UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

LUCAS FERNANDES DE MAGALHÃES

Simulando dados para ensinar avaliação de impacto de políticas públicas: uma abordagem prática

Belo Horizonte
2022

LUCAS FERNANDES DE MAGALHÃES

Simulando dados para ensinar avaliação de impacto de políticas públicas: uma abordagem prática

Monografia de especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística.

Orientador: Prof. Dr. Roberto da Costa Quinino

Belo Horizonte
2022

Magalhães, Lucas Fernandes de.

M189s Simulando dados para ensinar avaliação de impacto de políticas públicas [recurso eletrônico]: uma abordagem prática / Lucas Fernandes de Magalhães —2022.
1 recurso online (27 f. il, color.): pdf.

Orientador: Roberto da Costa Quinino
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 27.

1. Estatística. 2. Econometria. 3. Políticas públicas-Avaliação. I. Quinino, Roberto da Costa. II. Universidade Federal de Minas Gerais I. Instituto de Ciências Exatas, Departamento de Estatística .III.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 266 ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE LUCAS FERNANDESS DE MAGALHÃES.

Aos dezenove dias do mês de dezembro de 2022, às 16:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso do aluno **Lucas Fernandes de Magalhães**, intitulado: "Simulando dados para ensinar avaliação de impacto de políticas públicas: uma abordagem prática", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Roberto da Costa Quinino – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do candidato. Após a defesa, os membros da banca examinadora reuniram-se sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o candidato foi considerado Aprovado condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente o candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 19 de dezembro de 2022.

Roberto da Costa Quinino:80871291720
Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2022.12.21 19:32:51 -03'00'

Prof. Roberto da Costa Quinino (Orientador)
Departamento de Estatística / ICEX / UFMG

Mateus Morais Araújo
Fhemig - Fundação Hospitalar do Estado de Minas Gerais

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Lucas Fernandes de Magalhães, número de registro 2020705790, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que me entregou a versão final corrigida. O trabalho foi apresentado no dia 19 de dezembro de 2022 com o título "*Simulando dados para ensinar avaliação de impacto de políticas públicas: uma abordagem prática*".

Belo Horizonte, 27 de dezembro de 2022

Roberto da Costa Quinino:80871291720
Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2022.12.27 08:32:39 -03'00'

Prof. Roberto da Costa Quinino
Coordenador da Comissão do Curso de Especialização em Estatística

Prof. Roberto da Costa Quinino
Coordenador do curso de
Especialização em Estatística
Departamento de Estatística / UFMG

Simulando dados para ensinar avaliação de impacto de políticas públicas: uma abordagem prática

RESUMO

Um dos principais desafios metodológicos enfrentados pela Economia e Ciência Política consiste em avaliar o impacto de políticas públicas. A política de transferência condicional de renda aumentou os anos de escolaridade dos filhos das famílias benefi-ciadas pelo programa? O aumento do salário mínimo eleva o desemprego? A cobrança pelo despacho da bagagem reduziu o preço das passagens aéreas? Encontrar respostas para essas perguntas não é uma tarefa trivial. Geralmente, para se evitar vieses na análise, são utilizados desenhos experimentais ou quase-experimentais de difícil compreensão para alunos pouco versados em econometria e estatística. Neste trabalho proponho a utilização de bases simuladas para se ensinar de forma didática a intuição por trás desses desenhos e apresento exemplos práticos.

Palavras-chave: Avaliação de impacto. Econometria. Políticas Públicas.

Simulating data to teach impact evaluation of public policies: a practical approach

ABSTRACT

One of the main methodological challenges faced by Economics and Political Science is to evaluate the impact of public policies. Did the conditional income transfer policy increase the years of schooling of the children of the families benefiting from the program? Does increasing the minimum wage raise unemployment? Did charging for baggage handling reduce the price of airline tickets? Finding answers to these questions is not a trivial task. Usually, to avoid biases in the analysis, experimental or quasi-experimental designs are used that are difficult to understand for students with little knowledge of econometrics and statistics. In this work, I propose the use of simulated data to teach the intuition behind these designs in a didactic way and present practical examples.

Keywords: Impact evaluation. Econometric. Public policies.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Um dos principais desafios metodológicos enfrentados pela Economia e Ciência Política consiste em avaliar o impacto de políticas públicas - PPs. A política de transferência condicional de renda aumentou os anos de escolaridade dos filhos das famílias beneficiadas pelo programa? O aumento do salário mínimo eleva o desemprego? A cobrança pelo despacho da bagagem reduziu o preço das passagens aéreas? Encontrar respostas para essas perguntas não é uma tarefa trivial. Simplesmente calcular a diferença de médias do grupo que recebeu a política pelo grupo não beneficiado leva a resultados enviesados, tendo em vista que eles não são comparáveis. Para evitar o viés, recorre-se, geralmente, a um experimento ou à utilização de desenhos quase-experimentais.

