SÉRIE
Avaliação na Prática

AVALIAÇÃO DE IMPACTO

clear avaliação África lusófona e brasil

FGV EESP
ESCOLA DE ECONOMIA DE SÃO PAULO

global evaluation initiative

Avaliação de Impacto

Organizadores

André Portela Souza
Lycia Lima

Equipe técnica

Alei Fernandes Santos
Amanda Cappellazzo Arabage
Ana Carolina Marinato de Resende
Bernardo Ostrovski
Caio de Souza Castro
Fernando Gonçalves Marques
Michel Szklo
Patrícia Franco Ravaioli
Otávio Luiz Tecchio
Pedro Molina Ogeda
Victor Simões Dornelas

Citação sugerida:
LIMA, L.; SOUZA, A. P. (Orgs.). Avaliação de impacto. São Paulo: Fundação Getulio Vargas, 2025.

APRESENTAÇÃO

Este guia integra a série de publicações *Avaliação na Prática*, desenvolvida pelo FGV CLEAR com o objetivo de ampliar o acesso a conhecimentos sobre monitoramento e avaliação com foco em políticas públicas.

Fundado em 2015, o FGV CLEAR tem se dedicado ao fortalecimento da cultura de gestão orientada por evidências no Brasil e em países lusófonos. Com sede na Escola de Economia de São Paulo da Fundação Getulio Vargas (FGV EESP), o FGV CLEAR atua como centro regional da Iniciativa CLEAR (*Centers for Learning on Evaluation and Results*).

A Iniciativa CLEAR, criada em 2010, é um programa de desenvolvimento de capacidades em monitoramento e avaliação que congrega instituições acadêmicas e parceiros doadores de modo a contribuir para o uso de evidências na tomada de decisões em países em desenvolvimento. Ao todo, são seis centros regionais CLEAR, coordenados pela *Global Evaluation Initiative (GEI)*, um programa liderado pelo Banco Mundial e pelo Programa das Nações Unidas para o Desenvolvimento (PNUD).

O FGV CLEAR atua com governos, organizações multilaterais, sociedade civil e academia, oferecendo capacitação, assistência técnica, geração e disseminação de conhecimento científico na área.

Saiba mais sobre o FGV CLEAR e acesse outras publicações em: **www.fgvclear.org**

Sumário

CAPÍTULO 1.

Introdução

O presente guia tem como objetivo apresentar os principais conceitos e metodologias de avaliação de impacto no contexto das políticas públicas. Busca-se, assim, oferecer a gestores, técnicos e pesquisadores um arcabouço teórico introdutório que os capacite a compreender as diferentes metodologias de avaliação de impacto, bem como contratar avaliações robustas e acompanhar suas execuções, com vistas à mensuração precisa dos efeitos gerados por uma intervenção.

A avaliação de impacto corresponde a uma abordagem específica dentro do campo da avaliação de políticas públicas. Trata-se de um processo baseado em métodos rigorosos, com ênfase predominantemente quantitativa, cujo propósito central é isolar e medir os efeitos atribuíveis exclusivamente à política analisada. Em outras palavras, pretende-se entender em que medida a intervenção foi eficaz em enfrentar o problema social que motivou sua formulação.

Dada a complexidade dos problemas sociais — geralmente influenciados por múltiplos fatores —, é fundamental estimar com precisão qual parcela das mudanças observadas pode, de fato, ser atribuída à política em questão. Para isso, é necessário estabelecer uma relação causal entre a intervenção e os resultados alcançados. A pergunta que orienta essa análise é simples, mas poderosa: qual é o impacto da política sobre os resultados esperados?

Nesse contexto, "impacto" é definido como a diferença nos resultados entre dois cenários hipotéticos: um no qual a política foi implementada, e outro em que não foi. Quando a única diferença entre esses dois cenários é a presença ou não da política, pressupõe-se que, após a implementação, quaisquer diferenças observadas nos indicadores de interesse podem ser atribuídas à intervenção. Para realizar essa comparação, é preciso construir um cenário contrafactual — isto é, uma estimativa do que teria acontecido na ausência da política. Esse exercício exige a seleção criteriosa de um grupo de controle, que permita uma comparação válida e confiável.

Ao longo do guia, serão aprofundados os fundamentos da avaliação de impacto, além de discutidos os conceitos de validade interna e externa, essenciais para avaliar a credibilidade dos resultados e seu potencial de generalização. O guia também apresentará o conceito de viés de seleção - um dos obstáculos mais comuns à validade das conclusões -, além de oferecer

uma visão geral das principais metodologias de avaliação de impacto, com explicações acessíveis sobre as diferentes estratégias utilizadas para estimar efeitos causais a partir de comparações entre grupos. Por fim, discute-se como os impactos observados em uma política específica se comparam aos resultados de outras iniciativas semelhantes, oferecendo uma perspectiva mais ampla sobre sua efetividade relativa.

