AVALIAÇÃO DE POLÍTICAS PÚBLICAS:

POR ONDE COMEÇAR?

Um guia prático para avaliação de impacto

clear avaliação África lusófona e brasil

FGV EESP

FAPEMIG

nima NÚCLEO INTEGRADO DE MONITORAMENTO E AVALIAÇÃO

FJP FUNDAÇÃO JOÃO PINHEIRO

Ficha Técnica

Governador

Romeu Zema Neto

Vice-governador

Mateus Simões

FUNDAÇÃO JOÃO PINHEIRO

Presidente

Luciana Lopes Nominato

Vice-presidente

Mônica Moreira Esteves Bernardi

Organização da publicação

Carolina Proietti Imura

Marcos Assis

Carla Bronzo

Apoio

Karina Rabelo

Raquel Carlesso - bolsista de pesquisa

Redação do guia

Autores:

Gabriel Weber Costa (FGV EESP Clear)

Thiago Noce (Seplag/MG)

Claudio Burian Wanderley (DPP/FJP)

Reinaldo Carvalho (DPP/FJP)

Revisão técnica e colaboração:

Carolina Proietti Imura

Julio A. Racchumi Romero

Projeto gráfico/Revisão textual - Assessoria de Comunicação FJP

Aline Pereira (Projeto Gráfico/Revisão)

Heitor Vasconcelos (Revisor)

Tiago Alves (Assessor-Chefe)

Nicolau Campedelli (Assessor-Chefe)

Giovanna Santos (Design Gráfico/Diagramação)

Lista de Siglas

ATE: Average treatment effect (Efeito Médio do Tratamento)

ATT: Average effect of treatment on the treated (Efeito médio do tratamento sobre os tratados)

DCPPN: Diretoria Central de Planejamento, Programação e Normas

DID: Diferença em Diferenças

ECM: Efeito causal médio

EESP: Escola de Economia de São Paulo

EML: Efeito médio local

EMT: Efeito médio do tratamento

ENEF: Estratégia Nacional de Educação Financeira

Fapemig: Fundação do Amparo à Pesquisa de Minas Gerais

FGV: Fundação Getúlio Vargas

FJP: Fundação João Pinheiro

FRD: Fuzzy regression discontinuity (Desenho tipo Fuzzy)

GEI: Global Evaluation Initiative

IDH: Índice de Desenvolvimento Humano

ITT: Intention to treat (Intenção de Tratar)

IV: Instrumental variables (Variáveis Instrumentais)

LATE: Local average eff ect treatment (Efeito Médio Local do Tratamento)

M&a: Monitoramento e da avaliação

MaInd: Mapa de Indicadores

MaPR: Mapa de Processos e Resultados

MDE: Minimal detectable effect (Efeito Mínimo Detectável)

ODS: Objetivos de desenvolvimento sustentável

PNUD: Programa das Nações Unidas para o Desenvolvimento

Pronaf: Programa Nacional de Fortalecimento da Agricultura Familiar

Protejo: Projeto de Proteção aos Jovens em Território Vulnerável

PSM: Propensity Score Matching (Pareamento por escore de propensão)

RDD: Regression discontinuity design (Método de Regressão Descontínua)

SAPP-MG: Sistema de monitoramento e avaliação de Minas Gerais

SRD: Sharp Regression Discontinuity Design (Design de Regressão com descontinuidade Sharp)

SUMÁRIO

Avaliação de políticas públicas: por onde começar?

Um guia prático para avaliação de impacto

Apresentação FJP

A Fundação João Pinheiro (FJP), desde 2019, vem aumentando seus esforços para o incentivo do monitoramento e da avaliação (M&A) no governo do estado de Minas Gerais. São diversas estratégias combinadas com o objetivo de aprimorar a elaboração das políticas públicas estaduais, monitorar e avaliar ações, projetos e programas estratégicos para o governo e para a população mineira.

A partir de 2019, iniciamos o Ciclo de Assessoramento ao Plano Plurianual de Ação Governamental - PPAG, uma parceria com a Seplag, mais especificamente com a Diretoria Central de Planejamento, Programação e Normas (DCPPN), responsável pelos processos de elaboração e revisão dos planos plurianuais. Assim, a cada ano, a equipe da FJP tem oferecido aos gestores públicos aulas abertas e cursos a distância que buscam subsidiar o planejamento e acompanhamento cada vez mais qualificado das ações e dos projetos. Essa iniciativa tem sido bem avaliada por seus participantes e, gerando efeitos importantes e ampliando conhecimentos.

Em 2020, iniciou-se a primeira edição da pós-graduação em monitoramento e avaliação de políticas públicas ofertada para servidores públicos estaduais e outros interessados. Outra iniciativa em implementação no estado é a institucionalização do Sistema de Monitoramento e Avaliação de Minas Gerais (SAPP-MG), que culminou com a publicação do decreto 48.298, de 12/11/2021. Entre 2022 e 2024, o Sistema Estadual de Políticas Públicas de Minas Gerais (SAPP-MG) realizou a avaliação de 23 programas e ações do planejamento governamental do Executivo mineiro, com diferentes enfoques e estratégias metodológicas, além de ter desenvolvido ações de formação em avaliação e um painel de indicadores de monitoramento dos ODS. Todos os planos anuais de M&A, relatórios de divulgação dos resultados das avaliações estão disponíveis na página do SAPP no site da FJP (https://fjp.mg.gov.br/sistema-estadual-de-monitoramento-e-avaliacao-de-politicas-publicas/). Para 2025, outras 26 avaliações estão em execução.

O SAPP-MG conta com um sistema de governança e envolve um comitê estadual (instância máxima e deliberativa), composto por representantes do centro de governo como secretária-geral, Secretaria de Planejamento e Gestão, Fundação do Amparo à Pesquisa de Minas Gerais (Fapemig), Controladoria Geral do Estado e FJP, um comitê executivo e núcleos setoriais.

Para que todo esse esforço se realize e seja de fato perene e sustentável, entende-se como fundamental a criação de capacidades governamentais em M&A, por meio de cursos, capacitação, reuniões técnicas e publicações que possam ser acessadas a qualquer tempo e que auxiliem as equipes técnicas do governo no momento de elaboração ou avaliações de suas iniciativas. Até o momento, a FJP já disponibilizou três publicações: o Guia Prático da Metodologia do Marco Lógico[1], o Guia Prático do Mapa de Processos e Resultados (MaPR) e do Mapa de Indicadores[2] (MaInd) e, agora, o Guia de Avaliação de Impacto, em parceria com o Centro de Aprendizagem em Avaliação e Resultados para a África Lusófona e o Brasil (FGV Clear).

O Centro de Aprendizagem em Avaliação e Resultados para a África Lusófona e o Brasil (FGV Clear) foi estabelecido em 2015 e é um dos seis centros regionais da Iniciativa Clear, todos sediados em instituições acadêmicas de relevância internacional no sul global. Os centros Clear fazem parte da Iniciativa Global de

* Os autores e organizadores do Guia agradecem à Fapemig pelo fomento na elaboração e publicação do material, no âmbito do Projeto APQ03615-23, Chamada 03/2023.

[1] O guia do Marco Lógico, em sua versão on-line e gratuita, pode ser acessado em: http://fjp.mg.gov.br/wp-content/uploads/2021/04/01.06_AvaliacaoDePoliticasPublicas_GuiaMarcoLogico_FJP.pdf

[2] O guia do MaPR e MaInd, em sua versão on-line e gratuita, pode ser acessado em: http://fjp.mg.gov.br/wp-content/uploads/2022/03/03.06_Guia-MaPR-Layout-Final.pdf

Avaliação (GEI), uma rede global, coordenada pelo Banco Mundial e pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), que visa a atender a demanda crescente por tomada de decisão com base em evidência, por parte de governos, doadores e financiadores de políticas públicas e programas.

O FGV Clear tem sede no Brasil, na Escola de Economia de São Paulo (EESP) da Fundação Getulio Vargas (FGV), e é associado ao Centro de Estudos em Microeconomia Aplicada (C-Micro). Além do Brasil, atua em Angola, Cabo Verde, Guiné Bissau, Moçambique e São Tomé. Possui uma maneira única de fortalecer capacidades em M&A por meio da colaboração com parceiros locais, incluindo agências governamentais, instituições acadêmicas e terceiro setor.

Desde 2015, o FGV Clear tem atuado para fomentar e fortalecer a agenda de monitoramento e avaliação (M&A) de políticas públicas nos países africanos de língua portuguesa e no Brasil e, assim, contribuir para o uso de evidência na gestão governamental e para a institucionalização de sistemas de M&A. Minas Gerais, por meio da FJP, tem estabelecido intercâmbio constante com a equipe FGV Clear, e este guia é um dos produtos resultantes dessa parceria. A partir do detalhado manuscrito elaborado pela equipe FGV Clear, a equipe da FJP completou o documento criando conexões com o universo das políticas públicas mineiras.

A avaliação de impacto é um dos tipos de avaliação que podem ser desenvolvidos para que se conheçam os resultados e efeitos de determinada política pública, um projeto ou programa, gerando evidências para a tomada de decisão. A iniciativa deve ser mantida, modificada ou encerrada? O tipo de avaliação a ser aplicada em determinado projeto depende de alguns fatores: tempo de execução do projeto, tipo e qualidade de dados disponíveis, entre outros.

Governos em todos os seus níveis podem se beneficiar da realização de avaliações de impacto, que permitem identificar os efeitos reais das políticas públicas e fornecer informações para ajustes e melhorias. No Executivo, essas avaliações apoiam a alocação mais eficiente de recursos e a priorização de programas com desempenho melhor. No Legislativo, contribuem para a formulação de leis e o aperfeiçoamento de marcos regulatórios. Para os órgãos de controle, oferecem subsídios para auditorias mais qualificadas, ampliando a análise para além da conformidade legal e incorporando a avaliação dos resultados das políticas públicas.

É necessário considerar algumas exigências para a realização desse tipo de avaliação. Avaliações de impacto demandam uma estrutura técnica que inclui dados adequados, tempo para a maturação dos efeitos das políticas e dos investimentos financeiros e institucionais. Métodos como experimentos aleatórios ou técnicas econométricas avançadas exigem planejamento prévio e capacidade analítica específicos. Além disso, os custos devem ser considerados, tornando essencial a seleção criteriosa das políticas a serem avaliadas.

Devido às suas exigências técnicas e metodológicas, a realização de uma avaliação de impacto requer dados específicos e, sobretudo, a possibilidade de contrafactual (conceito que será explicado a seguir). Por isso, deve ser cuidadosamente planejada e conduzida por equipe especializada e experiente. Assim, as avaliações de impacto devem estar inseridas numa estratégia mais ampla ou em um sistema de monitoramento e avaliação que forneça informações complementares sobre eficiência, eficácia e efetividade das políticas públicas.

Este guia tem como objetivo apresentar aos gestores públicos o que são e como se aplicam as avaliações de impacto e suas diferentes técnicas. Para leitores com conhecimento prévio em econometria, o material poderá contribuir para a aplicação dessas técnicas em contextos avaliativos, desde que em colaboração com uma equipe avaliadora.

Boa leitura e bom trabalho!

Introdução - FGV Clear

Um tema recorrente nas ciências sociais aplicadas (mas não só) se refere aos impactos e efeitos de fenômenos socioeconômicos. Perguntas como os impactos da pobreza sobre a violência, da desigualdade sobre o crescimento econômico ou mesmo da presença da torcida sobre o comportamento do árbitro não são triviais de serem respondidas empiricamente. Avaliações de impacto são uma ferramenta poderosa para investigar a relação causal de fatos sociais (causa e efeitos). No âmbito das políticas públicas, as avaliações de impacto têm se tornado uma estratégia importante para melhorar a eficácia e eficiência das intervenções do governo.

Ao empregar métodos econométricos, uma avaliação de impacto nos permite saber se determinada intervenção tem relação causal sobre algum resultado de interesse, como a empregabilidade da população beneficiária ou a taxa de mortalidade de uma dada região. Mais do que isso, conseguimos estimar a magnitude desse impacto, chegando a conclusões como "a política aumentou em 10% a nota dos alunos em matemática" ou "o impacto da política foi de dois pontos percentuais de redução na taxa de hospitalização por acidentes". Esses resultados são relevantes para se decidir ampliar essa política para outras regiões ou grupos específicos. Ou seja, busca-se estimar esses resultados para se prever o que ocorrerá caso a política seja ampliada. Assim, mais do que identificar especificamente os efeitos sobre grupos específicos, faz-se necessário identificar o que ocorrerá caso a mesma política seja implementada em outros grupos ou situações. Em determinadas situações, pode ser que os resultados obtidos pela avaliação de impacto apresentem evidências de que a política pública é ineficaz, ou seja, talvez seja melhor interrompê-la.

O objetivo deste guia é apresentar os principais conceitos necessários para compreender e realizar[3] avaliações de impacto e discutir os métodos de avaliação mais utilizados, incluindo as hipóteses que estão por trás de sua aplicação e as formas de estimação de cada um. Ao longo do texto, discutiremos os métodos Avaliações Experimentais, Variáveis Instrumentais, Regressão Descontínua, Diferença em Diferenças e Pareamento. Para cada um desses, será utilizada uma abordagem teórica com ênfase na intuição de cada método, seguida de um exemplo prático adaptado de avaliações reais. Ao final do guia, concluímos apresentando o conceito de testes de hipóteses, necessário para analisar a relevância estatística dos impactos estimados.

Gabriel Weber Costa

[3] Para colocar em prática uma avaliação de impacto, o leitor deve ter conhecimento prévio e intermediário de econometria. Neste guia, buscar-se-á descrever algumas técnicas úteis para resolver problemas empíricos específicos que são recorrentes durante processos avaliativos.

