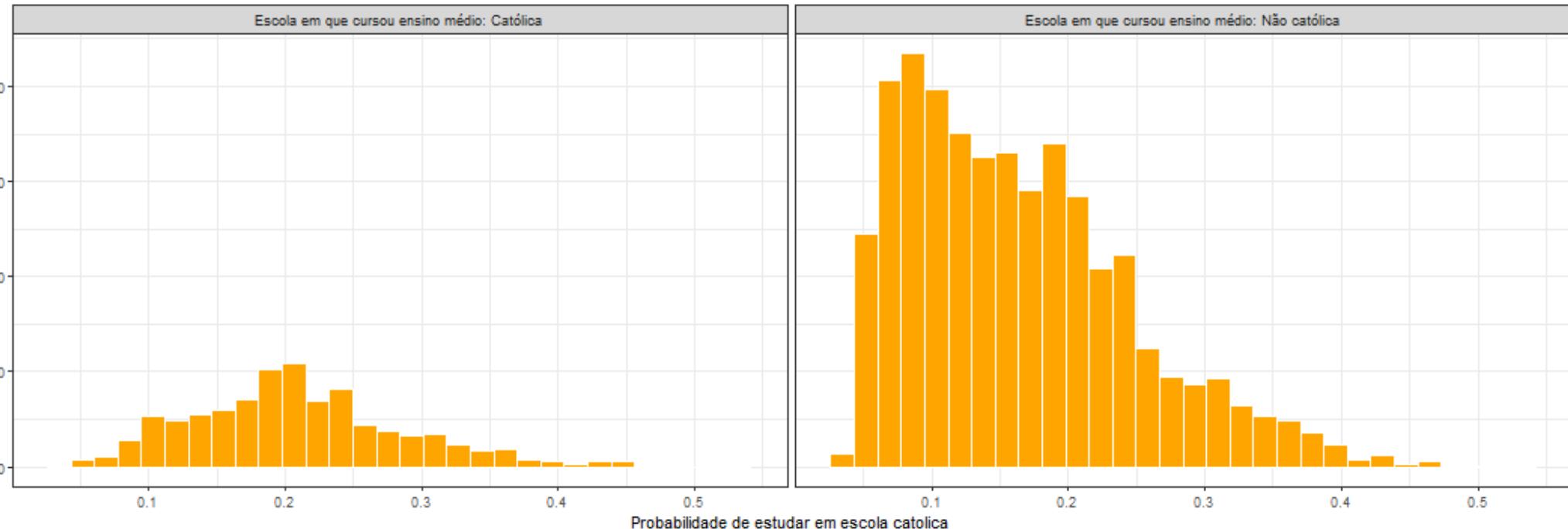
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.4092820	0.3263351	-10.447180	0.0000000
race_white	0.3001516	0.0870242	3.449058	0.0005625
w3income_1k	0.0063463	0.0008473	7.490225	0.0000000
p5hmage	0.0395347	0.0070820	5.582386	0.0000000
p5numpla	-0.2106296	0.1230279	-1.712047	0.0868880
w3momod_hsb	-0.5644212	0.0926141	-6.094336	0.0000000

Modelagem do Propensity Score - Regressão logística

- A raça do aluno (race_white) é significativa, sugerindo que alunos brancos têm uma maior probabilidade de estar em escolas católicas.
- A renda familiar (w3income_1k) também se mostrou altamente significativa, indicando que quanto maior a renda da família, maior a probabilidade de o aluno estar em uma escola católica.
- A idade da mãe (p5himage) também influencia positivamente essa probabilidade.
- Por outro lado, o número de lugares onde o aluno viveu (p5numpla) não apresentou significância estatística ao nível de 5%, mas tem um impacto negativo marginal.
- Finalmente, o nível educacional da mãe (w3momed_hsb) mostrou-se significativo, revelando que filhos de mães com ensino médio ou menos têm menor chance de frequentar escolas católicas, comparados a filhos de mães com nível superior.

Visualização dos Propensity Scores

A área de comum suporte é a faixa de valores dos *propensity scores* onde há sobreposição entre os grupos de tratamento e controle. Indivíduos fora dessa área são excluídos da análise, pois não possuem comparações adequadas no outro grupo, garantindo que apenas aqueles dentro dessa área possam ser comparados de forma válida para estimar o efeito do tratamento.



Aplicação do Matching

O objetivo do matching é equilibrar as características entre os dois grupos para tornar a comparação mais justa, como em um experimento randomizado.

No contexto de PSM, ele calcula as probabilidades de um indivíduo receber um tratamento com base em variáveis de confusão (ou covariáveis) e depois emparelha indivíduos com scores de propensão semelhantes nos grupos de tratamento e controle.

```
mod_match <- matchit(catholic ~ race_white +
                      w3income_1k +
                      p5hmage +
                      p5numpla +
                      w3momed_hsb,
                      method = "nearest",
                      data=ecls)

dta_m <- match.data(mod_match) # Nova base criada a partir do matching
```

Aplicação do Matching

Base após matching

Inicialmente, havia 930 indivíduos tratados e 4.499 no controle. Após o pareamento, todos os 930 tratados foram pareados com 930 controles, resultando em 3.569 controles que não tinham contrapartida no grupo de tratamento e foram descartados

```
resumo_mod_match <- summary(mod_match)  
kable(resumo_mod_match$nn)
```