Por tratar-se de um guia introdutório, muitos dos conceitos e abordagens aqui apresentados são explorados de forma sintética. Para os interessados em se aprofundar, o FGV CLEAR disponibiliza a série "Avaliação na Prática", reunindo uma série de publicações que desenvolvem esses temas com maior detalhamento teórico e técnico[1]

CAPÍTULO 2.

Avaliação de Impacto e Relação Causal

Na fase de formulação da política, são definidos quais objetivos pretende-se alcançar por meio dela. Assim, uma vez implementada, é tentador concluir que a política terá sido bem-sucedida (isto é, demonstrará ter impacto) se as mudanças pretendidas por meio dela passarem a ser observadas.

Considere, por exemplo, uma política de vacinação cujo objetivo seja reduzir a disseminação de uma determinada doença. À primeira vista, parece razoável afirmar que essa política não produziu impacto caso o número de casos registrados da doença não tenha diminuído em relação ao período anterior à sua implementação. Por mais plausível que possa parecer, essa conclusão pode estar equivocada. Suponha que a política de vacinação do país tenha sido implementada às vésperas de uma epidemia da doença originada em países vizinhos. Nesse contexto, o número de casos da doença provavelmente seria muito maior caso o programa jamais tivesse sido executado. Logo, o fato de a incidência da doença ter permanecido no mesmo nível de antes da política pode ser, na verdade, um atestado de seu sucesso, sendo as vacinas responsáveis por evitar que muito mais pessoas no país fossem acometidas pela doença. Em outras palavras, a vacinação seria a causa de não se observar um aumento no número de doentes.

Esse exemplo ilustra que, frequentemente, estabelecer uma **relação de causa e efeito** entre uma política e seu objetivo não é tão simples como verificar se tal objetivo foi ou não observado. Dado o contexto em que está inserido, o público-alvo da política estará sempre

[1] Confira todas as publicações em www.fgvclear.org/biblioteca/

sujeito a influências não relacionadas a ela (como o aparecimento de uma variante da doença que não seja coberta pela vacina, ainda no exemplo da vacinação). Portanto, atestar se uma política é ou não bem-sucedida exige uma maneira convincente de, a partir dos dados e informações disponíveis, evidenciar seu efeito em meio aos vários outros efeitos oriundos de fatores concomitantes a ela. A forma ideal de se entender os impactos de políticas é a partir da utilização de um grupo comparável ao grupo afetado pelas intervenções, permitindo isolar os efeitos atribuíveis à política.

2.1 Contrafactual e impacto

A avaliação de impacto tem, em seu cerne, a comparação de unidades que são contempladas pela política com outras que não o são. Esse procedimento baseia-se em uma construção abstrata, um modo de pensar sobre o comportamento das unidades em relação à política. Considera-se que cada unidade possui dois estados:

i. Ela participa ou recebe as intervenções previstas pela política, situação na qual é comum se dizer que a unidade recebe o tratamento da política, ou ainda, que é **tratada**.
ii. Ela não participa ou não recebe as intervenções previstas pela política, ou seja, é uma unidade **não tratada** ou de **controle.**

A partir desses dois estados, delineia-se uma estratégia para verificar se a política gera ou não algum impacto: comparar o que se sucede à unidade no estado em que ela recebe o tratamento da política com o que ocorre no estado em que não o recebe. A diferença entre os estados da unidade seria, assim, o efeito da política.

Embora simples em princípio, essa comparação é, como mencionado, um exercício abstrato. Isso porque os estados (i) e (ii) são mutuamente exclusivos – ou seja, se uma unidade é contemplada pela política, jamais se saberá o que ocorreria se ela não tivesse recebido tal tratamento (e vice-versa). Assim, o que impede uma verificação direta e precisa da existência de efeito da política é o fato de que um dos estados sempre será desconhecido. Esse estado desconhecido se denomina **contrafactual.**

Apesar de abstrato, o raciocínio desenvolvido acima explicita o que é fundamental para a construção de medidas factíveis do impacto de uma política: fornecer na realidade uma representação aproximada à situação contrafactual das unidades tratadas. Com vistas a esse fim, é convencional em avaliações de impacto de políticas se definir dois grupos de unidades:

i. **Grupo de tratamento:** composto pelas unidades que recebem o tratamento decorrente da política.

ii. **Grupo de controle (não tratados ou de comparação):** formado por unidades que não recebem o tratamento da política e que se acredita consistir em uma representação aproximada da situação contrafactual do grupo de tratamento.

Importante

A escolha de um bom grupo de controle, que represente o contrafactual do grupo de tratamento da maneira mais fiel possível, viabiliza o cálculo robusto de algumas medidas (ou parâmetros) interessantes para avaliar se a política provoca o impacto esperado. Uma medida frequentemente adotada é a de efeito médio: o valor médio do impacto da política sobre um conjunto de unidades ou, posto de outra forma, o impacto que se espera observar sobre uma unidade qualquer do conjunto.