Como estudar o guia de avaliação de impacto

Esta publicação tem como principal objetivo apresentar, de forma resumida, algumas das principais técnicas de avaliação de impacto. As ferramentas discutidas dispõem de potencial sólido de contribuição para a análise de programas públicos por meio do uso de evidências empíricas.

O texto foi organizado da seguinte forma: primeiro, apresentam-se conceitos básicos e fundamentais no uso dessa metodologia. Em seguida, demonstram-se a importância e a utilidade das ferramentas disponíveis. Em seguida, apresentam-se as técnicas mais utilizadas nas avaliações de impacto.

Buscou-se apresentar o tema da forma mais simples possível visando a ampliar o entendimento do assunto para o maior número de interessados. Para quem deseja apenas informações mais básicas destinadas a uma visão geral sobre as técnicas, as seções principais devem ser suficientes. Já para aqueles com formação em econometria é possível se aprofundar em cada técnica por meio da leitura de quadros adicionais nas próprias seções, em que são apresentados maiores detalhamentos da formalização matemática dos assuntos tratados. Para aqueles sem tal formação econométrica, é possível pular esses quadros sem perda de compreensão do tema tratado.

Conceitos de avaliação de impacto

Apresentaremos a seguir os conceitos mais fundamentais no contexto das avaliações de impacto: inferência causal, correlação, contrafactual, resultados potenciais, grupo de tratamento e grupo de controle (ou de comparação), viés de seleção, critérios de seleção e população e amostra.

Para os não estão familiarizados com esses termos, mas que querem demandar avaliações de impacto para seus projetos, é importante que conheçam esses termos e suas aplicações. Para quem tem conhecimento sobre pesquisas quantitativas, tais termos já devem ser familiares. São fundamentais para aplicação no contexto das avaliações do tipo *experimentais* ou *quasi-experimentais*.

Inferência causal

Na avaliação de impacto, procura-se responder a uma pergunta específica de causa e efeito: qual o impacto (ou efeito causal) de um programa sobre um resultado de interesse?

A inferência causal pode ser definida como um processo de identificação e compreensão das relações de causa e efeito entre variáveis e é fundamental para compreender os efeitos de intervenções, políticas ou programas. Além disso, ajuda a prever o resultado de mudanças nas variáveis, o que pode ser especialmente útil na elaboração de experimentos e na tomada de decisões.

Dessa forma, nota-se que a inferência causal consiste num processo em que as causas são inferidas a partir de dados. Por meio da formalização desse processo, é possível definir um tratamento ou política (P) a partir de um resultado Y.

Formalmente, P causa Y, se uma mudança em P provoca uma mudança em Y, mantendo todo o demais constante, então o efeito causal é a magnitude pelo qual Y muda numa unidade em P.

$$E[Y|P = 1] - E[Y|P = 0]$$

Essa fórmula estabelece que o impacto de um programa (P) sobre uma variável de resultado (Y) é a diferença entre o resultado (Y) como emprego do programa (ou seja, quando P = 1) e o mesmo resultado (Y) sem a presença do programa (isto é, quando P = 0).

Observação: *E(x)* - O valor esperado é um conceito fundamental na teoria da probabilidade e representa o valor médio que se espera obter de um evento ou processo aleatório que se repete muitas vezes.

Correlação

O primeiro ponto a ser esclarecido é que correlação não é causalidade" (Barnard 1982, p. 387). A correlação é uma medida do grau de intensidade da associação entre duas variáveis numéricas. Quando uma variável nos fornece informações sobre outra variável, dizemos que as duas variáveis estão correlacionadas. Por outro lado, quando não há correlação, o aumento ou a diminuição de uma variável não nos diz nada sobre o comportamento de outra variável.

O conceito de correlação é usado na estatística para descrever o grau em que duas variáveis se movem na mesma direção. A correlação é geralmente resumida em um número que varia de -1 a 1.

-1 significa correlação negativa perfeita

0 significa que não há correlação

1 significa correlação positiva perfeita

Na avaliação de impacto, uma das possíveis conclusões equivocadas é assumir que uma correlação – grau de associação entre duas variáveis – implique necessariamente uma relação de causalidade, ou seja, que a intervenção de um programa tenha causado a mudança nos resultados de interesses.

Contrafactual

O objetivo da avaliação de impacto é estimar se uma intervenção específica gerou efeitos sobre indicadores de resultados considerados relevantes e qual o tamanho desses efeitos. Ao final de tal avaliação, poderemos, com certo grau de confiança (como veremos ao longo do guia), afirmar, por exemplo, que "a política aumenta a probabilidade de os beneficiários serem contratados em x%", ou "a política reduziu a evasão dos beneficiários em y%", ou mesmo que "a política não apresentou impacto estatisticamente significativo nas dimensões analisadas".

A lógica de uma avaliação de impacto está em comparar a situação real, em que a intervenção analisada realmente ocorreu, com um cenário hipotético, caso a intervenção não tivesse acontecido. A esse cenário hipotético (que, vale enfatizar, não é observável), damos o nome de **contrafactual**. Conforme Duflo et al. (2007), o contrafactual representa aquilo que teria acontecido se a política não tivesse sido implementada.

Uma vez que o contrafactual, por definição, não pode ser observado, é necessário estimá-lo a partir de técnicas econométricas. Como a estimação do impacto de uma intervenção depende exatamente da comparação entre o que foi efetivamente observado e o que foi hipoteticamente estimado, é preciso ser cuidadoso na construção de um contrafactual válido. Só assim podemos ter segurança na comparação do cenário em que a intervenção existe com o cenário em que ela é ausente.

Estimar o cenário contrafactual exige a construção de um **grupo de comparação** (ou grupo de controle). É especialmente nesse ponto que diferentes métodos de avaliação de impacto surgem, partindo de hipóteses diversas, mais ou menos adequadas a cada situação, para construir um grupo de comparação tão convincente quanto possível.

Grupos de tratamento e de comparação

Dada a impossibilidade de se observar o mesmo indivíduo quando ele é tratado e quando não, o ponto de partida de uma avaliação de impacto é a construção de um grupo de comparação adequado, que permita a melhor *comparação* possível ao grupo de tratamento. Ou seja, buscar-se-á construir o que teria acontecido ao indivíduo tratado caso esse tratamento não tivesse ocorrido.

Grupo de tratamento: conjunto de indivíduos que receberam o tratamento

Grupo de comparação: formado a partir do conjunto de indivíduos que não receberam o tratamento. Também chamado **grupo de controle**

Garantir que o grupo de comparação seja uma boa representação do contrafactual significa que será possível comparar as variáveis de interesse da avaliação observadas para os dois grupos e, a partir de sua diferença, obter uma estimativa do impacto causal da intervenção estudada. Ou seja, o grupo de controle deve efetivamente reproduzir o que teria acontecido com o grupo tratado caso a política analisada não houvesse ocorrido. A construção desse grupo para cada avaliação específica é o principal problema a ser enfrentado numa avaliação de impacto.

É possível listar alguns critérios para determinar a qualidade de um grupo de comparação (Gertler et al., 2018):

1. As características dos grupos de tratamento e controle devem ser aproximadamente iguais em média;
2. O tratamento não deve afetar o grupo de controle, seja direta ou indiretamente. Também não deve haver exposição a intervenções externas de maneira distinta entre os grupos de tratamento e controle;
3. Ambos os grupos devem se comportar de maneira similar à eventual exposição ao tratamento.

Caso não se respeite esses critérios gerais, e se eleja um grupo de comparação ruim, a avaliação de impacto gerará uma estimativa de efeito *contaminada* por outros fatores além do próprio impacto (o chamado viés de seleção). Fica claro também que essas características são função básica do que se quer avaliar. Não é necessário que os grupos sejam iguais, mas que sejam iguais em função do que se quer avaliar. Isso implica afirmar que avaliações de programas distintos requerem grupos de controle distintos.

Resultados potenciais

Uma política pública geralmente é direcionada a pessoas físicas, famílias, firmas, municípios, regiões e países, entre outros. Por exemplo, podemos pensar em treinamento para pessoas em busca de emprego, bolsa de formação para estudantes universitários, crédito bancário para agricultores, programa de financiamento para municípios com baixo Índice de Desenvolvimento Humano (IDH) etc. No âmbito da avaliação de impacto, os destinatários da política pública são denominados unidades de tratamento. São unidades de tratamento todos aqueles *passíveis* de receber a intervenção. Aqueles que são potenciais destinatários da política e efetivamente concluem o tratamento são chamados de tratados. Já os agentes que não receberam o tratamento são chamados de não tratados.

Em termos de notação, podemos utilizar para denotar uma variável binária que indica esse tratamento, que é igual 1 se o indivíduo recebeu o tratamento (ou seja, a política em análise) e igual a 0 se não o recebeu. O quadro 1 apresenta um exemplo de representação para tratados e não tratados

Quadro 1: Notação utilizada para representar tratados e não tratados

Considere a seguinte notação para tratados (que participam do programa) e não tratados (não participam do programa).

Note que temos duas variáveis: T e Y. A primeira (T) pode ser denominada como variável categórica binária. Conta com duas categorias (ausência ou presença do programa). A segunda (Y) consiste numa uma variável quantitativa (assume valores numéricos com diversas possibilidades).

T_i=1, se a unidade de análise recebeu o tratamento (participou do programa)

T_i=0, se a unidade de análise não recebeu o tratamento (não participou do programa)

Suponha que num grupo de dez pessoas, seis tenham participado do programa e quatro não. Se os indivíduos 1, 3, 4, 5, 7 e 10 participarem do programa e os indivíduos 2, 6, 8 e 9 não participarem, podemos representar os conjuntos de tratados e não tratados da seguinte forma:

Conjunto A (tratados): $\{T_1=1,T_3=1,T_4=1,T_5=1,T_7=1,T_{10}=1\}$,

Conjunto B (não tratados): $\{T_2=0,T_6=0,T_8=0,T_9=0\}$

Nesse caso, os resultados dos indivíduos que receberam o tratamento (ou seja, participaram do programa) podem ser representados por Y_1, Y_3, Y_4, Y_5, Y_7, Y_{10} enquanto os dos não tratados (que não participam do programa) podem ser representados por .

Suponha que Y seja o salário mensal dos trabalhadores após a participação num programa de capacitação. Dessa forma Y pode ser representado da seguinte forma:

- Conjunto dos tratados (participaram do programa):

$Salário_{tratados} = \{Y_1=R\$\ 2.000,Y_3=R\$\ 2.300,Y_4=R\$\ 1.700,Y_5=R\$\ 2.400,Y_7=R\$\ 1.900,Y10=R\$\ 2.800\}$

- Conjunto dos não tratados (não participaram do programa):

$Salário_{não\ tratados} = \{Y_2=R\$\ 1.500,Y_6=R\$\ 2.200,Y_8=R\$\ 1.600,Y_{10}=R\$\ 2.000\}$

No caso da situação hipotética descrita no quadro 1, é comum pensar que a avaliação do impacto do programa pode ser feita comparando-se o *salário médio dos tratados com o dos não tratados*. Essa comparação, entretanto, pode ser inconsistente. É possível que outras variáveis contribuam para explicar o salário (além do treinamento). Para lidar com esse problema, recomenda-se o uso de variáveis de controle. O ideal seria cada trabalhador que participou do programa ser comparado com outro com as mesmas características, exceto o fato de ter participado do programa. Como isso é impossível, recomenda-se a construção de um grupo de controle "artificial", caracterizado como **contrafactual**. Com isso, para uma variável de interesse qualquer denotada , podemos definir os chamados **resultados potenciais**:

- Y_i^1 é o resultado potencial do indivíduo *i* quando ele recebeu o tratamento (T_i=1). Ou seja, mostra o valor que esperamos que a variável Y tenha para o indivíduo i caso este tenha sido objeto da política analisada (tenha recebido o tratamento);
- Y_i^0 é o resultado potencial do *mesmo* indivíduo *i* quando ele não recebeu o tratamento (T_i=0). Ou seja, mostra o valor que se espera que Y mostre para o mesmo indivíduo i caso ele não tenha sido objeto da política analisada (não tenha recebido o tratamento).

A definição de **impacto causal** individual do tratamento (Angrist e Pischke, 2009) é dada pelo parâmetro beta $\beta_i = Y_i^1 - Y_i^0$. Ele consiste na diferença do valor obtido para o indivíduo que recebeu o tratamento e a estimativa de tal valor caso ele não tivesse sido tratado (participado do programa). Assim, β_i é definido para cada indivíduo da análise (por isso o subscrito *i*), o que evidencia a discussão sobre o contrafactual, em que apenas é possível observar um dos seus componentes, nunca ambos simultaneamente. O indivíduo ou é tratado (observamos Y_i^1) ou não é (Y_i^0). É impossível ser tratado e não tratado simultaneamente!

Viés de seleção

Ao considerar um grupo de comparação qualquer, podemos denominar **viés de seleção** as diferenças existentes entre o grupo de tratamento e o grupo de comparação no caso hipotético da ausência de intervenção. Um caso clássico de viés de seleção, por exemplo, é a existência de mais motivação entre os indivíduos que se matriculam para realizar um curso do que entre os que escolhem não participar. Ao utilizar o conjunto de todos os que não receberam o tratamento como grupo de comparação, sem analisar as razões para que isso tenha acontecido, é possível que encontremos impactos causais positivos sobre indicadores de interesse como renda ou empregabilidade. Na verdade as diferenças encontradas podem ser devidas ao diferencial de motivação, e não somente à participação no curso.

Suponha que existem dois tipos de pessoa na condição relativa a cursar determinado treinamento. Existiriam os que receberiam um aumento salarial com isso e aquelas que não o receberiam. Suponha que essa informação seja privada, somente cada pessoa saberia em qual grupo ela se encontra. Nesse caso, somente os do primeiro grupo se matriculariam. Ao comparar a evolução salarial de ambos os grupos, poderíamos, erroneamente, inferir que o segundo grupo teria tido o mesmo ganho salarial caso houvesse também cursado o treinamento, o que seria equivocado.