	Control	Treated
All (ESS)	4499	930
All	4499	930
Matched (ESS)	930	930
Matched	930	930
Unmatched	3569	0
Discarded	0	0

Aplicação do Matching

Balanceamento das Covariáveis

Após o pareamento, as **diferenças** entre os grupos de tratamento e controle foram **drasticamente reduzidas**, o que poderia enviesar a estimativa do efeito do tratamento se essas diferenças não fossem ajustadas.

```
dta_m %>%
  group_by(catholic) %>%
  select(one_of(ecls_cov)) %>%
  summarise_all(funs(mean)) %>%
  kable()
```

catholic	race_white	p5himage	w3income_1k	p5numpla	w3momed_hsb
0	0.7720430	39.63441	86.00858	1.052688	0.2043011
1	0.7666667	39.77527	86.18063	1.073118	0.2053763

Análise de Resultados

- No início, ao realizar o teste t comparando as variáveis, todas apresentaram p-valor menor que 0,05.
- Após o pareamento, no entanto, não rejeitamos a hipótese nula, o que significa que, em relação a essas variáveis, o grupo de controle e o de tratamento são estatisticamente iguais.

variavel	Não.católico	Católico	p_value	statistic
race_white	0.7720430	0.7666667	0.7832882	0.2750763
p5hmage	39.6344086	39.7752688	0.5155120	-0.6504037
Renda familiar	86.0085828	86.1806253	0.9327609	-0.0843832
Quantidade de lugares morados	1.0526882	1.0731183	0.0978267	-1.6563263
Escolaridade da mãe	0.2043011	0.2053763	0.9542154	-0.0574218

Análise de Resultados

Com os dados pareados, é simples calcular os efeitos do tratamento. Uma opção é usar o **teste t**.

```
with(dta_m, t.test(c5r2mtsc_std ~ catholic))
```

```
##  
##      Welch Two Sample t-test  
##  
## data:  c5r2mtsc_std by catholic  
## t = 4.4425, df = 1847.3, p-value = 9.417e-06  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
##  0.1026983 0.2650474  
## sample estimates:  
## mean in group 0 mean in group 1  
##             0.4035580          0.2196851
```

- Os resultados mostram uma **diferença significativa** nas notas de matemática entre **alunos católicos e não católicos**, com um *p-valor* muito pequeno (0.000009417).
- Em média, os alunos não católicos têm notas significativamente maiores do que os alunos católicos.

Análise de Resultados

Uma outra maneira de analisar é usar uma regressão linear da variável de interesse, que é a média de matemática, em relação ao tipo de escola:

Term	Estimate	Std.Error	t.value	p.value
(Intercept)	0.4035580	0.0292666	13.789031	0.0e+00
catholic	-0.1838729	0.0413892	-4.442531	9.4e-06

- Em média, ser aluno de uma escola católica está associado a uma redução de 0.18387 nas notas de matemática, em comparação com alunos de escolas não católicas. O sinal negativo mostra que, após ajustar para a variável, os alunos de escolas católicas tendem a ter desempenho inferior.

Análise de Resultados

Para sofisticar um pouco mais o modelo, além de incluir apenas a variável católica, podemos controlar pelas demais variáveis para verificar se obtemos estimativas mais eficientes:

Term	Estimate	Std.Error	t.value	p.value
(Intercept)	-0.6936331	0.1947415	-3.561814	0.0003776
catholic	-0.1826328	0.0393196	-4.644835	0.0000036
race_white	0.2682888	0.0468558	5.725842	0.0000000
w3income_1k	0.0030883	0.0004647	6.645418	0.0000000
p5himage	0.0214028	0.0042952	4.983010	0.0000007
p5numpla	-0.1454435	0.0740612	-1.963830	0.0496991
w3momed_hsb	-0.3462700	0.0497587	-6.958982	0.0000000

- **O resultado se mantém:** alunos de escolas católicas têm desempenho em matemática pior do que alunos de escolas não católicas quando usamos uma estrutura de pareamento.

Análise de Resultados

Mesmo controlando todos os atributos de trajetórias que parecem relevantes, ainda assim **encontramos médias de desempenho em matemática diferentes**.

Esse resultado indica que **a escola tem uma diferença em relação aos alunos**; mesmo alunos com backgrounds iguais obtêm resultados diferentes.

Conclusão

- A diferença que observamos ao analisar apenas as médias inicialmente leva a uma conclusão equivocada.
- Se comparássemos apenas as médias entre os tipos de escola, diríamos que as médias dos dois tipos de escolas são iguais para os alunos.
- No entanto, ao realizarmos o pareamento — uma técnica para dados observacionais que segue uma lógica quase experimental — e compararmos dois grupos muito semelhantes, o resultado se revela diferente.
- Na verdade, alunos de escolas não católicas que têm trajetórias de vida semelhantes às dos alunos de escolas católicas apresentam resultados melhores, em média, nas notas de matemática.

Referências

R Tutorial 8: Propensity Score Matching: <https://simonejdemyr.com/r-tutorials/statistics/tutorial8.html>

Avaliação de políticas públicas B: https://www.youtube.com/watch?v=uMJeojTOcxc&ab_channel=CanaldaQuaest

HW11a - The Lost Homework on PSM and LASSO: <https://rpubs.com/metricsdawg/1037525>