No entanto, mesmo sendo complexo estimar efeitos causais no mundo, um desafio, talvez, ainda mais árduo consiste em ensinar a teoria que embasa a identificação dos efeitos causais. Para estudantes de ciências sociais, geralmente pouco versados em estatística e econometria, os experimentos e desenhos quase-experimentais parecem ser desenvolvidos em uma linguagem quase mística. Nesse sentido, utilizar exemplos práticos em sala de aula pode ser uma estratégia didática importante para se tentar contornar a resistência inicial dos alunos em matérias com alto nível de exigência metodológica (MEIRELES, SILVA e CORREA, 2017).

Neste trabalho, proponho a utilização de bases simuladas para se ensinar de forma prática a intuição por trás desses desenhos. Para se criar as bases simuladas, foi desenvolvida uma função em R que pode ser acessada pelo link: https://github.com/Lucasfm108/causalsim. A função permite que alunos e professores testem diversas situações, algumas em que os desenhos retornam os efeitos verdadeiros das PPs e outras em que, havendo violação de alguma suposição, retornam um efeito enviesado. O objetivo principal é que as bases simuladas sejam utilizadas como um recurso adicional em sala, tanto para aulas práticas quanto, especialmente, para exercícios avaliativos.

Nas próximas seções explicaremos brevemente as suposições que permitem a identificação do efeito causal por experimentos e pareamento, bem como daremos exemplos práticos da estimação dos efeitos usando as bases simuladas criadas a partir da função construída no R. Tendo em mente o público-alvo, alunos com pouca ou nenhuma formação em estatística e econometria, as explicações teóricas das suposições para identificação, bem como as estimações dos efeitos não demandarão nenhum conhecimento sofisticado em metodologia.

2 O PROBLEMA FUNDAMENTAL DA INFERÊNCIA CAUSAL

Em primeiro lugar, por efeito estamos nos referindo à mudança de uma variável, que geralmente atribuímos à letra “Y”, em função de outra tipicamente identificada pela letra“X”. Suponha que em uma amostra aleatória da população foram observados o nível de esco-laridade (X) e a renda anual (Y). Analisando os dados, identificou-se que o grupo com ensino superior possuía, em média, uma renda mensal maior do grupo que não cursou qualquer tipo de graduação. Denominamos de efeito a mudança em Y que parece estar associada à variação de X, que pode ser calculada a partir de uma simples diferença de média entre os dois grupos. Pela Tabela 1, o efeito seria de R$ 3.000,00.

Tabela 1 - Nível de escolaridade x Renda entre grupos incomparáveis

	Escolaridade	Renda.Média
Grupo 1	Ensino superior	4000
Grupo 2	Ensino Médio/Fundamental	1000

Fonte: Elaborado pelo autor a partir de dados simulados.

Mas, baseados neste exemplo, podemos afirmar com segurança que obter um diploma de nível superior “causa” um aumento médio na renda mensal de R$3.000,00? Não. Nada impede que outras variáveis, chamadas de variáveis de confusão [1] ou confusores (ANGRIST; PISCHKE, 2008; DUARTE, 2021), estejam interferindo nessa associação entre X e Y, como, por exemplo, a renda dos pais. É plausível que os pais dos alunos que ingressam em alguma faculdade possuam também uma renda superior a dos pais dos alunos sem graduação. E, como se sabe que a renda dos pais é tipicamente associada à renda dos filhos, uma parte da associação entre X e Y, na verdade, pode ser explicada pela renda dos pais, e não pelo diploma em ensino superior. Em outras palavras, o grupo que cursou a graduação, mesmo sem ter cursado, naturalmente teria uma renda maior que a do outro grupo em função de diversas variáveis para além do diploma em ensino superior.

Para Angrist e Pischke (2008), a impossibilidade de se afirmar que a diferença de médias entre os dois grupos foi causada pela intervenção decorre de um “problema de seleção”. Pessoas que buscam ingressar em alguma universidade já são tipicamente mais ricas do que aquelas que

[1] São variáveis que estão associadas tanto a X (Intervenção ou tratamento) quanto a Y (variável resposta).

se abstém de cursar alguma faculdade. Há, portanto, um viés de seleção que impede que se interprete a diferença de médias de renda entre os dois grupos como sendo a quantidade que representa o efeito causal da intervenção.

No fundo, isso nos remete a um problema de incomparabilidade entre grupos. Aquele que cursou a graduação não é diretamente comparável àquele que possui apenas ensino médio ou fundamental, seja por uma diferença de renda anterior, habitação, capital social entre outras variáveis.

Para contornar esse problema de incomparabilidade, idealmente, seria possível comparar a renda do grupo que cursou a graduação com a renda desse mesmo grupo, mas em uma dimensão paralela em que eles apenas obtiveram o ensino médio/fundamental. Esses dois grupos seriam perfeitamente iguais em tudo (forma de criação, estrutura familiar, renda dos pais, capital social, etc.), sendo a única diferença entre eles o nível de escolaridade. De posse das rendas de ambos os grupos, a mera diferença de médias retornaria o verdadeiro efeito do nível escolaridade (X) sobre a renda (Y), uma vez que nenhuma outra variável poderia estar interferindo nesse cálculo[2]. Esse processo de exclusão das potenciais explicações alternativas para o efeito de X sobre Y ($X \Rightarrow Y$), que se faz por meio da construção de grupos comparáveis, é denominado identificação (HUNTINGTON-KLEIN, 2021).