Suponha que se deseje inferir o impacto da frequência à pré-escola no aprendizado das crianças. Assim, poder-se-ia comparar o aprendizado das que frequentaram a pré-escola (grupo de tratados) com o das que não o fizeram (grupo dos não tratados). Mas isso pode ser equivocado. Vamos supor que exista falta de vagas nas pré-escolas. Vamos supor que o preenchimento dessas vagas se dá por meio de processo custoso para os pais (por exemplo, entrar de madrugada na fila de dia específico).

É possível que os pais dispostos a entrar em tal fila sejam exatamente aqueles preocupados com a educação de suas crianças. A diferença de desempenho observada entre os dois grupos poderia se relacionar com essa preocupação maior dos pais, não com a frequência à pré-escola.

O quadro 2 apresenta a formulação matemática do Viés de Seleção.

Quadro 2: Formalização matemática do viés de seleção

Mais formalmente, conforme discutido por Duflo et al. (2007), o viés de seleção pode ser representado pelos termos em vermelho na equação abaixo, que corresponde à diferença observada entre os grupos de tratamento e controle (Δ).[4]

$$\Delta = E[Y_i^1 - Y_i^0 \,|T_i=1\,] = \{E[Y_i^0 \,|T_i=1] - E[Y_i^0 \,|T_i=0]\}$$

O outro termo da equação, destacado em azul, é o impacto que se deseja estimar a partir da avaliação. Ele é conhecido como o parâmetro do Efeito Médio do Tratamento sobre os Tratados (ou ATT).

Assim, um grupo de comparação adequado será aquele capaz de tornar a diferença observada entre os grupos de tratamento e comparação () *igual* ao impacto causal desejado. Isto é, deverá ser o caso de zerar o viés de seleção, eliminando qualquer diferença entre os grupos na ausência de tratamento (de forma que $E[Y_i^0 \,|T_i=1] = E[Y_i^0 \,|T_i=0]$).

No exemplo do programa de treinamento sugerido acima, é possível que, mesmo na ausência dos treinamentos, as pessoas mais motivadas tivessem renda mais alta. Como nesse caso a motivação também foi um fator determinante da decisão de se inscrever, a comparação simples da renda entre inscritos e não inscritos incorreria em viés de seleção: $E[Y_i^0 \,|T_i=1] > E[Y_i^0 \,|T_i=0]$.

Critérios de seleção

Identificar um grupo de comparação adequado é um processo diretamente ligado à escolha do método de avaliação de impacto que será utilizado. No próximo capítulo, apresentamos as principais alternativas metodológicas, amplamente adotadas para estimar o impacto de políticas públicas.

Cada um desses métodos está ancorado em hipóteses que podem ser plausíveis a depender da situação específica sob análise. O que determina tal plausibilidade e, portanto, deve pautar a seleção do grupo de comparação, são os **critérios de seleção** da política avaliada.

Os critérios de seleção correspondem ao conjunto de regras que determinam quem serão os beneficiários da política em cada período. Segundo Gertler et al. (2018), essas regras podem estar relacionadas a três dimensões:

- Recursos disponíveis: em muitos casos, não há recursos suficientes para que uma determinada política alcance toda a população elegível. Quando isso ocorre, a decisão de qual parcela da população atender pode auxiliar na seleção de um grupo de controle a partir, por exemplo, de critérios geográficos;

[4] Na equação, assim como ao longo do guia, o operador E[x] corresponde ao valor esperado (ou esperança) de x, onde x é uma variável aleatória. O valor esperado é uma medida de tendência central, e pode ser entendido como a média populacional da variável em questão. Mais detalhes sobre os conceitos de valor esperado, variável aleatória e outros conceitos de probabilidade podem ser encontrados, por exemplo, no apêndice B de Wooldridge (2011).

- Critérios de elegibilidade: são um conjunto de características que definem qual é a população-alvo de uma política. Muitas vezes, eles podem estar relacionados a pontos de corte em variáveis observáveis, como a renda ou a idade, que podem ser explorados para selecionar um grupo de controle;
- Tempo de implementação: políticas implementadas de forma gradual, sem atender toda a população elegível no mesmo período, também geram uma oportunidade de selecionar um grupo de controle a partir dos indivíduos não atendidos em determinada fase de implementação.

É importante compreender os critérios de seleção profundamente para identificar grupos de comparação e estratégias de estimação adequados, que nos permitam obter estimativas confiáveis dos efeitos da política sob análise. Uma discussão detalhada sobre como as "regras operacionais" de um programa afetam a seleção do método de avaliação de impacto é apresentada no capítulo 11 de Gertler et al. (2018).

A tabela 1 resume como as principais metodologias de avaliação de impacto podem ser organizadas a partir dos três critérios apresentados.

Tabela 1. Relação entre os métodos de avaliação de impacto e as regras operacionais de um programa

Critérios de elegibilidade		**Excesso de demanda do programa (recursos limitados)**		**Sem excesso de demanda do programa (recursos suficientes)**	
		Ordenação [5] de acordo com índice de elegibilidade contínuo e ponto de corte	Sem ordenação de acordo com índice de elegibilidade contínuo e ponto de corte	Ordenação de acordo com índice de elegibilidade contínuo e ponto de corte	Sem ordenação de acordo com índice de elegibilidade contínuo e ponto de corte
Tempo de implementação	Implementação em fases ao longo do tempo	-Seleção aleatória -Regressão Descontínua	-Seleção aleatória -Variáveis Instrumentais -Diferença em Diferenças	-Seleção aleatória -Regressão Descontínua	-Seleção aleatória -Variáveis Instrumentais -Diferença em Diferenças
	Implementação imediata	-Seleção aleatória -Regressão Descontínua	-Seleção aleatória -Variáveis Instrumentais -Diferença em Diferenças	-Regressão Descontínua	-Variáveis Instrumentais -Diferença em Diferenças

Fonte: adaptado de Gertler et al. (2018)

[5] O índice de elegibilidade consiste numa variável contínua que pode classificar a população de interesse a partir de um ponto de corte para determinação dos indivíduos elegíveis ou não elegíveis. Por exemplo, o estudante deve obter pelo menos determinada nota de corte para entrar em determinado curso, o município deve contar com população até determinada quantidade de pessoas ou apresentar IDHM até determinado nível para se eleger para determinado programa público. Uma proposta deve alcançar determinado nível de votos em um plesbicito para ser implementada. Algumas técnicas de avaliação de impacto utilizam esse ponto de corte.

População e Amostra

A população é composta por todos os elementos (pessoas, objetos, organismos, registros) que participam do fenômeno definido e delimitado na análise do problema de pesquisa. A população tem a característica de ser estudada, medida e quantificada e deve ser delimitada claramente com relação a suas características de conteúdo, local e tempo.

A amostra é uma parte da população e pode ser definida como um subgrupo dela. Para selecionar a amostra, as características da população devem ser delimitadas primeiro.

Uma amostra representativa deve conter todas as características da população ou do universo para que os resultados sejam generalizáveis, deve ser proporcional ao tamanho da população e preferencialmente selecionada por procedimentos aleatórios/probabilísticos. Por outro lado, uma amostra não representativa consiste num subgrupo da população incapaz de refletir as características dessa população de forma fidedigna. Um exemplo clássico de amostragem não representativa é o uso de enquetes de opinião em programas de televisão. Como as pessoas que respondem essas pesquisas geralmente não refletem as características médias da população, não é possível generalizar os resultados.

Metodologias de Avaliação de Impacto

Avaliações Experimentais

O método de avaliação experimental é tido como o "padrão-ouro" das avaliações de impacto. Ele permite que as estimativas do impacto sejam tão confiáveis quanto possível (PRiME, 2019). Esse tipo de avaliação é possível quando o tratamento é distribuído de forma aleatória entre a população elegível, quando há algum tipo de sorteio para decidir quem receberá o tratamento oferecido.

Segundo Gertler et al. (2018), além de permitir uma avaliação com a metodologia mais robusta possível, tornar aleatória a designação do tratamento tem ainda as vantagens de ser de fácil comunicação ao público – conferindo transparência ao processo de seleção de beneficiários – e de ser uma maneira justa de alocar os recursos da política quando a população elegível é igualmente merecedora e não existem recursos suficientes para atender a todos. Por outro lado, é possível que existam questões éticas ou políticas que dificultem a realização de um sorteio[6].

Uma avaliação experimental precisa ser cuidadosamente planejada desde o começo da intervenção. De todo modo, quando esse planejamento é realizado com a devida antecedência e os protocolos da avaliação são seguidos (conforme discutido a seguir), a avaliação experimental resulta em estimativas confiáveis do impacto da política, deixando pouco espaço para questionamentos sobre a robustez da metodologia adotada e do resultado encontrado. Assim, toda essa confiabilidade advinda das avaliações experimentais é devida à equivalência estatística entre os grupos de tratamento e controle.

[6] Fica claro que se existem recursos para implementar um programa universal com claros impactos positivos (como por exemplo, a construção de cisternas no semiárido ou a implantação de Equipes de Saúde da Família em localidades distantes), isto deve ser feito. Restringir a aplicação para poder identificar efetivamente os impactos positivos do programa é uma violação ética grave. Ver Código de Nuremberg e Declaração de Helsinki.

A designação aleatória do tratamento faz com que os resultados potenciais e sejam *estatisticamente independentes* do status de tratamento . Caso a probabilidade de o indivíduo receber o tratamento seja completamente independente do resultado esperado (o que ocorre quando se sorteia quem será tratado), é bastante razoável esperar que os resultados obtidos pelos não tratados sejam equivalentes aos obtidos pelos tratados caso a intervenção não tivesse ocorrido. O quadro 3 mostra a formalização matemática deste conceito.

Quadro 3: Formalização matemática da independência estatística

Estatisticamente, dizemos que a independência entre os resultados potenciais e o status de tratamento garantiria que as esperanças condicionais de cada resultado potencial seriam iguais às esperanças não-condicionais:

$$E[Y_i^0 | T_i=1] = E[Y_i^0 | T_i=0] = E[Y_i^0]$$

$$E[Y_i^1 | T_i=1] = E[Y_i^1 | T_i=0] = E[Y_i^1]$$

Como exposto no Apêndice 1, as primeiras igualdades de cada uma das expressões acima são exatamente as necessárias para que a diferença observada entre os grupos de tratamento e de comparação seja igual ao impacto que se deseja estimar. Ou seja, garantir que o status de tratamento seja independentemente dos resultados potenciais leva a um grupo de comparação tão bom quanto possível, zerando o viés de seleção.

Além disso, a segunda igualdade de cada expressão gera uma característica adicional da independência entre resultados potenciais e status de tratamento. Nesse caso, a diferença observada (Δ) será igual não somente ao Efeito Médio do Tratamento sobre os Tratados (ATT), mas também ao parâmetro conhecido como Efeito Médio do Tratamento (*average treatment effect* - ATE), assim:

$\Delta = E[Y_i^1 - Y_i^0 \mid T_i=1]$	Efeito Médio do Tratamento sobre os Tratados (treatment on the treated – ATT)
$\Delta = E[Y_i^1 - Y_i^0]$	Efeito Médio do Tratamento (ATE)

A seção 1.1 apresenta a equação básica para a se estimar o impacto na avaliação experimental com a caracterização dos parâmetros de interesse, que trazem a estimativa da magnitude do impacto do programa ou ausência do impacto.

A seção 1.2 discute a importância da estimação do efeito mínimo detectável, cujo objetivo é calcular o poder estatístico de um teste para diferentes níveis de significância. Há situações em que são necessárias amostras cada vez maiores para a obtenção e conclusões consistentes sobre o impacto do programa.

Estimar o impacto

No caso de uma avaliação experimental, é possível estimar o impacto da intervenção a partir de uma regressão linear simples[7] conforme a equação:

$$Y_i = \alpha + \beta T_i + e_i$$

onde Y_i é a variável de interesse do indivíduo e é o indicador de tratamento definido anteriormente (ou seja, é igual a um para os tratados e zero para os não tratados). O termo e_i é o chamado termo de erro, e os coeficientes a serem estimados na regressão linear serão os seguintes:

- α é a média observada de Y_i para o grupo de controle;
- $(\alpha + \beta)$ é a média observada de Y_i para o grupo de tratamento;
- β é igual, portanto, à diferença de médias do resultado de interesse entre os grupos de tratamento e de controle – isto é, corresponde ao *impacto*.

Conforme discutiremos no apêndice sobre testes de hipóteses, é importante testar se o impacto estimado é *estatisticamente diferente de zero*. Esse procedimento é realizado de forma automática ao rodar uma regressão linear em softwares estatísticos como, por exemplo, Stata ou R.

Efeito mínimo detectável

Uma das principais preocupações quando se trata de avaliações de impacto relaciona-se à capacidade da avaliação de identificar o impacto causal da política em análise. Essa discussão se inicia na escolha de um grupo de comparação para estimar o contrafactual e eliminar o viés de seleção da comparação, mas continua válida mesmo após essa definição. O tamanho da amostra e a proporção de tratados, por exemplo, afetam o chamado **Efeito Mínimo Detectável** (*minimal detectable effect* – MDE), especialmente importante em avaliações experimentais.

O MDE mede, de modo geral, o quanto o tratamento precisa afetar a variável de interesse para que seja possível estimar esse impacto conforme discutido por Duflo et al. (2007). Valores mais baixos de MDE significam que a avaliação de impacto possui probabilidade maior de identificar o impacto (caso ele exista) mesmo que o efeito real seja pequeno, e por isso são desejáveis. Por outro lado, caso o desenho da avaliação forneça valores elevados de MDE, é maior a possibilidade de que o ato de estimar não consiga identificar o impacto da intervenção estudada (ou seja, o impacto estimado será estatisticamente igual a zero). O quadro 4 apresenta os cálculos estatísticos referentes ao efeito mínimo detectável.