2.2 Indicadores de impacto, parâmetros de interesse e intervalos de confiança

Como, na prática, o cálculo do impacto causal é realizado a partir da comparação entre grupos — e não entre unidades individuais —, ele se baseia em parâmetros como o do efeito médio. Assim, em uma avaliação quantitativa de uma política, presume-se que ao menos um aspecto relevante do objetivo pretendido possa ser mensurado numericamente. Esses aspectos permitem a construção de **indicadores de impacto**, ou seja, variáveis numéricas que, com base no desenho da política, espera-se que sejam influenciadas por sua implementação.

No caso de uma política de vacinação, por exemplo, o número de casos diagnosticados da doença constitui um indicador de impacto: se a vacinação de fato previne o acometimento pela doença, é esperado que a quantidade de diagnósticos positivos diminua após o início da política. Assim, o impacto da política na redução da incidência da doença pode ser verificado por meio da comparação entre o número de casos registrados após a campanha de vacinação e o número que seria observado na situação contrafactual, da qual o grupo de controle deve ser uma representação aproximada.

Diversas estratégias empíricas podem ser utilizadas para estimar o efeito médio sobre um indicador de impacto. Como será discutido na próxima seção, algumas estratégias são mais adequadas do que outras, a depender das características da implementação da política e dos dados disponíveis. No entanto, essas estratégias diferem não apenas em termos de aplicabilidade, mas também quanto à interpretação do parâmetro estimado. Embora o parâmetro de interesse seja o efeito médio da política, é fundamental atentar-se ao conjunto de unidades ao qual essa média se refere. Para abordar o assunto, a **Figura 6.1.** apresenta 4 tipos possíveis de efeitos:

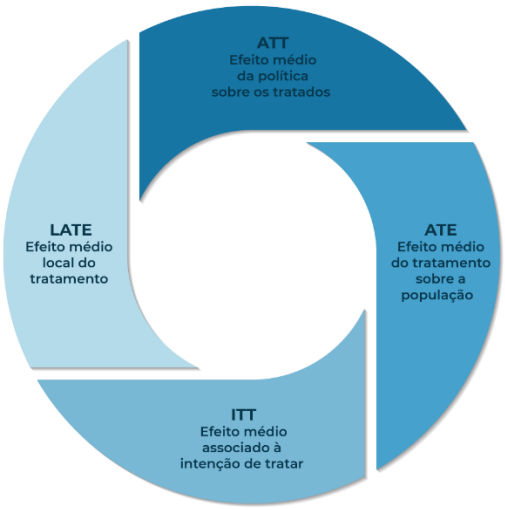


Figura 6.1. Efeito médio da política

Quando os grupos de tratamento e de controle são representativos de toda uma população, é possível estimar o **efeito médio do tratamento sobre a população (ATE)**[2] a partir da diferença de médias do indicador de impacto de interesse entre os dois grupos. No entanto, nem sempre os grupos de tratamento e controle são representativos da população. Nesses casos, é possível estimar outros tipos de efeitos de tratamento. Um deles é o **efeito médio do tratamento sobre os tratados (ATT)**, que pode ser calculado quando não se pode assumir que o grupo de tratamento seja generalizável para toda a população — ou seja, quando não é possível inferir que o comportamento de qualquer unidade tratada seria o mesmo que o da população em geral. Nesses contextos, não se pode afirmar qual seria o impacto da política caso ela fosse estendida ao restante da população.

Quando há dificuldade em assegurar que as unidades para as quais a política foi ofertada efetivamente receberam o tratamento, o parâmetro estimado pode ser interpretado como o **efeito médio associado à intenção de tratar (ITT)**[3]. Ou seja, esse parâmetro estimado desconsidera a possibilidade de que algumas unidades às quais a política foi direcionada não tenham, de fato, sido tratadas (caso de cumprimento parcial), e se baseia estritamente na definição original dos grupos: as unidades designadas ao grupo de tratamento são consideradas como tendo recebido a política, independentemente de isso ter ocorrido na prática — o mesmo valendo, de forma análoga, para o grupo de controle.

[2] Em inglês, frequentemente, recebe a denominação *average treatment effect* (ATE).
[3] Em inglês, frequentemente, recebe a denominação *intention to treat* (ITT).

Ainda nesse contexto de cumprimento parcial por parte das unidades, pode ser razoável estimar um parâmetro mais restrito de efeito médio do tratamento, dependendo das características específicas da avaliação. O **efeito médio local do tratamento (LATE)**[4] refere-se ao impacto calculado considerando apenas as unidades que seguiram a designação original dos grupos — ou seja, aquelas que foram designadas para o grupo de tratamento e efetivamente receberam a política, bem como aquelas designadas para o grupo de controle e que, de fato, não a receberam.