Tabela 2 - Nível de escolaridade x Renda entre grupos perfeitamente comparáveis

	Escolaridade	Renda.Média
Grupo 1	Ensino superior	4000
Grupo 1 - Contrafactual	Ensino Médio/Fundamental	3500

Fonte: Elaborado pelo autor a partir de dados simulados.

Como ilustra a Tabela 2, ao utilizarmos realmente grupos comparáveis, estima-se um efeito menor[3] (R$ 500,00) do que na primeira análise, indicando que a mera diferença de médias

[2] Formalmente, quando ambos os grupos são comparáveis, a distribuição do tratamento T é ortogonal ou independente dos resultados potenciais $Y(1)$ e $Y(0)$. Isto é, recorrendo ao exemplo anterior das dimensões paralelas, a probabilidade de alguém cursar ou não cursar uma faculdade não estaria associada a sua futura renda.

[3] A rigor, o efeito apontado pela comparação entre o grupo que recebeu o tratamento e ele mesmo na situação contrafactual de não ter recebido é denominado *average treatment effect on treated - ATT*

entre grupos não-comparáveis (Tabela 1) gera uma estimativa enviesada do efeito do nível de escolaridade, certamente afetada por confusores.

O problema óbvio, que nos impede na vida real de estimar o efeito verdadeiro, é que, enquanto viagens interdimensionais não forem acessíveis a pesquisadores, em um dado momento podemos observar um indivíduo que possui ensino superior e outro que não possui ensino superior, mas nunca um mesmo indivíduo possuindo e não possuindo ao mesmo tempo um ensino superior, o que impede a observação de grupos perfeitamente comparáveis. Eis o problema fundamental da inferência causal (MORGAN e WINSHIP, 2007).

Todavia, embora contrafactuais perfeitos, tal como os ilustrados pela Tabela 2, não existam, às vezes é possível encontrar nos dados uma boa aproximação deles. Ou seja, é possível encontrar em algumas situções grupos que, embora não sejam perfeitamente iguais, são semelhantes a ponto de permitirem a exclusão das explicações alternativas para o efeito de X sobre Y (identificação). Nas próximas seções iremos explicar alguns desenhos utilizados justamente para se "fabricar" grupos comparáveis que permitem a estimação dos efeitos das PPs.

3 EXPERIMENTO

Desenhos experimentais recorrem à aleatorização/randomização do tratamento entre os elementos da amostra, nome dado à variável de interesse que imagina-se exercer algum impacto sobre a variável dependente ou resposta. Quando há a aleatorização do tratamento, os dois grupos criados (GT - grupo de tratamento e GC - grupo de controle) são, em média, comparáveis tanto nas variáveis observadas (aquelas que constam da tabela) quanto nas não observadas, sendo a única diferença entre eles a ausência/presência do tratamento.

No contexto de avaliação de impacto, a PP analisada é o próprio tratamento e a sua randomização entre os indivíduos da amostra faz com que aqueles beneficiados sejam comparáveis, em média, aos que não receberam o benefício. Sendo a única diferença entre os grupos a presença ou ausência de tratamento, a diferença de médias do indicador de impacto entre eles retorna o efeito verdadeiro da PP.

Para comprovar a comparabilidade entre os grupos criados através da randomização do tratamento e o fato de que a diferença de médias é um estimador que retorna o efeito verdadeiro quando o tratamento é aleatorizado, pode-se recorrer à simulação de dados.

Figura 1 - Código em R da função causalsim

```
bd <- causalsim(n=10000,treat_effect=500, prob_treat = 0.5,
                covar_dist=c("binom","binom","norm"),
                covar_prob=c(0.4,0.75),
                covar_mean=c(3000),
                covar_sd=c(500),
                covar_effect = c(100,150,3),
                intercept=500)
```

Fonte: Elaborado pelo autor.

A função **causalsim**[4] acima serve para criar uma base de dados contendo *n* unidades, uma coluna *id* diferenciando cada uma delas, uma coluna para indicar quem recebeu o tratamento (*treat*), o valor da variável dependente/resposta (*outcome*) para cada unidade e uma série de

[4] O código da função pode ser acessado no link: https://github.com/Lucasfm108/causalsim. Ressalte-se que boa parte desta função foi construída usando funções do pacote "faux" (DEBRUINE, 2021).

covariáveis (*x_1* a *x_p*,sendo *p* covariáveis definidas). Abaixo segue uma breve descrição de cada um dos argumentos:

- n: Indica o total de unidades;
- treat_effect: o efeito verdadeiro do tratamento sobre a variável dependente/resposta;
- prob_treat: a probabilidade de cada unidade receber o tratamento;
- covar_dist: a função de distribuição de cada uma das covariáveis, podendo ser uma normal ou binomial. Recebe um vetor texto com os valores “norm” e “binom”;
- covar_prob: caso seja atribuída a alguma das covariáveis a distribuição binomial, deve- se indicar a probabilidade de cada uma delas em um vetor numérico;
- covar_mean e covar_sd: caso seja atribuída a alguma das covariáveis a distribuição nor- mal, deve-se indicar a média e o desvio-padrão delas;
- covar_effect = efeito das covariáveis sobre a variável dependente/resposta, podendo ser 0;
-intercept: é o termo constante na função que define os valores da variável dependente/resposta.