[7] Uma boa referência sobre o uso de regressão linear, a partir do método de Mínimos Quadrados Ordinários, é o capítulo 2 de Wooldridge (2011). Outra sugestão é o livro de Flávia Chein (2019).

Quadro 4: Formalização matemática do efeito mínimo detectável

Em linhas gerais, no caso canônico de avaliações experimentais, o *minimal detectable effect* (MDE) pode ser calculado com a seguinte expressão:

$$MDE = (t_{1-\kappa} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

onde:

- t_{1-k} e t_a são os valores da distribuição t de Student para o poder do teste e o nível de significância escolhidos pelo avaliador. É usual que sejam utilizados os valores de poder igual a 80% ($t_{0,2}$ = 0,84) e nível de significância igual a 5% ($t_{0,05}$ = 1,96). Um poder de teste mais baixo e um nível de significância mais alto diminuem o valor de MDE, mas têm consequências para o processo de inferência[8];
- σ^2 é a variância populacional do termo de erro e_i. Na maior parte das aplicações, esse valor precisa ser estimado ou padronizado para 1;
- N é igual ao tamanho da amostra de avaliação. Tamanhos de amostra maiores diminuem o MDE;
- P é a proporção de indivíduos tratados na amostra. Proporções de tratados mais próximas de 50% diminuem o valor de MDE.

Alternativamente, é possível rearranjar a expressão do MDE para obter o tamanho de amostra necessário para atingir um valor desejado de MDE, que pode então ser definido de antemão. Assim:

$$N = \frac{(t_{1-\kappa} + t_{\alpha})^2\sigma^2}{P(1-P)\,MDE^2} * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

De todo modo, é essencial levar em consideração os valores de MDE ao planejar uma avaliação experimental, já que problemas como tamanho de amostra e proporção de indivíduos tratados não podem ser alterados após o ato de tornar aleatórios os indivíduos já ter acontecido.

Por fim, há algumas situações em que o cálculo do MDE se torna mais complexo. Primeiro, em desenhos de avaliação em que o sorteio é realizado em blocos (estratos) ou de forma agregada (com o uso de *clusters*). Em segundo lugar, nos casos em que o sorteio não é seguido à risca, quando nem todos os indivíduos sorteados para o grupo de tratamento recebem o benefício e/ou alguns indivíduos sorteados para o grupo de comparação o recebem[9]. E, por fim, quando são incluídas variáveis dependentes na análise de regressão. Esses casos especiais são detalhados por Duflo et al. (2007) e por Djimeu e Houndolo (2016). O conceito de efeito mínimo detectável também é discutido no capítulo 15 de Gertler et al. (2018).

[8] Os conceitos de poder do teste e nível de significância serão discutidos no Apêndice 8.

[9] Quando isto ocorre, é possível que a aleatorização citada não ocorra. Isto porque existiriam razões para que as unidades sorteadas para serem tratadas não o fossem ou para que as unidades sorteadas para não serem tratadas acabassem por receber o tratamento.

Exemplo 1: Aplicação da técnica de seleção aleatória

No artigo "*Returns to capital in microenterprises: evidence from a field experiment*", Mel et al. (2008) estimam o efeito médio do tratamento (ATE) em uma avaliação do impacto de uma intervenção que buscava aumentar o estoque de capital de microempreendedores no Sri Lanka.

Intervenção avaliada

Os autores selecionaram 408 microempresas no Sri Lanka, dividindo-as em cinco grupos de maneira aleatória, um grupo tendo sido tomado como o de controle e quatro grupos tendo recebido diferentes tipos de benefícios: a) 100 dólares em equipamentos, b) 200 dólares em equipamentos, c) 100 dólares em dinheiro d) 200 dólares em dinheiro.

Metodologia de avaliação de impacto

Como o status de tratamento designado pelo processo de sorteio foi seguido à risca, os autores puderam estimar o ATE do aumento do capital sobre as variáveis de interesse: estoque de capital, lucro real e horas trabalhadas pelo proprietário.

Resultados

A avaliação revelou que:

- Houve um aumento estatisticamente significante no estoque de capital para todos os tipos de tratamento;
- Houve um aumento estatisticamente significante entre 2% e 14% no lucro real da empresa para três dos grupos de tratamento (com exceção da intervenção de 100 dólares em equipamentos);

Houve um aumento significativo nas horas trabalhadas pelo proprietário, entre os beneficiados com 100 dólares (em dinheiro ou em equipamentos).

Variáveis instrumentais

Retomando o arcabouço dos *resultados potenciais*, podemos pensar no problema do *viés de seleção* como um exemplo de **endogeneidade**[10] (Imbens e Wooldridge, 2007). Quando o status de tratamento dos indivíduos beneficiados por uma política não provém de um processo aleatório, não é possível garantir que a diferença observada entre os grupos de indivíduos tratados e não tratados será igual ao impacto causal da política. Isso acontece devido ao viés de seleção, por exemplo, que representa a existência de fatores omitidos. A princípio, não conseguimos considerá-los em nossas análises.

Como descrito anteriormente, uma possível fonte de viés de seleção é a **diferença de motivação.** Ela pode estar mais presente no grupo de indivíduos tratados do que entre os não tratados, afetando simultaneamente a propensão a participar da política entre os indivíduos motivados e o resultado de interesse da avaliação.

Outro caso de viés de seleção é quando os indivíduos participam de determinada política a partir de uma escolha baseada no próprio resultado de interesse. Assim, aulas de reforço escolar

[10] Endogeneidade, entretanto, também ocorre por outros fatores além do viés de variável omitida (por exemplo, simultaneidade, erro de medida, erro de especificação do modelo, entre outros). Endogeneidade por simultaneidade, por exemplo, acontece quando uma terceira variável impacta simultaneamente as duas variáveis analisadas, gerando uma falsa correlação entre elas. Isto é bem comum em algumas áreas específicas das políticas públicas, como segurança e saúde.

são dadas aos alunos com piores notas (ou seja, com maior dificuldade de obterem notas melhores). Políticas de prevenção de doenças são aplicadas nos grupos com maior probabilidade de ficarem doentes. Transfere-se renda exatamente para os mais pobres (ou seja, com menor capacidade de conseguir renda maior). Ou seja, como construir um grupo de controle quando quem é o beneficiário da política é exatamente quem apresenta os piores resultados da variável que se quer tratar? Esses são exemplos da chamada *endogeneidade* - o grupo de tratamento é escolhido de acordo com o problema que se quer resolver com a política em questão.

O método de variáveis Instrumentais (*instrumental variables* – IV) busca resolver o problema de endogeneidade – mais especificamente em nosso contexto, o viés de seleção – a partir de uma fonte de variação exógena. A chamada variável instrumental (ou simplesmente instrumento) deve ser relacionada ao status de tratamento dos indivíduos, mas não deve afetar o resultado de interesse de qualquer outra maneira.

Para o entendimento sobre o papel das variáveis instrumentais, imagine um modelo de regressão no qual o rendimento do trabalho depende do número de anos de estudo. Estimar esse modelo sem considerar outras variáveis que influenciam o comportamento da variável de resposta (renda do trabalho) pode gerar problemas de viés. No caso em questão, existem variáveis não mensuráveis, como a habilidade inata, que contribuem para a variação do salário. Nesse caso, a solução inclui a utilização de algum instrumento, como o uso da variável "escolaridade da mãe", que se relaciona tanto com a variável explicativa (escolaridade do trabalhador) como com a variável explicada (rendimento também do trabalhador). Ela deve ser "exógena", de forma a não introduzir uma nova fonte de endogeneidade à análise. Variáveis exógenas são aquelas determinadas fora do modelo, não são definidas pelas variáveis analisadas do próprio modelo. No caso em questão, dificilmente a escolaridade da mãe foi definida a partir dos possíveis impactos que ela teria sobre a escolaridade ou a renda dos possíveis filhos que a mãe teria no futuro. Outro exemplo é a curva de demanda de um bem, na qual a quantidade demandada depende do preço. Nesse caso, tanto o preço quanto a quantidade são consideradas variáveis endógenas (o preço é a variável de decisão, pois a variação no preço implica na variação na quantidade). Variáveis que influenciam na quantidade demandada, como a renda, podem ser consideradas exógenas.

Respeitadas essas condições, a variável instrumental é incorporada à avaliação de impacto e conseguimos estimar o impacto causal da política para um grupo específico de indivíduos, como veremos a seguir.

Hipóteses

As condições descritas acima para a existência de uma variável instrumental adequada são as chamadas **hipóteses do método**. Elas devem ser respeitadas simultaneamente para que o método de Variáveis Instrumentais seja válido e possam ser descritas como a seguir:

1. **Alocação independente:** a variável instrumental é exógena, ou seja, é definida fora do modelo, como explicado anteriormente;
2. **Restrição de exclusão:** a variável instrumental afeta o resultado de interesse apenas por intermédio de sua relação com o status de tratamento[11];
3. **Monotonicidade:** a variável instrumental afeta todos os indivíduos no mesmo sentido.

[11] O uso de variáveis instrumentais é bem mais amplo do que o tratamento específico de viés de seleção, podendo ser utilizado para diversos problemas de endogeneidade. O exemplo utilizado relativo à educação mostra isto.

As duas primeiras condições buscam resolver exatamente o problema da endogeneidade descrito (por óbvio, caso não sejam válidas, a endogeneidade se mantém). A última condição garante que o resultado encontrado seja consistente. Queremos identificar o impacto de uma variável X sobre Y usando como Z instrumento. Mas se Z tem uma relação direta com X em uma parte da amostra (aumento de Z aumenta X e vice-versa) e inversa com outra parte (aumento de Z diminui X e vice-versa), o resultado fica inconclusivo (no primeiro grupo, o aumento de Z levaria a identificar os impactos sobre Y de um aumento de X. No segundo grupo, o aumento de Z levaria a identificar os impactos sobre Y de uma diminuição de X).

Avaliações experimentais com variáveis instrumentais

Encontrar uma variável instrumental que respeite as hipóteses necessárias do método é bastante desafiador. Alguns exemplos de instrumentos clássicos utilizados na literatura incluem índices pluviométricos ou variações no tipo de colonização entre determinadas regiões. Em alguns contextos, essas variáveis são ditas plausivelmente exógenas, mas ainda assim não é impossível encontrar potenciais fontes de endogeneidade que gerem dúvidas sobre a robustez do método.

Entretanto, um contexto em que naturalmente temos acesso a uma variável instrumental exógena é o dos experimentos aleatórios. Nesses casos, o instrumento de fato foi sorteado, de forma que não precisamos nos preocupar com qualquer endogeneidade.

Utilizamos uma variável instrumental sorteada para conseguir identificar o impacto causal quando o sorteio não foi acatado de maneira perfeita, quando pelo menos alguns dos indivíduos da amostra não cumpriram o status que lhes foi designado. Isso acontece quando existe, por exemplo, alguma falha de fiscalização, permitindo que pessoas do grupo de comparação recebam o benefício, ou quando simplesmente não é possível controlar quem participa ou não de determinada intervenção. A situação em que o status do sorteio não é cumprido para toda a amostra é conhecida como *cumprimento parcial*.

É possível também planejar a própria avaliação experimental incorporando o cumprimento parcial no desenho do experimento. Isso acontece nas situações em que a participação em um programa é tornada disponível para toda a amostra (por exemplo, serviços públicos que não devem excluir o acesso de nenhuma parcela da população, intervenções online de amplo acesso etc.). Nesses casos, para possibilitar uma avaliação de impacto, é possível realizar uma *aleatorização de encorajamento*. Nela se proporciona algum incentivo ao grupo de pessoas sorteadas para que elas participem. À medida que o incentivo aumenta a proporção dos que recebem o tratamento, é possível identificar o impacto. Assim como antes, a estratégia de estimação utilizará a alocação aleatória do incentivo como variável instrumental. Os exemplos 2 e 3 apresentam aplicações da técnica de variáveis instrumentais. O quadro 5 mostra a formalização matemática do uso de variáveis instrumentais.

Exemplo 2: Aplicação da técnica de seleção variáveis instrumentais

No artigo "*Rescuing At-Risk Youth: Experimental Evidence from a Human Capital Investments Program in Brazil*", Barros et al. (2019) estimam o Efeito do Tratamento em uma avaliação de impacto de um programa de formação e inclusão social de jovens expostos à situação de violência, utilizando uma amostra do Rio de Janeiro.

Intervenção avaliada

O Projeto de Proteção aos Jovens em Território Vulnerável (Protejo) é um dos programas administrados pelo Ministério da Justiça para prevenir e controlar crimes no Brasil. O programa é voltado para as pessoas com maior vulnerabilidade social e tem como objetivo aumentar o senso de cidadania, protegê-los da violência precoce e promover educação e trabalho. O protejo atua ofertando cursos que combinam elementos de educação vocacional e técnica, além de atividades para desenvolver habilidades socioemocionais.

Para a avaliação, que focou na edição de 2010 do programa, foram consideradas 19 comunidades no Rio de Janeiro. A seleção dos participantes foi realizada utilizando-se um sorteio que levava em consideração três critérios: sexo, vulnerabilidade social e educação. A partir dessas características, foram formados estratos; caso não houvesse excesso de demanda (mais inscritos do que vagas) em um estrato, todos os participantes receberiam a oferta do programa. Aos selecionados, foram ofertadas 800 horas de atividades divididas em 12 módulos entre novembro e julho do ano seguinte. Além disso, 400 horas foram dedicadas para o desenvolvimento socioemocional.

Metodologia de avaliação de impacto

Apesar da utilização de um sorteio para alocar as vagas do programa, a adesão não foi perfeita, houve cumprimento parcial: alguns sorteados para o grupo de tratamento não participaram do programa, outros, sorteados para o grupo de comparação, conseguiram acessar o tratamento. Assim, os autores puderam estimar o Efeito Médio Local do Tratamento, considerando como resultados de interesse variáveis relacionadas à formação de família, interação social, vitimização, engajamento na comunidade, educação, habilidade não cognitiva e mercado de trabalho.