Outro aspecto a ser ressaltado é que os efeitos médios do tratamento calculados em uma avaliação de impacto correspondem a estimativas obtidas a partir de uma amostra da população. Por essa razão, os valores encontrados estão sujeitos a algum grau de incerteza quanto ao seu verdadeiro valor populacional. Essa incerteza é expressa por meio dos erros-padrão e dos intervalos de confiança que acompanham as estimativas. De forma simplificada, o intervalo de confiança representa uma faixa de valores ao redor da estimativa pontual que provavelmente contém o verdadeiro efeito médio que a política teria se fosse aplicada à população como um todo, considerando-se um determinado nível de confiança estatística.[5]

Quanto menor for o erro-padrão estimado, mais estreito será o intervalo de confiança associado, o que confere à avaliação maior precisão e poder informativo. Além disso, quando o valor zero está contido no intervalo de confiança, não é possível afirmar, com o nível de confiança adotado, que o impacto da política seja estatisticamente diferente de zero - isto é, não se pode rejeitar a hipótese de ausência de impacto da intervenção analisada.

2.3 Validade Externa

Ao estimar o impacto de uma política, obtém-se uma evidência sobre o efeito causal da intervenção sobre uma variável de resultado. No entanto, interpretar esse resultado de maneira adequada requer mais do que avaliar sua magnitude ou a direção: é essencial discutir se os resultados observados podem ser generalizados para além do grupo analisado — em outras palavras, considerar sua **validade externa**.

A validade externa diz respeito à possibilidade de generalizar os resultados observados para outros contextos, populações ou períodos. Por exemplo, ao avaliar o impacto de uma política implementada em um conjunto específico de municípios, é natural questionar se esse impacto se manteria caso o programa fosse expandido para todas as cidades do país. Da mesma forma,

[4] Em inglês, frequentemente, recebe a denominação *local average treatment effect* (LATE).

[5] É usual considerar os níveis de confiança de 95% e de 99%. Para maiores detalhes sobre intervalos de confiança e significância estatística, ver os capítulos 14 e 15 em Meyer (1983). Um detalhamento aplicado a estimadores econométricos pode ser visto em Wooldridge (2010b, cap. 4). Keuzenkamp e Magnus (1995) também fazem uma discussão a respeito do tema.

caso a política atinja apenas um grupo específico da população, é importante refletir sobre o quanto os resultados obtidos são aplicáveis a outros grupos. Essa preocupação torna-se especialmente relevante ao comparar os efeitos de políticas semelhantes implementadas em diferentes locais ou momentos.

A possibilidade de generalização dos resultados está diretamente relacionada à seleção das unidades analisadas. Quando o grupo estudado é representativo da população de interesse e o critério de seleção para o tratamento não depende fortemente do esforço individual para aderir à política, há maior confiança na validade externa da estimativa. No entanto, como destacam Athey e Imbens (2017), nenhum desenho de seleção, independentemente da metodologia adotada, garante validade externa por si só. Por exemplo, em políticas voltadas a indivíduos, uma das principais questões é o comparecimento ou adesão efetiva daqueles selecionados para receber a intervenção. Nada garante que os resultados observados se apliquem a toda a população, pois o grupo que efetivamente recebe o tratamento pode divergir do grupo originalmente designado ao tratamento, caso algumas pessoas não atendam à seleção, conforme explicado por Andrews e Oster (2019).

Além da forma como as unidades foram selecionadas, a capacidade de generalizar os resultados de uma avaliação também depende do parâmetro de efeito médio estimado. Cada parâmetro captura o impacto da política sobre um grupo específico, o que impõe diferentes limites à sua validade externa. O ATE representa o efeito médio do tratamento sobre toda a população de interesse. Por fornecer uma medida abrangente, ele é o parâmetro mais útil quando o objetivo é estimar qual seria o impacto da política se fosse implementada em larga escala.

O ATT, por sua vez, estima o efeito apenas sobre aqueles que efetivamente receberam o tratamento, sendo apropriado quando o objetivo é compreender o impacto da política sobre o público efetivamente alcançado. No entanto, ele não permite inferir o que ocorreria com os não tratados, o que limita seu potencial de generalização. Já o LATE corresponde ao efeito médio sobre um subconjunto específico da população — aqueles que efetivamente seguiram as definições dos grupos de controle e tratamento. Por ser restrito a um grupo particular, esse parâmetro tem validade externa ainda mais limitada. Portanto, ao analisar a validade externa dos resultados de uma avaliação, é fundamental compreender a quem o efeito estimado se refere.

2.4 Validade interna, hipóteses e análises de sensibilidade

Por mais que as estratégias de avaliação de impacto procurem produzir, com base em dados, as melhores informações para apoiar o processo de tomada de decisão, elas sempre partem da presunção de validade de um conjunto de hipóteses. Quando há confiança de que a

metodologia foi aplicada corretamente e de que os pressupostos necessários são plausíveis, é possível afirmar que existe **validade interna**. Ou seja, dentro do contexto avaliado, o resultado estimado pode ser interpretado como um **efeito causal** da política sobre o indicador de impacto de interesse.