Tabela 3 - Primeiras 10 linhas da tabela criada pela função causalsim

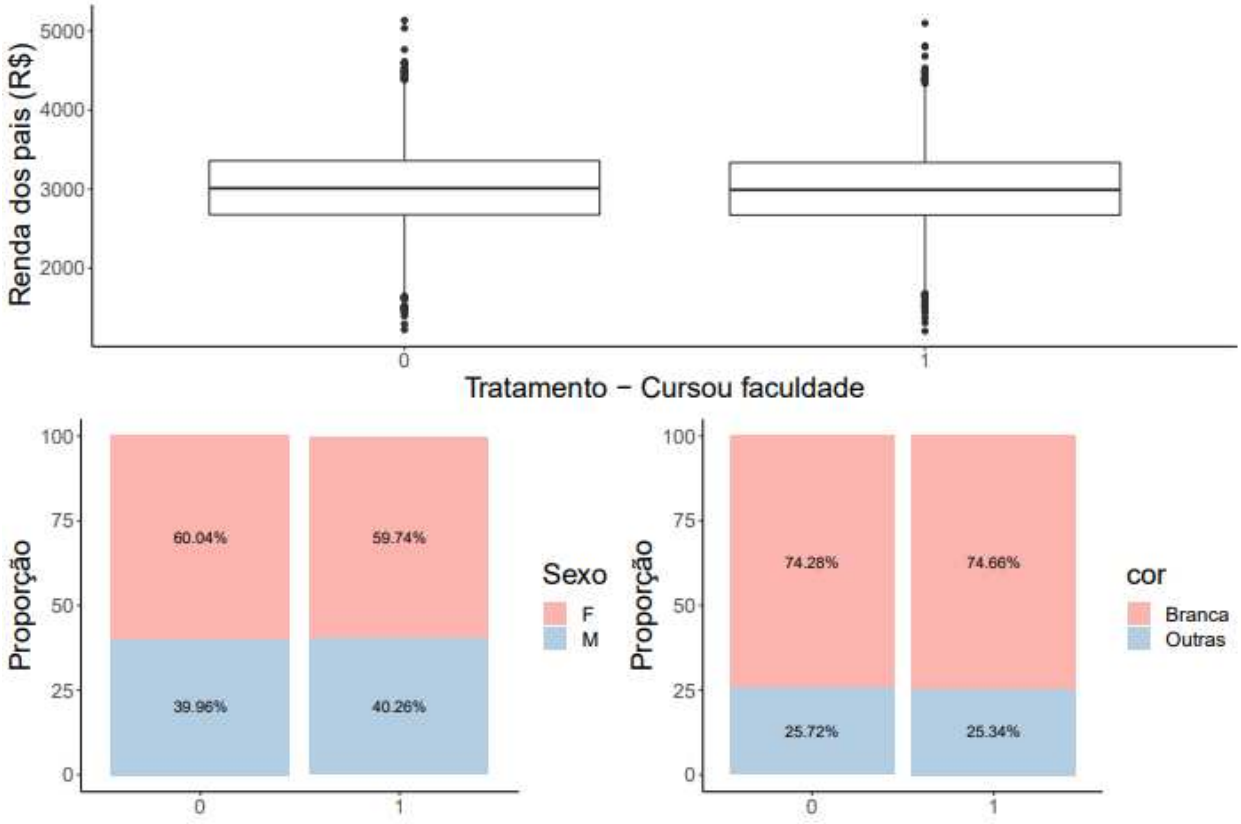
id	cursou_faculdade	renda_pais	sexo	cor	renda
1	0	2547.207	M	Outras	8241.584
2	1	3738.367	M	Outras	12313.885
3	1	2787.214	F	Branca	9511.376
4	0	3922.704	M	Branca	12517.615
5	0	3671.769	M	Outras	11615.712
6	1	2534.741	M	Branca	8853.051
7	0	2759.791	F	Branca	8931.138
8	1	2553.459	F	Branca	8809.555
9	1	1778.773	M	Branca	6586.776
10	1	2798.329	M	Branca	9643.252

Fonte: Elaborado pelo autor a partir de dados simulados.

Os nomes das colunas e de algumas categorais que retornam na função foram alterados simplesmente para encaixá-la no exemplo anterior sobre o efeito do ensino superior sobre a renda futura. Imagine um experimento no qual alunos recém-saídos do ensino médio aceitaram participar de uma experiência ousada na qual alguns deles foram selecionados aleatoriamente

para cursar uma faculdade (Grupo de tratamento) e outros ingressaram no mercado de trabalho diretamente, sem continuar o ensino (Grupo de controle). Dez anos após a aleatorização dos alunos, foi realizado um survey por meio do qual foram medidas a renda do entrevistado (variável resposta) e algumas covariáveis: sexo, cor e renda dos pais.

Figura 2- Visualização da associação entre o tratamento e as covariáveis



Fonte: Elaborado pelo autor.