Resultados

Alguns dos resultados da avaliação são:

- Aumento de 24 pontos percentuais (p.p.) na probabilidade de se estar empregado no setor formal em 2014 e 13 p.p na probabilidade em 2015;
- Aumento de 15 p.p. na probabilidade de se ter filho;
- Aumento de 11,5 p.p na probabilidade de sofrer bullying;
- Entre os alunos que estavam na escola e saíram em 2012, houve um amento de 9.9 p.p. na probabilidade de voltar a estudar em 2013;
- Sem mais efeitos significativos sobre variáveis de educação ou relacionadas a habilidades cognitivas.

Exemplo 3: Aplicação da técnica de variáveis instrumentais

No artigo "*The impact of high school financial education: evidence from a large-scale evaluation in Brazil*", Bruhn et al. (2016) estimam o impacto da oferta de aulas de educação financeira durante o Ensino Médio no Brasil.

Intervenção avaliada

O governo brasileiro lançou em 2010 a Estratégia Nacional de Educação Financeira (Enef) com o objetivo de promover a educação financeira no país. Um dos seus primeiros projetos foi um programa de educação financeira para as escolas públicas de ensino médio. O objetivo do curso era ser multidisciplinar e ser usado como tema de fundo em outras matérias. Os livros didáticos utilizados nesse curso contêm material suficiente para uma carga horária entre 72 e 144 horas de ensino.

Apenas cinco estados (Minas Gerais, São Paulo, Rio de Janeiro, Ceará e Tocantins) e o Distrito Federal aderiram ao projeto piloto. Para participar, as escolas desses estados deveriam se voluntariar; no total 815 aceitaram participar. Outras 101 escolas desses estados parceiras em outro estudo também foram consideradas na pesquisa. Entre todas elas, o recebimento do programa de educação financeira foi alocado de forma aleatória.

Metodologia de avaliação de impacto

Apesar de ter sido feito um sorteio para definir quais escolas deveriam receber o programa, houve cumprimento parcial: algumas escolas que foram sorteadas para o grupo de tratamento não receberam o programa, enquanto outras, sorteadas para o grupo de controle, acabaram recebendo o tratamento. Dessa forma, os autores decidiram estimar o Efeito da Intenção de Tratar (ITT) considerando variáveis de interesse diversos aspectos dos alunos: 1) conhecimento financeiro, medido por meio de um teste realizado pelas escolas; 2) taxa de aprovação; 3) comportamento e atitudes financeiras.

Resultados

A avaliação revelou que:

- Houve aumento de 5% a 7% no conhecimento financeiro dos alunos de escolas sorteadas para o grupo de tratamento;
- A taxa de aprovação desses alunos aumentou em 1,2 ponto percentual;
- Foram identificados efeitos positivos sobre diversas características relacionadas ao hábito de poupar dos alunos (se o aluno se considera poupador, se está guardando dinheiro para comprar algo, se tem uma poupança formal, quanto está poupando), além de efeitos positivos em outros comportamentos financeiros.

Quadro 5: Formalização matemática da estimação do impacto através do uso de variável instrumental com efeitos homogêneos do tratamento

Quando se deseja estimar o efeito médio do tratamento com existências de outros fatores, além das características observáveis presentes no vetor X_i , que afetam a participação do programa e os resultados potenciais, precisamos de uma variável exógena Z_i que afeta a decisão de participação e não está correlacionada com fator algum não observável relacionado ao resultado potencial.

Supondo que o efeito homogêneo do tratamento, temos:

$$Y_i = \alpha + \beta T_i + e_i$$
$$T_i = \{1, se\ \gamma + \delta Z_i + \varphi_i \geq 0\ 0, \qquad caso\ contrario$$

em que T_i é igual a 1 se o indivíduo recebeu tratamento, e 0 se o indivíduo é não tratado. Assume-se que a unidade tratada dada certa condição depende de variáveis observáveis e não observáveis. Além disso, assumimos que $Cov(Z_i, e_i) = 0$ e $Cov(\pi_i, e_i) \neq 0$, isto é, não existe relação linear entre o instrumento e o termo aleatório da equação principal e que os termos aleatórios das duas equações são relacionados.

Para estimar esse modelo, podemos usar o método de mínimos quadrados em dois estágios. No primeiro estágio, estimamos um modelo de probabilidade linear entre o com e obtemos:

$$\widehat{T}_i = \hat{\gamma} + \hat{\delta} Z_i$$

No segundo estágio, estimamos uma regressão linear dos resultados de interesse , tendo:

$$\widehat{Y}_i = \hat{\alpha} + \hat{\beta} \widehat{T}_i$$

Nas equações anteriores, a variável Z_i não afeta diretamente os resultados de interesse Y_i, mas só por meio da sua relação com a participação ou não no tratamento.

Assumindo que o feito de tratamento é homogêneo, isto é, $\beta = \beta_i$ para todo indivíduo i. No caso de tratamento homogêneo, o resultado do indivíduo depende apenas da sua participação ou não no programa, e não está relacionada a como a participação no programa é afetada pelo instrumento Z_i.

No caso de tratamento homogêneo, o efeito médio do tratamento (EMP) é igual ao efeito médio do tratamento sobre os tratados (EMPT), assim:

$$EMP = EMPT = \beta$$

Observação: *quando as unidades supõem que os ganhos de participação podem diferir para certos grupos, eles irão levar em consideração essa informação na hora de decidir se participam ou não no programa. Dito de outra forma: o efeito de tratamento não é homogêneo e o ato de estimar deve levar em consideração os ganhos individuais* β_i *(Ver GERTLER, P. J. et al., 2017).*

Regressão descontínua

Na ausência de um sorteio que determine a participação dos indivíduos em uma determinada política a ser avaliada, ainda é possível avaliar o impacto utilizando métodos não experimentais. Conforme discutimos na seção "Conceitos de Avaliação de Impacto", a escolha do método deve ser baseada nos *critérios de seleção* da política em análise.

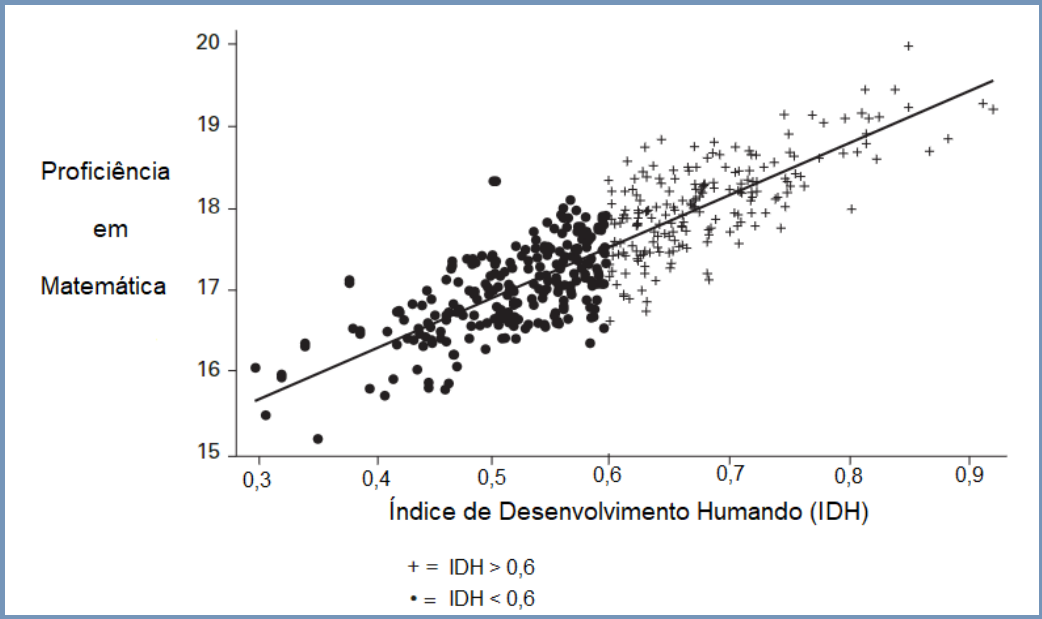
O método de Regressão Descontínua (RDD) pode ser utilizado para estimar o impacto causal de uma política quando há critérios de elegibilidade explícitos estabelecidos a partir de um índice contínuo e um ponto de corte bem definido (Imbens e Lemieux, 2008). Dessa forma, os indivíduos são ordenados com base nesse índice, conhecido como a **variável de atribuição** (*running variable* em inglês), de forma que apenas aqueles acima (ou abaixo) do ponto de corte são considerados elegíveis. Um exemplo usual de aplicação do RDD é o de exames de admissão em escolas e universidades. Nesses casos, um mecanismo de admissão possível é aquele em que os alunos são ordenados com base em suas notas (a variável de atribuição) e somente aqueles com notas acima de um determinado valor (o ponto de corte) são aprovados (ou seja, comparar-se-ão os alunos com notas ligeiramente acima dessa nota de corte com aqueles ligeiramente abaixo). Programas de combate à pobreza cujas famílias beneficiárias devam ter renda familiar per capita até determinado valor também são um bom exemplo (comparar-se-iam as famílias com renda familiar per capita ligeiramente menor deste valor de corte com aquelas ligeiramente maior).

A intuição do RDD está em comparar indivíduos logo acima e logo abaixo desse ponto de corte bem definido. Ao considerar que todos os indivíduos suficientemente próximos do ponto de corte são similares entre si, a diferença observada de resultados entre os grupos de elegíveis e não elegíveis poderá ser atribuída ao impacto da política (Lee e Lemieux, 2010). Vale notar, entretanto, que o impacto estimado a partir do método de regressão descontínua não é generalizável para toda a população e só é válido para a subamostra de indivíduos na vizinhança do ponto de corte. Os exemplos 4, 5 e 6 trazem aplicações do método de regressão descontínua. O quadro 6 mostra a formalização matemática deste método.

Exemplo 4: Aplicação do método de regressão descontínua

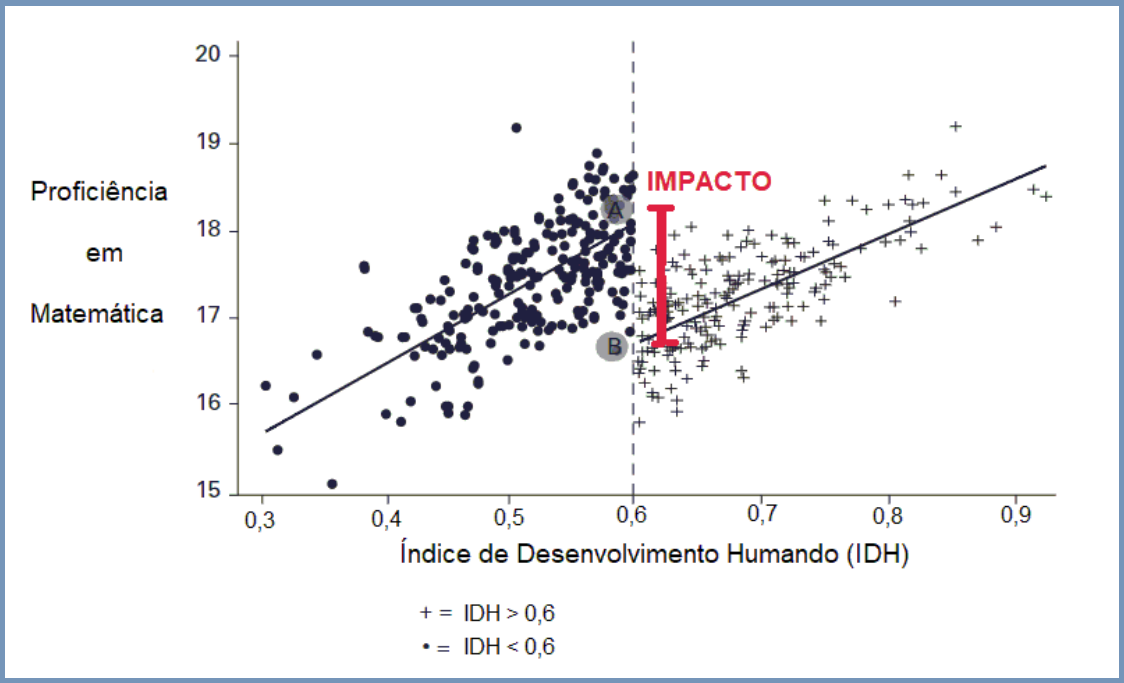
Imagine um programa social que vise a melhorar o desempenho dos estudantes de municípios mais carentes. O programa destina-se à alocação de recursos a serem gastos em educação em municípios com Índice de Desenvolvimento Humano (IDH) menores. Antes do início do programa, é provável que cidades mais carentes apresentem desempenhos piores, como na figura 1. No caso, o critério de elegibilidade é o baixo IDH. Foi escolhido como ponto de corte o escore 0,60. O grupo de municípios com escore de 0,59 provavelmente é muito semelhante ao grupo de municípios com 0,61 em diversos fatores. Nesse caso o que os diferencia é o fato de participar ou não do programa.

Figura 1 – Proficiência em Matemática versus IDH (menor e maior)



Após a implementação do programa, os avaliadores poderiam usar uma regressão descontínua para estimar o impacto do mesmo (figura 2). Operacionalmente, calcula-se a diferença entre os resultados, como o desempenho em matemática, dos municípios localizados em ambos os lados do ponto de corte de elegibilidade (no caso, o IDH de 0,6). Os municípios considerados de alto IDH (acima de 0,6) podem ser utilizados como grupo de comparação, o que gera uma boa estimativa do contrafactual. Note que os municípios eram muito parecidos no início do estudo (linha de base). Dessa forma, é razoável imaginar que a diferença de desempenho tenha se dado em função da implementação do programa.