A validade interna, portanto, depende diretamente da adequação do desenho de avaliação e da **consistência das estimativas econométricas** obtidas. Isso significa que só será razoável confiar em uma estimativa de impacto se ela for compatível com as hipóteses do método utilizado.[6] Quando essas condições são atendidas, podemos considerar que o impacto estimado reflete, de fato, a relação de causa e efeito no grupo analisado.

Importante

A comunicação dos resultados da avaliação de impacto deve explicitar que ela pressupõe um conjunto de hipóteses. Caso isso não ocorra, corre-se o risco de transmitir uma falsa impressão de certeza, o que pode levar a decisões equivocadas por parte dos formuladores de políticas ou de outros públicos interessados.

Tais hipóteses, cuja validade muitas vezes não pode ser verificada com total certeza, podem ser mais ou menos plausíveis a depender do contexto analisado, mas inevitavelmente imprimem algum grau de incerteza à análise. Reconhecendo essas incertezas inerentes às avaliações de impacto, Manski (2013) recomenda a realização de **análises de sensibilidade** com o objetivo de examinar em que medida as estimativas obtidas dependem das hipóteses assumidas.

Como ilustração desse tipo de análise, considere, novamente, o exemplo do programa de vacinação. Suponha que a avaliação conclua que a política evitou o surgimento de 500 casos da doença por 100 mil habitantes. Essa estimativa, no entanto, pressupõe que a epidemia registrada em países vizinhos também teria atingido o país avaliado — sendo bloqueada pela vacina. Todavia, se essa hipótese não for verdadeira (isto é, se a epidemia não chegou a se alastrar para o país), então o impacto real da política seria significativamente menor do que o estimado — possivelmente até nulo.

A hipótese de que a epidemia teria afetado o país da mesma forma que os vizinhos é, portanto, uma das premissas às quais o resultado é particularmente sensível. Pode-se argumentar, no entanto, que é mais plausível assumir que a epidemia entrou no país do que o contrário. Nesse sentido, a estimativa de queda de 500 casos por 100 mil habitantes tende a ser mais crível do que a hipótese de ausência total de impacto da política.

[6] Os principais métodos de avaliação de impacto e suas hipóteses serão apresentados na próxima seção.

A análise de sensibilidade promove justamente esse tipo de discussão: ela convida os gestores públicos e a sociedade a refletirem sobre quais valores possíveis para o efeito da política são mais verossímeis, à luz das hipóteses consideradas em cada cenário. Esse exercício reforça a transparência da avaliação de impacto e contribui para decisões mais informadas e responsáveis.

CAPÍTULO 3.

Metodologias para estimativa do impacto

Mesmo que seja possível, em princípio, produzir informações importantes a partir dos dados de indicadores de impacto sem depender da observação do contrafactual do grupo de tratamento, é preciso assegurar que elas realmente expressem o efeito da política e não de fatores concomitantes; em outras palavras, é preciso avaliar como obter estimativas do impacto da política que não sejam enviesadas.

O viés (muitas vezes chamado de viés de seleção) ocorre, nesses casos, quando fatores que não estão sendo levados em consideração afetam a decisão por participar do programa ou do processo de seleção, fazendo com que as estimativas do impacto estejam divergindo do real impacto que a política teve. Neste capítulo, discutimos detalhadamente os tipos de métodos disponíveis para obter estimativas de impacto que buscam sempre evitar qualquer possibilidade de estimativa com viés de seleção.

Conforme explicado anteriormente no **CAPÍTULO 2**, um dos interesses em avaliar quantitativamente uma política é o de poder inferir em relações de causas e efeito a partir de evidências estatísticas.

Sendo assim, é importante que sejamos capazes de responder questões como: "Qual o impacto médio da política sobre o grupo ao qual ela foi direcionada?". Para isso, podemos utilizar uma série de métodos econométricos, sendo cada um deles mais adequado para identificar efeitos causais em diferentes contextos e fornecer parâmetros diferentes de interesse. Por exemplo, um método pode estimar o efeito médio sobre os tratados (ATT), enquanto outro pode estimar o efeito médio local do tratamento (LATE).

O principal fator para a escolha do método na avaliação *ex post* de impacto é o **desenho do processo de seleção** da política. A forma como os beneficiários são selecionados indica qual ou quais métodos são adequados para a análise de impacto. Por exemplo, o método utilizado para avaliar uma política cujos beneficiários são escolhidos de maneira aleatória (via sorteio) será diferente do método mais adequado para políticas em que participação é determinada por características observáveis (tais como idade e salário das pessoas ou o tamanho da população de uma cidade.