Como os alunos foram selecionados aleatoriamente para ingressar em uma faculdade ou permanecer sem ensino superior, nenhuma das covariáveis possui uma associação com a distribuição do tratamento. Em outras palavras, os grupos de tratamento e controle são comparáveis, sendo a única diferença entre eles o nível de escolaridade, como ilustra a Figura 2.

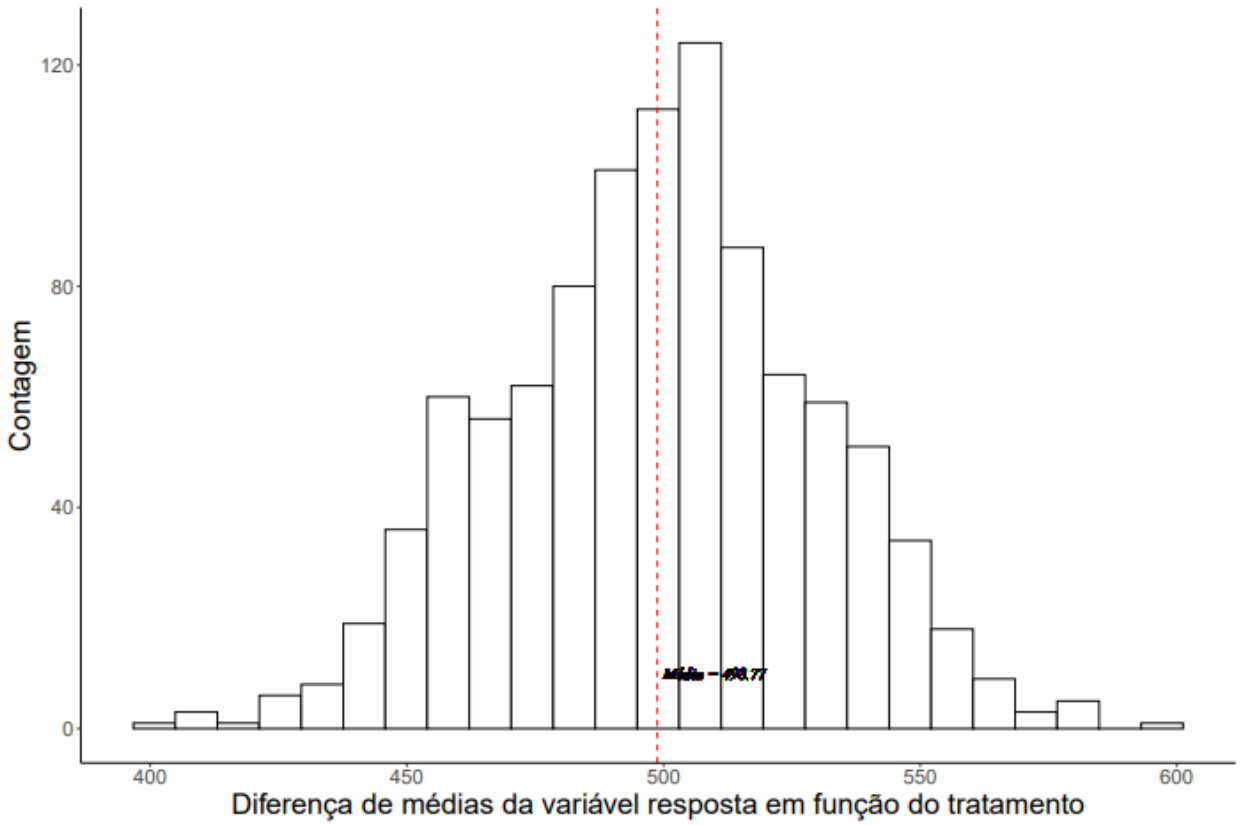
Tabela 4 - Comparação da renda entre os grupos de tratamento e controle

Cursou faculdade?	Renda (R$)
Não	9689,99
Sim	10146,99

Fonte: Elaborado pelo autor a partir de dados simulados.

O valor estimado pela diferença de médias em algumas simulações pode se distanciar um pouco do efeito verdadeiro em função do erro amostral. Todavia, gerando uma quantidade razoável de simulações, a média das diferenças de médias converge para o efeito verdadeiro, como ilustra a Figura 3.

Figura 3 - Distribuição das diferenças de médias de 1.000 simulações de tamanho n=10000



Fonte: Elaborado pelo autor.

Naturalmente, o mesmo valor pode ser estimado usando a regressão linear com ou sem a inclusão das covariáveis, conforme a Tabela 4. Aliás, essa é uma boa oportunidade para explicar na prática duas intuições relevantes: a) quando se tem um regressor binário e uma variável resposta contínua, o coeficiente $\beta 1$ da regressão linear é igual à diferença de médias entre os dois grupos[5]; b) em um desenho experimental, mesmo que uma covariável possua um efeito sobre a variável resposta, incluí-la ou não no modelo não afeta o coeficiente do tratamento, apenas o torna mais preciso, reduzindo o erro padrão da estimativa (DOLAN; GREEN, 2016).