Figura 2 – Proficiência em Matemática versus IDH (menor e maior) após a intervenção



Exemplo 5: Aplicação do método de regressão descontínua

No artigo "*Voting technology, political responsiveness and infant health: evidence from Brazil*", Fujiwara (2015) utiliza o método de regressão descontínua para estimar o impacto da implementação da urna eletrônica nas eleições brasileiras sobre a quantidade de votos válidos computados nas eleições.

A hipótese do autor é que a urna pode auxiliar principalmente as pessoas com menores níveis de educação, pois a nova tecnologia não permite rasura, facilitando o voto ao apresentar a foto do candidato quando o número inserido é correto e mostrando uma mensagem clara de que o número não é válido quando isso ocorre.

Intervenção avaliada

Nas eleições de 1998, o Brasil passou a adotar o voto por meio de urna eletrônica. Entretanto, devido ao ainda baixo número de urnas disponíveis, apenas as cidades com mais de 40.500 eleitores puderam utilizar essa nova tecnologia, enquanto as cidades menores continuaram no sistema antigo do voto em papel.

Metodologia de avaliação de impacto

Devido à forma como a urna eletrônica foi implementada, criou-se arbitrariamente um ponto de corte bastante claro, em que apenas cidades com um número de eleitores acima de 40.500 receberam a nova tecnologia. Isso proporciona a utilização do método de regressão descontínua do tipo *sharp*, já que todas as cidades acima do ponto de corte foram tratadas e nenhuma abaixo dele recebeu o tratamento. A população de eleitores por município é, portanto, a variável de atribuição nesse estudo, com o ponto de corte igual a 40.500. Então, a amostra final é composta por municípios com populações de eleitores logo acima e logo abaixo desse número.

A principal variável de interesse no estudo foi o número de votos válidos. O voto é considerado válido caso ele possa ser atribuído a um determinado candidato ou partido. Por outro lado, ele é considerado não valido em caso de rasura ou nome/número inexistente (no voto em papel) ou caso seja digitado número inexistente ou se aperte a tecla branco (no voto na urna eletrônica).

Resultados

Alguns resultados dessa avaliação são:

- A porcentagem de votos válidos aumentou 12 pontos percentuais com o uso de urnas eletrônicas;
- Análises complementares mostraram que essa maior porcentagem de votos válidos foi proveniente majoritariamente de eleitores com menor nível educacional;
- A adoção da urna eletrônica também foi associada a um aumento nos gastos públicos em saúde, o que é condizente com maior representatividade da população pobre nas eleições.

Exemplo 6: Aplicação do método de regressão descontínua

No artigo "*Lighting and Homicides: Evaluating the Effect of an Electrification Policy in Rural Brazil on Violent Crime Reduction*", Arvate et. al. (2018) utilizam o método de regressão descontínua para estimar o efeito da iluminação pública sobre a incidência de crimes violentos no Brasil.

Intervenção avaliada

O Programa Luz Para Todos é um programa federal que começou em 2003 com o objetivo de expandir e universalizar o acesso a eletricidade. Na teoria, para serem elegíveis ao programa, os municípios deveriam atender a alguns critérios estabelecidos pelo Ministério de Minas e Energia. Entretanto, na prática, o critério que acabou predominante para definir a implementação nos municípios foi o de cobertura elétrica, no qual municípios onde menos de 85% da população possuíam acesso a eletricidade foram considerados elegíveis.

Metodologia de avaliação de impacto

As regras de elegibilidade do programa proporcionaram a utilização do método de regressão descontínua. A variável de atribuição foi a taxa de cobertura elétrica prévia da população. Entretanto, como o programa previa ainda outros critérios e, portanto, a cobertura elétrica não determinava por completo o status de tratamento dos municípios, o RDD empregado foi do tipo *fuzzy*. Assim, para estimar o efeito da iluminação pública sobre a taxa de homicídios, os autores utilizaram os municípios que estavam um pouco acima do ponto de corte como grupo de comparação daqueles que estavam abaixo do ponto de corte de 85%.

Resultados

A avaliação revelou que:

- Houve grande redução na taxa de homicídios em estradas rurais na Região Nordeste. O efeito foi equivalente à redução de 91,7 homicídios a cada 100.000 habitantes, considerando um município qualquer do Nordeste que elevasse o acesso à energia elétrica de 0 para 100% da população;
- A concentração do efeito encontrada na Região Nordeste foi considerada consistente com o objetivo do programa, que atuava fortemente na região.

Quadro 6: Formalização matemática da estimação de impacto através de regressão descontínua

Como se sabe, a Regressão descontinua (RDD) consiste em um experimento natural que nos fornece variáveis instrumentais para estimar o efeito causal local do tratamento e existem várias formas de se chegar a estimativas graficamente e via regressões lineares locais e regressões paramétricas. Além disso, existem dois desenhos Sharp (SRD) e Fuzzy (FRD).

Para estimar o impacto usando esse método, além da variável de resultado e a variável que indica se a unidade pertence ao grupo de tratamento , define-se, , *forcing variable*, que

determina as unidades que serão designadas ao tratamento; variável *dummy* que define qual lado do limiar se encontra o indivíduo, ou seja, qual será o limite para "designar a unidade ao tratamento" ("c" também é denominado ponto de descontinuidade),

a) No desenho do tipo *Sharp* (SRD), temos

$$Z_i = \{1\ se\ X_i \geq c\ \ 0, \qquad se\ X_i < c$$

Considerando o "compliance" é perfeito temos então:

$$T_i = \{1\ se\ X_i \geq c\ \ 0, \qquad se\ X_i < c$$

Logo o efeito causal médio (ECM) no ponto de descontinuidade é dado por

$$ECM\ =\ E[Y_i(T_i = 1) - Y_i(T_i = 1)|X_i = c]$$

Na prática, para calcular o efeito médio do tratamento (EMT), podemos usar uma regressão linear local, a qual consiste numa diferença de médias de Y_i de observações à direita e à esquerda de dentro de um intervalo $X_i \in [c - h;\ c + h]$. .

Assim o modelo de regressão, dado que $Z_i = T_i$, será

$$Y_i = \alpha + \beta T_i + \gamma_1(X_i - c) + \gamma_2 T_i(X_i - c) + e_i; \quad i: c \leq X_i \leq c + h$$

Em que o efeito médio do tratamento em $X_i = c$ é β e deve ser suficientemente pequeno

b) No desenho do tipo *Fuzzy* (FRD), temos que o tratamento é determinado parcialmente pela descontinuidade em Z_i, dado que probabilidade de receber tratamento não muda de 0 para 1 no ponto de corte , mas acontece apenas um salto na probabilidade de receber tratamento no ponto em que $Z_i = C$. Logo, para estimar o efeito médio local, a razão entre o salto observado nos resultados em torno do ponto corte e o salto observado na probabilidade de participação no programa:

$$EML = \frac{E[Y_i|X_i = x]\ - E[Y_i|X_i = x]}{E[T_i|X_i = x]\ - E[T_i|X_i = x]}$$

No entanto, na pratica é possível usar a regressão linear para estimar o EML:

O primeiro estágio consistiria em estimar

$$T_i = \theta + \rho Z_i + \lambda_r(X_i - c)Z_i + \lambda_I(X_i - c)(1 - Z_i) + \eta_i\ ;\ i: c \leq X_i \leq c + h$$

E o segundo estágio:

$$Y_i = \alpha + \beta T_i + \gamma_r(X_i - c)Z_i + \gamma_I(X_i - c)(1 - Z_i) + e_i\ ;\ i: c \leq X_i \leq c + h$$

Em que o efeito médio do tratamento em $X_i = c$ é β e h deve ser suficientemente pequeno

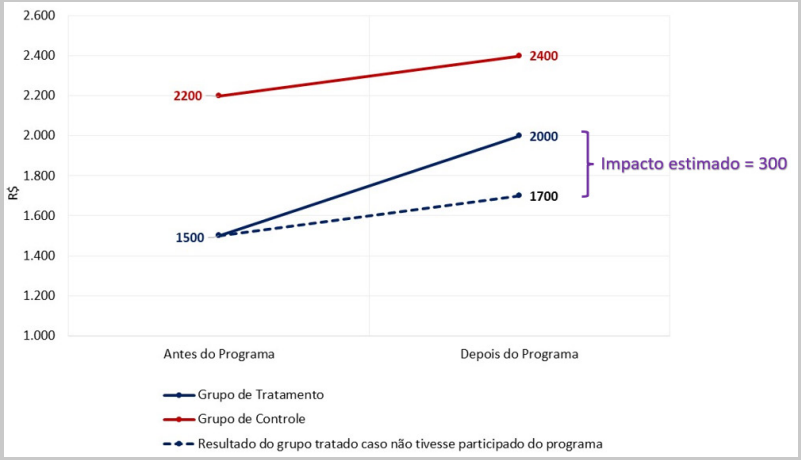
Diferença em diferenças

Tanto no caso das avaliações experimentais, como no caso da regressão descontínua, somos capazes de estimar o *impacto causal* da política em análise ao adotar métodos que exploram seus *critérios de seleção* – quais sejam, um procedimento de sorteio e um ponto de corte claro em um índice de elegibilidade, respectivamente. Entretanto, nem sempre temos acesso a um índice contínuo que determina a elegibilidade dos indivíduos ou sabemos com exatidão de que forma uma determinada política selecionou seus beneficiários. Nesses casos, ainda é possível aferir o impacto causal desejado pela utilização de métodos que impõem hipóteses adicionais sobre a estimação (Gertler et al., 2018)

O método de diferença em diferenças (DID) estima o impacto de uma política ao comparar os grupos de indivíduos tratados e não tratados em dois momentos do tempo: antes e depois de sua implementação. Para isso, além da necessidade de termos acesso a dados pré-intervenção (inclusive para o indicador de resultado de interesse), o DID supõe que esses grupos distintos de indivíduos se comportariam da mesma maneira na ausência da política – mais especificamente, apresentariam a mesma evolução do indicador de resultado ao longo do tempo. Os exemplos 7 e 8 apresentam aplicações do método de diferenças em diferenças. O quadro 7 mostra as hipóteses essenciais para a utilização desse método. Já o quadro 8 mostra sua formalização matemática.

Exemplo 7: Aplicação do método de diferenças em diferenças

Suponha que determinada prefeitura crie um programa de capacitação para jovens de famílias mais pobres. A ideia é que jovens mais qualificados poderiam melhorar sua remuneração. Suponha que um analista da prefeitura resolva avaliar o impacto do programa. Imagine que, ao selecionar um grupo de jovens que passou pelo treinamento (grupo de tratamento) e outro grupo que não passou (grupo de controle), o salário médio dos tratados seja de R$ 2.000, enquanto o dos não tratados seja de R$ 2.400. O analista pode concluir, ingenuamente, que o programa não funcionou. Uma forma de lidar com esse tipo de situação é medir o salário médio dos jovens do grupo de tratamento e do grupo de controle antes e depois do programa. Suponha que, antes do programa, os tratados recebiam em média R$ 1.500 e, após o programa, passaram a receber R$ 2.000 (diferença de R$ 500), enquanto quem não participou do programa recebia em média R$ 2.200 e depois passou a receber R$ 2.400 (diferença de R$ 200). Nesse caso podemos afirmar que o impacto do programa foi de R$ 300 (R$ 500 menos R$ 200, ou seja, **a diferença da diferença**).



Note que nesse caso o salário médio dos dois grupos aumentou. Pode ser que outro fator tenha contribuído para isso, como a instalação de uma grande indústria no município ou a melhoria da conjuntura econômica nacional e local. Mas o fato de ter participado do programa proporcionou maior aumento nos salários dos jovens do grupo de tratamento.

Exemplo 8: Aplicação do método de diferenças em diferenças

No artigo "*Avoidable environmental disasters and infant health: Evidence from a mining dam collapse in Brazil*", Carillo et al. (2020) utilizam o método de diferença em diferenças para avaliar o impacto do colapso de uma barragem de rejeitos da mineração em Mariana, Minas Gerais, sobre a saúde das mulheres grávidas.

Intervenção avaliada

Em 15 de novembro de 2015, ocorreu o rompimento da barragem de Fundão, no município de Mariana, Minas Gerais. O desastre ambiental foi de larga escala, com quase 52 milhões de metros cúbicos de lama despejados pelo caminho. Além das 19 vidas perdidas, centenas de casas foram destruídas e o sistema de água e esgoto interrompidos para várias cidades. A infraestrutura local foi severamente atingida, comprometendo o comércio regional, o turismo e outras atividades, causando uma perda massiva de empregos na região.

As comunidades foram afetadas de diferentes formas, com mais de um milhão de pessoas sendo expostas à lama ao longo dos 600 quilômetros.

Metodologia de avaliação de impacto

Os autores adotaram o método de diferença em diferenças. O grupo de tratamento foi constituído por mulheres residentes nos 36 municípios afetados diretamente pelo desastre grávidas no momento do colapso. Por sua vez, o grupo de comparação foi construído a partir da população de grávidas residentes em outros municípios do estado que teriam sido atingidos pelo rompimento hipotético de outras barragens na região (caso isso tivesse ocorrido). A estratégia de estimar, então, leva em consideração a comparação entre esses dois grupos populacionais, antes e depois do colapso da barragem, configurando uma diferença dupla. Note que o acidente foi aleatório (não previsto). Ao mesmo tempo, ambos os grupos (tratamento e controle) moravam em área passível de ocorrência destes choques (ou seja, não havia auto seleção da amostra dos dois grupos). Assim, garantiu-se que não existia nem viés de seleção nem endogeneidade nos testes feitos.

As variáveis de interesse do estudo são relacionadas à saúde gestacional, incluindo duração da gestação, peso ao nascer e mortalidade infantil.