Dessa maneira, é muito importante que, na etapa de formulação da política, sejam pensados os critérios de elegibilidade para o público-alvo, os critérios de seleção para participação (sorteio, regra etc.), os critérios de priorização e os planos de monitoramento e avaliação, com destaque para as metodologias de avaliação de impacto. Conforme a implementação da política ocorre, cabe verificar se está sendo feito um acompanhamento dos grupos de interesse para sabermos a qualidade das informações utilizadas para a estimativa do efeito da política. Portanto, as etapas de formulação e implementação da política pública se tornam essenciais para a definição da metodologia a ser utilizada na avaliação.

Para tanto, este capítulo apresenta os métodos para estimativa do impacto no contexto da avaliação de políticas públicas. São eles: método experimental (seção 3.1), regressão descontínua (seção 3.2), pareamento (seção 3.3), diferença-em-diferenças (seção 3.4) e de controle sintético (seção 3.5). O capítulo discute o contexto em que devem ser utilizados, as hipóteses necessárias para a obtenção dos resultados[7] e o parâmetro de impacto recuperado por meio de cada método, além de exemplos de estimativas de impacto que os utilizem. Por fim, a seção 3.6 traz uma análise sobre a validade externa de cada método.

3.1 Método experimental[8]

[7] Para utilização de qualquer um dos métodos apresentados, é necessário que a hipótese de SUTVA (do inglês, Stable Unit Treatment Value Assumption) seja válida. As hipóteses adicionais necessárias para a validade de cada modelo são apresentadas ao longo deste capítulo e detalhadas no Apêndice deste guia, junto à hipótese de SUTVA.

[8] Para uma revisão de conceitos básicos de estatística ou econometria, algumas referências são: para estatística, Bussab e Morettin (2013) ou Meyer (1983) (tradução da 2ª edição); para econometria, Pereda e de Oliveira Alves (2018), Wooldridge (2010b) (tradução da 4ª edição) ou Gujarati (2011) (tradução da 5ª edição). Alguns livros e materiais mais avançados são recomendados para o ferramental econométrico utilizado nas avaliações de impacto, sendo eles os livros de Angrist e Pischke (2008), Cunningham (2018), Menezes Filho e Pinto (2017) e Wooldridge (2010a), além dos trabalhos de Athey e Imbens (2017b) e de Abadie e Cattaneo (2018). Alguns guias disponíveis em língua portuguesa, e que tratam de tópicos relacionados podem ser utilizados também como referências úteis, além de fornecerem mais exemplos de aplicações, são: Instituto Jones dos Santos Neves (2018) e Brasil (2018).

De forma geral, as estratégias empíricas consideradas ideais para estimar o efeito médio da política são as denominadas experimentais. Em essência, estas avaliam se a política gera impacto por meio de um experimento controlado, no qual indivíduos de uma população são aleatoriamente selecionados para receber o tratamento ofertado pela política. O principal mérito de um experimento controlado aleatório é sua capacidade de eliminar o **viés de seleção**, ao possibilitar a comparação entre a evolução de indivíduos tratados e não tratados após a execução da política.

Em uma política de qualificação profissional, é razoável supor que indivíduos menos qualificados tem maior interesse em participar, a fim de elevar o seu desempenho profissional, do que outros indivíduos mais qualificados. Como os menos qualificados estariam mais interessados em participar do programa, dizemos que há uma seleção amostral para a participação na política, daí o nome "viés de seleção". Note que se a política de qualificação for realizada sorteando os participantes, como no método experimental, esse problema não ocorrerá, pois a participação agora será mediante um sorteio e não via interesse.

Intuitivamente, os métodos experimentais funcionam por dois principais motivos. Ao aleatorizar quem recebe o tratamento: (i) garantimos que, em média, os indivíduos alocados nos grupos de tratamento e controle são estatisticamente idênticos em relação às suas características observáveis e não observáveis e (ii) reduz-se significativamente a probabilidade de que fatores concomitantes à política — que também possam influenciar o indicador de impacto — estejam presentes ou ocorram de forma desproporcional em apenas um dos grupos. Por exemplo, se 55% da população de interesse é composta por mulheres, espera-se que tanto o grupo de controle quanto o de tratamento tenham aproximadamente 55% de mulheres entre seus integrantes. Dessa forma, diferenças observadas no indicador de impacto entre os dois grupos não podem ser atribuídas ao fator "sexo", e sim ao tratamento em si. Essas e outras questões serão abordadas na subseção dedicada ao método de estimativa experimental.

O contexto para aplicação desse método ocorre quando há algum tipo de sorteio para definir quais unidades irão compor o grupo de tratamento e quais farão parte do grupo de controle. O termo "método experimental" refere-se ao uso de um desenho metodológico que se assemelha aos experimentos de laboratório, nos quais, por exemplo, um grupo recebe determinado medicamento e outro recebe um placebo — sendo a distribuição feita de forma aleatória.

No caso de políticas públicas, um grupo de indivíduos é selecionado aleatoriamente para receber o benefício da política, enquanto outro grupo não o recebe. A proposta é comparar os dois grupos com base em indicadores de impacto (relacionados aos objetivos da política) em um momento posterior à sua implementação.