Tabela 5 - Resultados dos modelos de regressão linear

	Mod. sem covariável	Mod. com covariável
Ensino superior	457.004***	499.972***
	(30.317)	(0.020)
Sexo (Ref.: Masculino)		100.010***
		(0.020)
Cor (Ref.:Branca)		149.988***
		(0.023)
Renda dos pais		3.000***
		(0.00002)
Intercepto	9,689.987***	499.931***
	(21.470)	(0.064)
Observations	10,000	10,000
R2	0.022	0.9999996
Adjusted R2	0.022	0.9999996
Residual Std. Error	1,515.861 (df = 9998)	1.000 (df = 9995)
F Statistic	227.226*** (df = 1; 9998)	5,870,918,289.000*** (df = 4; 9995)
Notas:		*p<0.1; **p<0.05; ***p<0.01

Fonte: Elaborado pelo autor com base em dados simulados

[5] Em outras palavras, o teste-t é igual a uma regressão linear simples com um regressor binário.

O experimento é o padrão-ouro para estimação de efeitos causais, justamente porque a aleatorização do tratamento produz grupos comparáveis, facilitando a superação do problema da identificação. Entretanto, em muitos casos, como no próprio exemplo explorado, é inviável a randomização da política pública. É improvável que alunos aceitem participar do grupo de controle (isto é, não cursar uma faculdade) em prol de um experimento mesmo sob o risco de prejudicar seu próprio futuro.

Nessas situações, não há garantia de que os dois grupos, aquele que se beneficiou de uma política e o grupo de controle, sejam comparáveis, impedindo a estimação do efeito ver-dadeiro por meio de uma simples diferença de médias. Quando isso ocorre, a saída é recorrer aos desenhos quase-experimentais, como o pareamento que será explicado a seguir. Nesses desenhos, embora o tratamento não seja aleatorizado, são “fabricados” grupos comparáveis, de tal modo que a distribuição do tratamento fica “como se fosse randomizada”.

4 PAREAMENTO

Para se explicar a aplicação do desenho de pareamento usando dados simulados, basta acrescentar o argumento *treat_covar* na função **causalsim** que introduz uma correlação ($[-1, 1]$) entre a covariável e o tratamento. Desde que haja uma associação entre a covariável e o tratamento, bem como um efeito da primeira sobre a variável resposta, tem-se ai uma confusora, fato que impede a estimação do efeito verdadeiro pela simples diferença de médias entre os grupos de tratamento e controle.

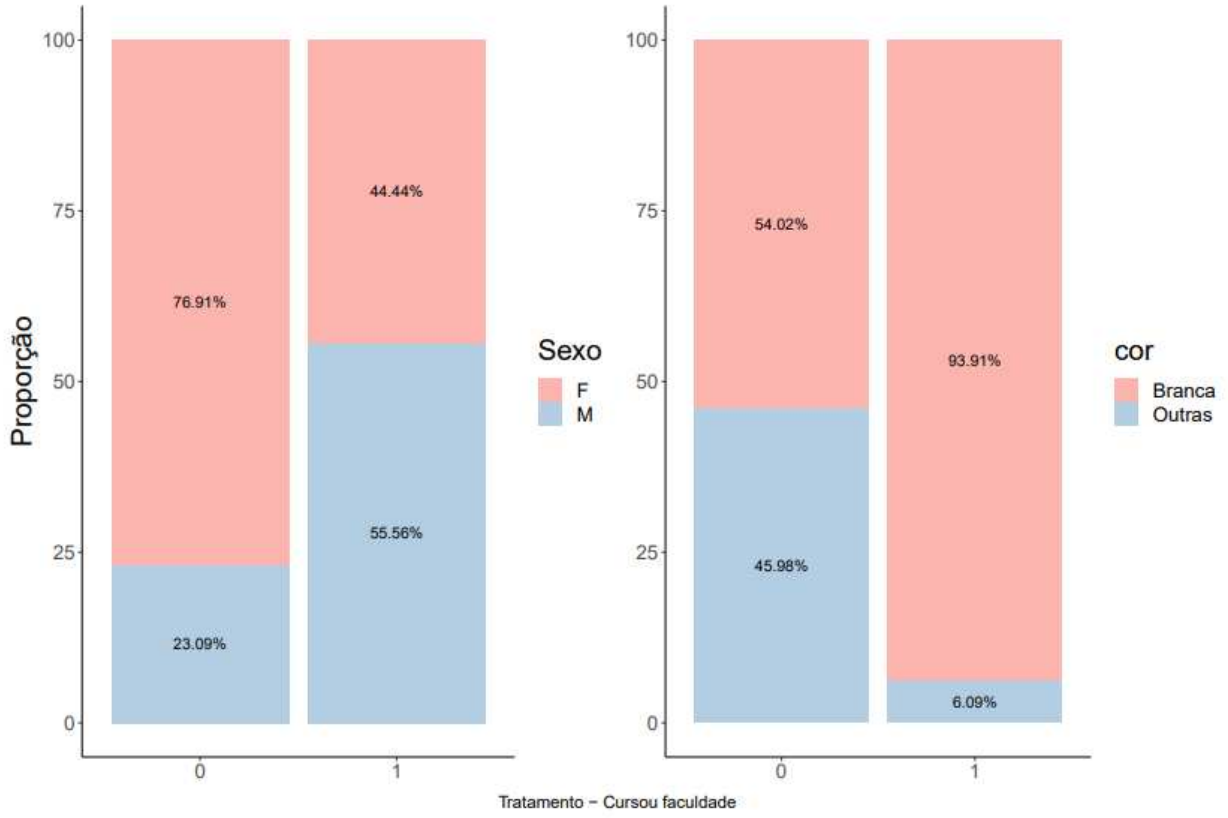
Figura 4 - Função causalsim para criar grupos não comparáveis

```
bd <- causalsim(n=10000,treat_effect=1000, prob_treat = 0.5,
                covar_dist=c("binom","binom"),
                covar_prob=c(0.4,0.75),
                covar_effect = c(400,600),
                treat_covar=c(0.4,0.6),
                intercept=500)
```