Resultados

Os principais resultados da avaliação foram:

- Impacto negativo sobre o peso ao nascer, equivalente a uma redução média de 24 gramas;
- Houve um aumento de 89% na mortalidade infantil;
- Não foi encontrado efeito sobre a duração gestacional.

Quadro 7: Hipóteses para o método de diferenças em diferenças

As hipóteses necessárias para que o método de diferença em diferenças seja válido são as seguintes:

1. **Tendências paralelas**: na ausência da política, a tendência observada do indicador de resultado para o grupo de comparação seria igual à observada para o grupo de tratamento.
2. **Não há fatores simultâneos à intervenção** que afetem os grupos de tratamento e comparação de forma diferencial.

A hipótese 1 significa que podemos utilizar a evolução do indicador de resultado observada para o grupo de comparação como um *contrafactual* da evolução do grupo de tratamento. A hipótese 2 garante que qualquer diferença observada ao se comparar as evoluções do indicador de resultado entre esses grupos pode de fato ser atribuída à política, será igual ao *impacto* que desejamos estimar. Dito de outra maneira: não existiriam diferenças não-observáveis entre os dois grupos que impactariam o resultado final.

Vale notar que as hipóteses necessárias para realizar uma avaliação de impacto utilizando o DID são mais fortes do que as hipóteses dos métodos apresentados anteriormente. Isso significa que os resultados melhoram caso seja possível refinar a qualidade dos grupos tratados e de controle pela utilização de métodos como o da regressão descontínua ou quando o tratamento não foi distribuído de maneira aleatória.

Além disso, essas hipóteses não são testáveis. Em alguns casos, todavia, é possível analisar a plausibilidade das tendências paralelas ao comparar a evolução ao longo do tempo entre os grupos de tratados e não tratados em períodos anteriores à política. Para isso, é necessário ter acesso a dados de mais do que apenas um período pré-intervenção. Também é possível realizar testes "placebo" ao utilizar um grupo de tratamento falso (não afetado pela política) ou um indicador de resultado falso (também não afetado). Essas análises adicionais são discutidas em Gertler et al. (2018). Caso exista a disponibilidade de dados de mais de um período, esse "placebo" também pode se relacionar à data de intervenção (ou seja, faz-se o mesmo exercício com uma data aleatória qualquer onde não ocorreu o tratamento. Caso se encontre diferenças estatisticamente significativas, isso indica a possibilidade de o resultado obtido não ser confiável).

Quadro 8: Formalização matemática da estimação do impacto no método de diferenças em diferenças

O impacto estimado pelo método de Diferença em Diferença será dado ao comparar as situações dos grupos de tratamento (GT) e de comparação (GC), antes e depois da intervenção. Essa comparação em duas diferenças dá nome ao método e pode ser representada da seguinte forma:

$$DID = (GT_1 - GT_0) - (GC_1 - GC_0)$$

onde e se referem ao indicador de resultado do grupo de tratamento depois e antes da intervenção respectivamente; e GC_1 e GC_0 são os valores do indicador de resultado para o grupo de controle depois e antes da intervenção respectivamente.

Intuitivamente, ao comparar os momentos *antes* e *depois*, estamos controlando pelos fatores fixos no tempo daquele próprio grupo, já que o grupo é comparado com ele mesmo. No segundo momento, comparamos essas evoluções *depois-antes* entre os grupos de tratamento e comparação, controlando por quaisquer fatores que variem no tempo de forma similar para os dois grupos.

O impacto recuperado pelo DID corresponde ao efeito médio do tratamento sobre os tratados (ATT) (Angrist e Pischke, 2009). Esse parâmetro pode ser estimado a partir de uma regressão linear por mínimos quadrados ordinários da seguinte equação:

$$Y_{it} = \alpha + \gamma T_i + \delta D_t + \tau(T_i * D_t) + e_{it}$$

Sugestão (texto original continua abaixo):

Onde:

Y_{it} : indicador de resultado para o indivíduo i no período t (antes ou depois da política),

T_i: indica se o indivíduo i faz parte do grupo de tratamento,

D_t : variável binária com valor igual a 1 no período pós-intervenção e 0 no período pré-intervenção,

e_{it} : termo de erro.

onde T_i indica se o indivíduo i faz parte do grupo de tratamento, D_t é uma variável binária com valor igual a 1 no período pós-intervenção e 0 no período pré-intervenção, e Y_{it} é o indicador de resultado para o indivíduo i no período t (antes ou depois da política). Ao multiplicar as variáveis binárias de tratamento e de tempo ($T_i * D_t$), conseguimos separar o grupo tratado exatamente no período após a política (identificando a mudança ocorrida após a intervenção). O termo de erro é dado por e_{it}. O coeficiente de interesse na regressão é dado por T, que corresponde ao impacto (ATT). Ele e os demais coeficientes a serem estimados (α, γ, δ) possuem relação direta com as diferenças entre os grupos e entre os períodos, conforme pode ser visto na tabela 1.

Tabela 1. Coeficientes estimados pelo método de Diferença em Diferenças

	Depois ($D_t = 1$)	Antes ($D_t = 0$)	Diferença Depois - Antes
Grupo de tratamento ($T_i = 1$)	$\alpha + \gamma + \delta + \tau$	$\alpha + \gamma$	$\delta + \tau$
Grupo de controle ($T_i = 0$)	$\alpha + \delta$	α	δ
Diferença Tratamento - Controle	$\gamma + \tau$	γ	τ

Fonte: Reproduzido de FGV EESP Clear (2018).

Pareamento

Quando uma avaliação experimental não foi implementada e as regras de seleção do programa que se deseja estudar não permitem a utilização de uma regressão descontínua, ainda é possível realizar uma avaliação de impacto. Além do método de diferença em diferenças, abordado anteriormente, a aplicação de um procedimento de pareamento da amostra também possibilita construir um grupo de comparação nessas situações em que métodos tidos como mais robustos não são possíveis.

Os métodos de pareamento podem ser aplicados em diversas formas de seleção de beneficiários do programa. Para isso, é necessário que exista um grupo de comparação. O método tem como estratégia o uso de técnicas estatísticas capazes de criar um grupo de comparação artificial (Gertler et al, 2018). A ideia é encontrar o "melhor par" entre os elementos do grupo de controle para cada unidade do grupo de tratamento. Dessa forma, procuram-se indivíduos com características parecidas em tudo (em todas as variáveis explicativas), exceto o fato de ter participado do programa.

Exemplo 9: Aplicação do método de pareamento (Gertler et al, 2018)

A figura X ilustra uma situação de pareamento exato com base em quatro características: idade, sexo, meses de desemprego e se possui ou não diploma de ensino médio. Entre as unidades tratadas e não tratadas é possível notar dois indivíduos com as mesmas características: 19 anos, sexo masculino, três meses de desemprego e ausência de diploma de ensino médio. Note que a única diferença entre eles é o fato de o primeiro indivíduo ter sido beneficiário de determinado programa enquanto o segundo indivíduo não foi.

Figura X: Ilustração de três situações de pareamento completo

Unidades tratadas			
Idade	Sexo	Meses de desemprego	Diploma do ensino médio
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Unidades não tratadas			
Idade	Sexo	Meses de desemprego	Diploma do ensino médio
24	1	8	1
38	0	1	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

Portanto, o método do pareamento utiliza características observáveis dos indivíduos da amostra, como idade, sexo e outras características socioeconômicas consideradas relevantes, para atribuir a cada indivíduo beneficiado o seu "grupo de comparação individual" a partir de um grande grupo de indivíduos não tratados. Esse *pareamento* da amostra dá nome ao método e possibilita

a construção do grupo de comparação ao incorporar na amostra final de avaliação apenas os indivíduos pareados entre si. Dessa forma, sob algumas hipóteses, garante-se que os grupos de tratamento e de controle sejam tão parecidos quanto possível, a não ser pela existência de uma intervenção ao primeiro.

Um desafio característico do processo de pareamento consiste no denominado problema da multimensionalidade. À medida que o número de características (variáveis) aumenta fica cada vez mais difícil encontrar pares exatos. Para lidar com esse problema foi criado o método de pareamento por escore de propensão. Nele se estima a probabilidade de se pertencer ao grupo de tratados ou de controle com base nas variáveis disponíveis[12]. Ou seja, busca-se construir um grupo de controle cuja probabilidade de participação no programa seja igual àquela observada para o grupo de tratamento. Ao fazer isso, é possível que o viés de seleção dos dois grupos seja eliminado.

As características a serem consideradas para realizar o pareamento da amostra devem ser pré-definidas, não podem ter sido afetadas pelo próprio tratamento, já que isso incorporaria viés na avaliação. Além disso, segundo Gertler et al. (2018), devem ser escolhidas variáveis que estejam relacionadas à probabilidade de os indivíduos serem tratados ou não, assim como variáveis que afetem as expectativas de resultados futuros. O exemplo 10 apresenta uma aplicação do método de pareamento. O quadro 9 mostra a formalização matemática desta técnica.

Exemplo 10:

No artigo "*Efeitos do Programa Territórios da Cidadania sobre indicadores econômicos e sociais nos municípios de Minas Gerais* ", Neder e Lopes (2015) utilizam o método de pareamento para avaliar o impacto do programa

Intervenção avaliada

O programa teve por objetivo promover o desenvolvimento de áreas do país com baixos índices de desenvolvimento e dinâmica econômica. Ele foi criado em 2003 pelo governo federal através dos Consórcios Intermunicipais de Segurança Alimentar e Desenvolvimento Local (Consads) e dos Territórios Rurais. Os municípios beneficiários da política contavam com características rurais e baixos índices de desenvolvimento econômico e social.

As ações do programa incluem a integração de políticas públicas regionais como financiamentos do Programa Nacional de Fortalecimento da Agricultura Familiar (Pronaf), Programa Luz para Todos, Bolsa Família, Farmácia Popular, Brasil Sorridente, Cisternas etc. Diferentes ações de ministérios e governos estaduais e municipais são combinadas para fortalecer a produção agropecuária local.

Metodologia de avaliação de impacto

Os autores adotaram o método de pareamento por escore de propensão. Foram utilizadas diversas variáveis econômicas e sociais dos municípios tratados (que receberam o programa) e não tratados. Algumas das variáveis utilizadas foram produto interno bruto per capita, índice de desenvolvimento humano, população rural, valor adicionado do setor

[12] O Quadro 9 apresenta maiores detalhes da técnica de pareamento por escore de propensão.

agropecuário, proporção de domicílios abaixo da linha de pobreza, proporção de domicílios com água canalizada, proporção de domicílios com esgoto, entre outras.

Resultados

Os principais resultados da avaliação foram:

- Aumento do valor adicionado do setor agropecuário entre os municípios que participaram do programa.
- Melhoria das condições de saúde da população dos municípios beneficiários.

Quadro 9: Formalização matemática do pareamento por escore de propensão

Apesar de existirem diversas maneiras de realizar o pareamento da amostra, uma das mais recorrentes, em grande parte por conta de sua facilidade de implementação que não gera déficits à qualidade do pareamento, é o pareamento por escore de propensão (ou *propensity score*). Conforme definido anteriormente, esse escore nada mais é do que a probabilidade de participação na política (a probabilidade de que o indivíduo seja tratado).

Podemos parear os indivíduos do grupo de tratamento que tenham aos indivíduos não tratados que apresentem o escore de propensão mais próximo possível. Esse procedimento é de implementação relativamente simples, já que pareia a amostra a partir de uma única variável (), mas ainda permite a construção de um grupo de comparação adequado, desde que respeitadas as hipóteses do método.

Fica claro também que a probabilidade de se participar do tratamento deve ser função de variáveis observáveis para ambos os grupos. Caso isto não ocorra, o método simplesmente não tem sentido.

Ainda assim, há alguns critérios que devem ser observados ao realizar o pareamento da amostra com base no escore de propensão. Conforme discutido por Caliendo e Kopeinig (2008), é preciso decidir, por exemplo, quantos indivíduos não tratados serão pareados a cada quantos tratados, qual será o nível de tolerância para diferenças do escore de propensão etc. Essas propriedades podem ser resumidas desta forma:

- **Vizinho mais próximo**: no pareamento por vizinho mais próximo, define-se um número de indivíduos não tratados (um ou mais) que serão pareados a cada indivíduo do grupo de tratamento. Então, são escolhidos os indivíduos (tantos quantos definido de antemão) seguindo a regra de proximidade do escore de propensão.
- ***Caliper*:** no pareamento por *caliper*, que pode ser aliado a outros critérios, define-se o nível de tolerância máximo permitido, em termos da diferença entre os escores de propensão de um indivíduo tratado ao seu par (ou pares) não tratado. A ideia do *caliper* é evitar pares distantes, que possivelmente não configurem uma boa comparação.
- **Raio**: define-se um nível máximo de distância entre os pares a serem formados. Difere do *caliper* por considerar pareados todos os indivíduos não tratados que tenham o escore de propensão dentro dessa margem.

- ***Kernel***: no pareamento por *kernel*, atribui-se um peso a cada indivíduo da amostra, calculado com base no escore de propensão. Intuitivamente, indivíduos não tratados com escores de propensão mais próximos do grupo de tratamento receberão peso maior.

Alguns desses critérios podem ser combinados, como o caso do vizinho mais próximo e do *caliper*, por exemplo. Frequentemente, pode ser útil comparar os resultados da estimação utilizando critérios distintos como uma espécie de teste de robustez. Obter resultados similares mesmo com variações na implementação do método pode ser um indicativo de confiabilidade.

Estimação do Impacto

Em toda aplicação do método de pareamento, a primeira etapa da estimação é exatamente calcular o score (a probabilidade de participar da política no caso do pareamento por escore de propensão) de cada elemento da amostra. O escore de propensão deve ser estimado, por exemplo, com o uso de um modelo logit ou probit (esses modelos podem ser consultados no capítulo 17 de Wooldridge, 2011) em que a variável dependente é o status de tratamento de cada indivíduo, em função das características observáveis relevantes. É importante notar que se deve trabalhar somente com variáveis relativas ao período antes do tratamento.