Como a atribuição ao grupo de tratamento ou controle não está associada a nenhuma característica prévia das unidades, é possível afirmar que as diferenças observadas nos indicadores entre os grupos decorrem do recebimento da política — desde que ambos estejam expostos a contextos semelhantes após a implementação, sendo a participação em um ou outro grupo a única diferença relevante entre eles.

Imagine que o governo de uma cidade decida implementar uma política visando o desenvolvimento da infraestrutura do ambiente escolar, por meio da compra de computadores para uso dos alunos da rede pública. Como o orçamento é limitado, não será possível distribuir o equipamento para todas as escolas ao mesmo tempo, sendo preciso que algumas fiquem sem receber esses insumos nesse primeiro momento. Para lidar com essa limitação, o governo realiza um sorteio para definir quais escolas receberão os computadores. Após a implementação da política, é possível comparar indicadores relacionados ao desempenho dos alunos entre as escolas sorteadas para receber os computadores e aquelas que não foram contempladas. Caso seja observada uma diferença positiva[9] a favor das escolas que receberam os equipamentos, isso constituiu um indício favorável à efetividade da política pública.

Levando em consideração o exemplo anterior, imagine que apenas escolas da rede pública daquela cidade estavam aptas a receber a política. Portanto, o grupo de comparação, nesse caso, inclui apenas escolas da rede pública daquela localidade, mas que não foram sorteadas para receber a política. Caso incluíssemos escolas privadas ou escolas de outras cidades, estaríamos compondo o grupo de comparação com unidades que não poderiam receber a política, o que descaracterizaria o uso do método experimental para as estimativas de impacto.

O cálculo matemático para a obtenção do impacto é bastante simples, pois basta tirar a média da(s) variável(is) de interesse para cada um dos grupos (tratamento e comparação) no período

[9] Supondo que o cálculo seja feito subtraindo o valor do indicador referente às escolas que não receberam daquele das escolas que receberam.

pós-intervenção e calcular a diferença entre elas. Esse procedimento é exatamente igual à estimativa do seguinte modelo econométrico:

Equação 1 – Modelo econométrico para o método de aleatorização

$$y = \alpha + \beta D + u,$$

sendo y o indicador de impacto, D uma variável binária que indica se a observação pertence ao grupo de tratamento (igual a 1, nesse caso) ou ao grupo de controle (igual a 0, nesse caso), α o coeficiente de intercepto e u o termo de erro aleatório. O parâmetro β é o que temos interesse em estimar, pois ele indica o **Efeito Médio do Tratamento** (ATE) e irá retornar o impacto médio da política.

Supondo agora que algumas unidades sorteadas para o tratamento decidam não adotá-lo, enquanto outras que não foram sorteadas consigam, de alguma maneira, adotar o tratamento. Nessa situação, a variável D deixa de indicar se a unidade recebeu de fato o tratamento e passa a indicar se a unidade foi sorteada para receber o tratamento. Nesse caso, em vez de o parâmetro β representar o ATE, ele vai estimar o **Efeito da Intenção de Tratar** (ITT, na sigla em inglês).

Em situações como essa, se tivermos informações corretas sobre o status de tratamento de cada unidade, podemos definir uma nova variável, *Z,* que indica se a unidade foi, de fato, tratada. Com isso, é possível calcular a diferença entre a probabilidade de ter recebido efetivamente o tratamento entre os grupos sorteados e não sorteados. Esse procedimento é equivalente à estimação do seguinte modelo:

$$Z = \gamma + \delta D + \varepsilon$$

Ao fazer a divisão de β por δ, obtém-se o efeito médio de tratamento para aquelas unidades que só receberam o tratamento porque foram sorteadas para recebê-lo. Neste caso, este novo parâmetro é o **Efeito Médio de Tratamento Local** (LATE).

Considerando novamente o exemplo anterior de escolas que recebem ou não computadores novos, imagine que algumas escolas, mesmo tendo sido sorteadas para receber os computadores, decidam não os utilizá-los. Ao mesmo tempo, outras escolas que não foram sorteadas resolvam, por conta própria, adquirir os computadores e implementar a política.

Nesse cenário, é possível estimar dois estimar dois parâmetros de tratamento: o do intensão de tratar (ITT) e o de tratamento médio local (LATE).

Quando se dispõe de dados sobre as características das unidades tratadas e de controle em períodos anteriores à intervenção, é possível verificar se, de fato, os dois grupos de análise são semelhantes com base nessas características. Esse procedimento é conhecido como **teste de balanceamento do tratamento** e funciona como uma checagem da comparabilidade entre os grupos. A ideia é bastante simples: se a seleção dos participantes foi realizada de forma aleatória, espera-se que as características dos grupos de tratamento e controle sejam semelhantes, ao menos em variáveis que não envolvam diretamente o indicador de impacto analisado.