Fonte: Elaborado pelo autor.

A partir da base simulada gerada pelo código acima pode-se observar que os dois grupos não são comparáveis (Figura 3) e, portanto, não há uma identificação do efeito. Há uma proporção maior de homens com curso superior em relação a mulheres e de brancos em relação a não-brancos. Ou seja, não é possível afastar a hipótese de que a diferença de médias de renda entre o grupo de tratamento e controle decorram unicamente da diferença da proporção de sexo ou cor em ambos os grupos.

Figura 5 - Visualização da associação entre o tratamento e as covariáveis



Fonte: Elaborado pelo autor.

Tabela 6 - Comparação da renda entre GT e GC não comparáveis

Cursou faculdade?	Renda (R$)
Não	916,49
Sim	2285,74

Fonte: Elaborado pelo autor a partir de dados simulados

Entretanto, embora os grupos de tratamento e controle não sejam diretamente comparáveis, é possível encontrar nos dados grupos comparáveis. É sabido que as únicas variáveis que afetam a renda do entrevistado são o tratamento (ter cursado faculdade), o sexo e a

cor[6]. Dessa forma, existem 4 (quarto) estratos na base de dados que são perfeitamente comparáveis: homens brancos com nível superior vs homens brancos sem nível superior, homens não brancos com nível superior vs homens não-brancos sem nível superior e assim por diante, conforme exemplifica a Tabela 7.

Tabela 7 - Comparação da renda entre estratos comparáveis

Estrato	sexo	cor	Grupo de Tratamento	Grupo de controle	Proporção	Dif. De médias
1	F	Branca	2100.02	1100.02	0.4144	1000.00
2	F	Outras	1500.06	500.02	0.1904	1000.04
3	M	Branca	2500.01	1500.03	0.3276	999.98
4	M	Outras	1899.93	899.96	0.0676	999.97

Fonte: Elaborado pelo autor a partir de dados simulados.

Eis a estratégia do pareamento: estratificar a base de dados por todas as covariáveis até criar grupos perfeitamente comparáveis e, só então, calcular a soma das diferenças de médias em cada estrato ponderando pelo seu tamanho[7]. A Tabela 7 exemplifica esse processo, permitindo a estimação do efeito verdadeiro de R$ 1000.

Figura 6 - Função causalsim para criar associação espúria

[6] Sabe-se disso pois assim foi definido na função **causalsim.**

[7] Quando há uma variável contínua ou discreta, não é possível estimar o efeito por meio da estratificação (*subclassification*) em razão do problema da dimensionalidade (CUNNINGHAM, 2021). Com muitos estratos, boa parte deles conterão apenas unidades tratadas ou sem tratamento, violando a suposição de suporte comum. Nessa situação, para se estimar o efeito utiliza-se algoritmos de matching.

```
bd <- causalsim(n=10000,treat_effect=0, prob_treat = 0.5,
                covar_dist=c("binom","binom", "norm"),
                covar_prob=c(0.4,0.75),
                covar_mean=2000,
                covar_sd=400,
                covar_effect = c(300,200, 2.5),
                treat_covar=c(0.7,0.6, 0.5),
                intercept=500)
```

Fonte: Elaborado pelo autor a partir de dados simulados.

Por meio da simulação de dados e recorrendo ao pareamento também é possível exemplificar casos em que o efeito estimado pela diferença de médias trata-se apenas de uma associação espúria, como na simulação acima. Embora aparentemente haja um efeito do tratamento sobre a renda de R$ 1246.97, após o pareamento percebe-se que não há efeito algum (2º Modelo da Tabela 8). Ocorre que uma associação espúria entre o tratamento e a variável resposta foi gerada pela correlação entre o tratamento e as covariáveis, as quais, estas sim, possuem um efeito sobre a renda.

Tabela 8 - Resultados dos modelos de regressão linear - associação espúria

	Mod. sem covariáveis	Mod. com covariáveis
Ensino superior	1.246,98***	0,02
	(17,58)	(0.03)
Sexo (Ref.: Masculino)		299,98***
		(0,02)
Cor (Ref.:Branca)		-200,01***
		(0,02)
Renda dos pais		2.500***
		(0.00)
Intercepto	5.144,62***	700.01***
	(12,39)	(0,05)
Observações	10.000	10.000
R2	0.335	0.9999991
Adjusted R2	0.335	0.9999991

Residual Std. Error	879.156 (df = 9998)	0.991 (df = 9995)
F Statistic	5,029.265*** (df = 1; 9998)	2,958,171,790.000*** (df = 4; 9995)
Notas		*p<0.1; **p<0.05; ***p<0.01

Fonte: Elaborado pelo autor a partir de dados simulados.