Na sequência, uma vez formado o grupo de comparação com base nos critérios de pareamento definidos, estimar o impacto pode ser realizado, por exemplo, a partir de uma regressão linear (em que podem ser considerados ainda pesos individuais baseados no escore de propensão, de acordo com o critério de pareamento escolhido) do resultado de interesse Y_i em função do indicador de tratamento T_i e das características observáveis X_i, como na equação abaixo:

$$\gamma Y_i = \alpha + \beta T_i + \lambda X_i + e_i + \tau$$

onde o parâmetro de interesse, que revelará o impacto causal, é dado por β.

Por fim, no quadro 10 explicitamos o que é um teste de hipóteses na estatística.

Quadro 10: Testes de hipóteses

Em todos os métodos de avaliação de impacto discutidos nos capítulos anteriores, o resultado obtido é um coeficiente que corresponde ao impacto causal mensurado por aquele método, podendo ser o efeito médio do tratamento (ATE), o efeito médio do tratamento entre os tratados (ATT), o efeito médio local do tratamento (Late) ou o Efeito da Intenção de Tratar (ITT). Entretanto, como estamos tratando de um processo de *estimação*, há sempre um componente de incerteza sobre qualquer desses parâmetros estimados.

O chamado **teste de hipóteses** é uma maneira estatística de se aferir a significância de um determinado parâmetro. Exemplos clássicos de aplicação de testes de hipóteses envolvem testar se a média amostral de uma variável (calculada utilizando apenas uma amostra da população) é equivalente à média observada para a população.

No caso das avaliações de impacto, estamos interessados em testar se o coeficiente estimado do impacto causal de uma determinada política é diferente de zero. Caso a estimativa de impacto seja corroborada por esse procedimento, dizemos que o impacto estimado é **estatisticamente significante**, ou **estatisticamente diferente de zero**. Como são milhares de fatores aleatórios a impactar as medidas estatísticas utilizadas (que podem mudar tanto o valor real do que se quer medir como gerar algum viés no próprio mecanismo de medição), pode-se testar se o valor encontrado é diferente ou não de zero com determinado nível de certeza. É isso que é feito quando se faz tal teste.

Conforme discutido por Bussab e Morettin (2017) e, no contexto de avaliação de impacto, por Duflo et al. (2007), podemos formular o teste de hipóteses a partir da chamada *hipótese nula* () e de uma *hipótese alternativa* () correspondente. Em geral, a hipótese nula é aquela que gostaríamos de rejeitar. Ela diz que o impacto é igual a zero. Nesse caso, uma possível hipótese alternativa é de que o impacto seja diferente de zero. Considerando que o parâmetro do impacto seja dado por β, podemos formalizar essas hipóteses assim:

$$\gamma Y_i = \alpha + H_0\colon \beta = 0 \text{ (hipótese nula)}$$
$$H_1\colon \beta \neq 0 \text{ (hipótese alternativa)}$$

Como veremos na sequência, o *teste de hipóteses* será capaz de fornecer evidências para que possamos rejeitar a hipótese nula (nos levando a aceitar a hipótese alternativa). Nesse contexto, é possível pensar em duas situações em que chegamos à conclusão errada:

- **Erro do tipo I:** rejeitar H_0 quando H_0 é verdadeira (falso positivo)
- **Erro do tipo II:** não rejeitar H_0 quando H_0 é falsa (falso negativo)

Em uma avaliação de impacto, cometer um erro do tipo I corresponde a dizer que a política avaliada possui um impacto relevante, quando na verdade o impacto é igual a zero (por essa razão, conhecido como *falso positivo*). O caso contrário, o erro do tipo II, corresponde a dizer que a política não possui impacto, quando a realidade é que o impacto existe (*falso negativo*). Podemos formalizar a probabilidade de cometer os erros do tipo I e II como a seguir:

$$P(Erro\ I) = P(Rejeitar\ H_0 \mid H_0\ verdadeira) = \alpha$$
$$P(Erro\ II) = P(Não\ rejeitar\ H_0 \mid H_0\ falsa) = 1 - \kappa$$

A probabilidade de cometer um erro do tipo I, que chamamos de α, é também conhecida como P-Valor. Definimos um nível máximo aceitável desse valor para que aceitemos que Ho deva ser rejeitada. Esse nível máximo é o chamado **nível de significância** do teste de hipóteses. Esse valor é escolhido pelo avaliador e usualmente é fixado em α = 5. Quando rejeitamos a hipótese nula, dizemos que o impacto estimado é considerado estatisticamente significante ao nível de significância de 5% (ou, simplesmente, significante a 5%). Já é usual (uma vez que os programas tendem a calculá-lo) reportar diretamente o P-Valor dos testes feitos para o leitor.

Podemos também definir o poder do teste realizado (κ). Ele variaria entre zero e um. Quando é zero, sempre rejeitamos a hipótese alternativa, mesmo que ela seja verdadeira (no nosso caso, concluímos que não houve impacto quando efetivamente houve). Quando é um,

só rejeitamos a hipótese alternativa quando ela é efetivamente falsa (ou seja, só concluímos que uma política não apresenta impacto quando ela efetivamente não o faz). Esse poder (κ) variaria inversamente com a probabilidade de se fazer o erro tipo II. Esse conceito, assim como o nível de significância, é importante para o cálculo do efeito mínimo detectável (MDE), apresentado no capítulo anterior deste guia. Naquele contexto, um valor usual para o poder é κ + 80%.

Estatística t e Valor-p

Para realizar um teste de hipóteses, adotamos uma regra de decisão com base no *nível de significância* escolhido. Dizemos que a hipótese nula é rejeitada quando a seguinte condição é satisfeita:

$$|t| = |\frac{\widehat{\beta}}{\sigma}| > c$$

onde t é a estatística do teste conhecida como **estatística t**, que é igual à razão entre o valor absoluto do impacto estimado ($\widehat{\beta}$) e o desvio-padrão do estimador (σ). A regra de decisão nos leva a rejeitar a hipótese nula quando o valor absoluto da *estatística t* é maior do que um valor crítico (c), determinado a partir do nível de significância do teste pela equivalência. Caso o parâmetro estimado esteja, em unidades de desvio-padrão, mais longe de zero que determinado valor crítico, pode-se inferir que a probabilidade de esse mesmo parâmetro ser efetivamente zero é extremamente baixa se não nula[13] – quanto maior a distância, menor a probabilidade de ele ser nulo.

Para uma amostra de avaliação grande o suficiente, o valor crítico correspondente ao nível de significância de 5% será tal que c = 1,96 . Será possível rejeitar a hipótese nula de um coeficiente igual a zero – e, portanto, dizer que o impacto estimado é estatisticamente significante – quando a estatística t calculada for maior que 1,96 (em valores absolutos).

Um conceito relacionado ao da estatística t é o do **p-valor, já descrito**. Esse valor corresponde à probabilidade de significância do teste, indicando a probabilidade de que ocorram valores da estatística t mais extremos do que o observado sob a hipótese de impacto igual zero (Bussab e Morettin, 2017).

Intuitivamente, o valor-p é o menor nível de significância ao qual rejeitaríamos a hipótese nula. Na prática, isso significa que, ao obter um valor-p menor que 5%, por exemplo, é possível dizer que o impacto estimado é estatisticamente significativo a 5%. Em geral, é possível obter os valores da estatística t e do valor-p automaticamente ao realizar uma regressão em um software estatístico como o Stata ou o R.

[13] Este teste pode ser feito em relação a qualquer valor que se queira. Genericamente, pode-se testar se o valor estimado é estatisticamente igual a determinado valor X. Neste caso, trabalha-se com a distância, em desvios-padrão, entre o valor estimado e X, ou seja, |-X|.

Considerações sobre o guia

Este guia buscou apresentar os principais conceitos e técnicas utilizadas nas avaliações de impacto, com o objetivo de fortalecer a cultura avaliativa no setor público. A FJP e o FGV Clear esperam contribuir para a disseminação dessas práticas, oferecendo insumos teóricos e práticos que auxiliem gestores e gestoras públicas a compreender os diferentes tipos de avaliação, a estruturar boas demandas e a interpretar seus resultados de forma qualificada. Não é necessário que os responsáveis pelas políticas públicas sejam especialistas em avaliação, mas é fundamental que saibam como utilizá-la para melhorar suas decisões e aprimorar programas e políticas.

Foram discutidas, de forma objetiva, as principais técnicas utilizadas na avaliação de impacto de programas públicos. As avaliações experimentais, muitas vezes consideradas ideais, permitem isolar os efeitos de uma política pública de outros fatores externos. No entanto, sua aplicação costuma ser inviável na maioria dos contextos governamentais. Felizmente, há alternativas metodológicas robustas, como os modelos quase-experimentais, entre eles diferenças em diferenças, variáveis instrumentais, regressão descontínua e pareamento, que possibilitam análises rigorosas mesmo em cenários mais desafiadores.

Espera-se que este guia tenha esclarecido conceitos e metodologias tanto para profissionais com formação mais voltada à pesquisa quantitativa quanto para gestores e gestoras interessados em qualificar suas demandas por avaliações de impacto. A adoção de práticas avaliativas bem fundamentadas pode contribuir muito para o aprimoramento das políticas públicas e para a geração de valor público com base em evidências.

Referências bibliográficas

ANGRIST, J.; PISCHKE, J. Mostly harmless econometrics: an empiricists guide. 2009.

ARVATE, P.; FALSETE, F.O.; RIBEIRO, F.G.; SOUZA, A.P. Lighting and homicides: Evaluating the effect of an electrification policy in rural Brazil on violent crime reduction. Journal of quantitative criminology, v. 34, n. 4, p. 1047-1078, 2018.

BARROS, R.P.; BIDERMAN, C.; LIMA, L.; SOUZA, A. P. Rescuing At-Risk Youth: Experimental Evidence from a Human Capital Investments Program in Brazil. 2019.

BRUHN, M.; LEÃO, L.D.S.; LEGOVINI, A.; MARCHETTI, R.; ZIA, B. The impact of high school financial education: Evidence from a large-scale evaluation in Brazil. American Economic Journal: Applied Economics, v. 8, n. 4, p. 256-95, 2016.

BUSSAB, W. O.; MORETTIN, P. A. Estatística básica. Editora Saraiva, 2017.

CALIENDO, M.; KOPEINIG, S. Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22(1), pp.31-72, 2008.

CARRILLO, B.; DA MATA, D.; EMANUEL, L.; LOPES, D.; SAMPAIO, B. Avoidable environmental disasters and infant health: Evidence from a mining dam collapse in Brazil. Health economics, v. 29, n. 12, p. 1786-1794, 2020.

CHEIN, FLAVIA. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas / Flávia Chein. -- Brasília: Enap, 2019.

DJIMEU, E.W.; HOUNDOLO, D.G. Power calculation for causal inference in social science: sample size and minimum detectable effect determination. Journal of Development Effectiveness, 8(4), pp.508-527, 2016.

DUFLO, E.; GLENNERSTER, R.; KREMER, M. Using randomization in development economics research: A toolkit. Handbook of development economics, v. 4, p. 3895-3962, 2007.

FAHEL, M.C.X.; FRANÇA, B.C.; MORAES, T. O efeito da condicionalidade educação do Bolsa Família em Minas Gerais: uma avaliação por meio da PAD/MG. Revista Brasileira de Monitoramento e Avaliação, vol.2, p.4-25, 2011.

FGV EESP Clear. Centro de Aprendizagem em Avaliação e Resultados. Curso ForMA de Avaliação de Impacto Avançada. Materiais de curso. São Paulo, 2018.

FUJIWARA, T. Voting technology, political responsiveness, and infant health: Evidence from Brazil. Econometrica, v. 83, n. 2, p. 423-464, 2015.

GERTLER, P. J. et al. Avaliação de Impacto na Prática, Segunda edição. World Bank Publications, 2018. Disponível em: <https://openknowledge.worldbank.org/bitstream/handle/10986/25030/9781464808890.pdf>

IMBENS, G.; LEMIEUX, T. Regression discontinuity designs: A guide to practice. Journal of econometrics, v. 142, n. 2, p. 615-635, 2008.

IMBENS, G.; WOOLDRIDGE, J. "What's New in Econometrics" Lecture Notes. National Bureau of Economic Research (NBER), 2007. Disponível em: <http://www.nber.org/WNE/WNEnotes.pdf>

LEE, D. S.; LEMIEUX, T. Regression discontinuity designs in economics. Journal of economic literature, v. 48, n. 2, p. 281-355, 2010.

MEL, S.; MCKENZIE, D.; WOODRUFF, C. Returns to capital in microenterprises: evidence from a field experiment. The Quarterly journal of Economics, v. 123, n. 4, p. 1329-1372, 2008.

MENEZES FILHO, N.; PINTO, C.C.X. (organizadores). Avaliação Econômica de Projetos Sociais, 3ª Edição. Fundação Itaú Social, 2017.

NEDER, Henrique Dantas, LOPES, Tiago Camarinha. EFEITOS DO PROGRAMA TERRITÓRIOS DA CIDADANIA SOBRE INDICADORES ECONÔMICOS E SOCIAIS DOS MUNICÍPIOS DE MG: uma abordagem de mensuração com métodos de pareamento. Revista de Políticas Públicas, v. 19, n. 2, 2015. ISSN 0104-8740.

PRiME. Program in Rural Monitoring and Evaluation. Impact Evaluation, 2nd Edition. Presentation. Dacar, 2019. Disponível em: <https://www.primetraining.global/materials/fundamentals>

WOOLDRIDGE, J. Introdução à econometria: uma abordagem moderna. São Paulo: Cengage Learning, 2011.