Infelizmente, na vida real nunca se sabe ao certo se todas as confusoras foram controladas ou se alguma delas foi omitida do modelo. Mesmo estratificando a base por *n* variáveis, nada impede que haja uma variável *n* + 1 que foi deixada de lado pelos pesquisadores. Caso isso ocorra, é impossível estimar o efeito verdadeiro. Essa situação também pode ser verificada por meio de uma simulação. Quando se acrescenta a renda dos pais como uma variável confusora, a estratificação unicamente pelo sexo e cor não permite a estimação do efeito verdadeiro, como aponta o resultado do primeiro modelo de regressão linear sem a inclusão da renda dos pais na Tabela 9.

Tabela 9 - Resultados dos modelos de regressão linear

	Mod. sem renda dos pais	Mod. com renda dos pais
Ensino superior	1.985,76***	999,99***
	(20,66)	(0,03)
Sexo (Ref.: Masculino	283***	300***
	(18,74)	(0,21)
Cor (Ref.:Branca)	206,55***	199,99***
	(22,49)	(0.03)
Renda dos pais		2500***
		(0,00)
Intercepto	5.012,63***	499,98***
	(17,67)	(0,55)
Observations	10.000	10.000
R2	0.613	0.9999995
Adjusted R2	0.613	0.9999995
Residual Std. Error	868.672 (df = 9996)	0.994 (df = 9995)
F Statistic	5,270.911 (df = 3; 9996)	4,928,837,313.00 (df = 4;9995)
Notas		*p<0.1; **p<0.05; ***p<0.01

Fonte: Elaborado pelo autor a partir dos dados simulados.

Deve-se reforçar que a estimativa incorreta do primeiro modelo da Tabela 9 não é culpa de um erro amostral, o qual poderia ser resolvido com uma amostra maior. Na verdade, se trata de um problema de identificação. Mesmo com uma amostra de tamanho infinito não seria

possível estimar o efeito verdadeiro sem a observação de todas as variáveis de confusão incluídas no modelo.

CONSIDERAÇÕES FINAIS

A finalidade principal da função **causalsim** apresentada neste trabalho consiste em facilitar a criação de dados simulados que permitam um ensino mais didádito sobre avali-ação de impacto. Estudantes de ciências sociais pouco versados em estatística e econometria geralmente encontram dificuldades para comprender a intuição por trás da identificação e es-timação do efeito causal em experimentos e desenhos quase-experimentais. Nesse sentido, espera-se que a utilização de exemplos práticos por meio de dados simuladas possa ser uma estratégia importante para superar a resistência inicial dos alunos em matérias com alto nível de exigência metodológica.

A replicação de artigos científicos com dados abertos também é uma boa solução que deve ser adotada em paralelo. Porém, a base simulada tem a vantagem de não violar nenhuma suposição (randomização perfeita no experimento, controle por todas variáveis de confusão no caso do pareamento, etc.) dos desenhos, já que há um domínio sobre o processo de geração de dados (*data generating process* - DGP), sendo, portanto, ideal para o primeiro contato dos estudantes com o estudo sobre avaliação de impacto.

Por fim, deve-se registrar que a função ainda está em desenvolvimento e passando por pequenas correções com o objetivo de permitir a simulação de dados adaptados para a aplicação do desenho de variável instrumental e diferenças em diferenças. Assim que esses ajustes forem finalizados, pretende-se publicar um pacote em R.

REFERÊNCIAS

ANGRIST, Joshua D; PISCHKE, J¨orn-Steffen. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, 2009.

CUNNINGHAM, Scott. Causal Inference: The Mixtape. Yale University Press, 2021.

DEBRUINE, L. faux: Simulation for Factorial Designs. doi: 10.5281/zenodo.2669586, R package version 1.1.0, https://debruine.github.io/faux/, 2021.

DOLAN, L.; GREEN, D.; LIN, W. 10 Things to know about covariate adjustment. EGAP: Evidence in Governance and Politics., n. 1, 2016.

DUARTE, Guilherme Jardim. Causalidade. In: SHIKIDA, Claudio D.; MONASTERIO, Leonardo; NERY, Pedro Fernando (eds.). Guia Brasileiro de Análise de Dados: Armadilhas e Soluções. Brasilia: Enap, 2021.

HUNTINGTON-KLEIN, N.. The Effect: An Introduction to Research Design and Causality (1st ed.). Chapman and Hall/CRC., 2021.

MEIRELES, Fernando; SILVA, Denisson e CORREA, Filipe. Simulações de Monte Carlo no ensino de Ciência Política. Revista Brasileira de Ciência Política. 2017, n. 24, pp. 223-254.

MORGAN, Stephen L.; WINSHIP, Christopher. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, 2007.