Considerando novamente o exemplo anterior de escolas que recebem ou não alguns computadores novos, imagine que tenhamos disponíveis, para o período anterior à intervenção, as seguintes variáveis: número de alunos por turma, quantidade de salas de aula, escolaridade média dos professores (em anos) e distância até o centro da cidade (em quilômetros). A intuição é que, se o sorteio foi bem executado, as médias dessas variáveis nos dois grupos serão muito próximas. Caso contrário, diferenças substanciais podem indicar problemas na aleatorização. Por exemplo, se a média de escolaridade dos professores nas escolas que receberam os computadores for de 10 anos, enquanto nas escolas que não receberam for de 15 anos, isso constitui uma evidência importante de que a seleção pode ter sido influenciada por essa variável — sugerindo que o sorteio não foi bem-sucedido em gerar grupos equivalentes.

É importante notar que o tamanho da amostra utilizada pode ajudar a distorcer as médias das variáveis utilizadas para observar o balanceamento. Por exemplo, se utilizamos 20 observações em cada grupo de análise, é possível termos distorções como a exemplificada, por mais que o sorteio tenha sido feito de maneira aleatória e realmente não haja nenhum problema com a seleção dos participantes. Por outro lado, amostras maiores que revelem diferenças em muitas variáveis entre os dois grupos podem estar sofrendo de problemas no sorteio do tratamento. Em suma, para que o processo de aleatorização gere grupos de tratamento e controle similares, é necessário que o tamanho do universo de onde se parte para fazer o sorteio seja suficientemente grande[10].

[10] Para referências bibliográficas sobre essa questão, veja a nota de rodapé 96.

Box 3.1 Subsídios para produtores de milho em Moçambique: Programa de Suporte aos Produtores

a) Contexto: A Revolução Verde foi um conjunto de inovações tecnológicas que visava o desenvolvimento do setor agrícola ao redor do mundo a partir de meados do século XX. No entanto, nos países da África Subsaariana, a adoção dessas inovações ocorreu de forma mais tardia. Foi apenas a partir de 2003, com a Declaração de Maputo, que o setor agrícola da região começou a ser transformado pelas tecnologias da Revolução Verde. Mais recentemente, alguns países africanos passaram a adotar (ou a revisar) os programas de subsídio à produção agrícola, configurando uma espécie de "segunda onda" desse movimento no continente. Essa nova fase busca incorporar os aprendizados da experiência anterior para implementar tais ações de forma mais eficaz.

Para estimar os possíveis impactos dessas inovações, Carter, Laajaj e Yang (2019) analisaram o Programa de Suporte aos Produtores em Moçambique. A política forneceu subsídios a 25.000 pequenos agricultores, por meio de vouchers que podiam ser utilizados em estabelecimentos credenciados. Esses vouchers ofereciam descontos na compra de fertilizantes e sementes modificadas para uso durante a safra 2010/2011.

b) Método: O estudo avaliou o impacto para o caso de 32 vilarejos na província de Manica. As famílias elegíveis para o sorteio precisavam cumprir alguns critérios básicos, sendo que apenas metade dos agricultores ganhariam o subsídio. Os vouchers eram nominais e possuíam data de validade, o que garantia que apenas os beneficiários sorteados pudessem utilizá-los — e dentro do período estabelecido para a avaliação. Com isso, o grupo de tratamento foi composto pelos agricultores elegíveis que receberam o voucher, enquanto o grupo de controle reuniu os elegíveis que não o receberam.

c) Resultados: Os resultados indicam que o subsídio temporário para produtores de milho em Moçambique incentivou a adoção de tecnologias da Revolução Verde (como o uso de fertilizantes, que aumentou em 78%; e de sementes modificadas, 49%), além de elevar a produção de milho (21%). Os efeitos encontrados mostraram-se persistentes para anos posteriores, mesmo com o fim do subsídio (no caso dos fertilizantes, aumento de 30%), indicando que a política gerou impactos duradouros. O estudo ainda encontrou um efeito positivo no bem-estar das famílias, mensurado por um aumento de 9% no consumo para o período após a implementação. Além disso, foram identificados efeitos de transbordamento, com impactos positivos nas redes de contato dos agricultores subsidiados. Por fim, os autores evidenciaram que a política também influenciou de forma positiva as expectativas de produção, ao incentivar a adoção das novas tecnologias.

Referência Bibliográfica. CARTER, Michael; LAAJAJ, Rachid; YANG, Dean. Subsidies And The African Green Revolution: direct effects and social network spillovers of randomized input subsidies in Mozambique. **National Bureau of Economic Research**, 2019.

Se, por um lado, os métodos experimentais são eficazes na produção de estimativas robustas do impacto de uma política, por outro, frequentemente enfrentam desafios éticos, logísticos e operacionais relacionados à sua implementação e acompanhamento.
Além disso, ao serem informadas sobre os objetivos e o contexto da política, as pessoas participantes do experimento podem alterar seu comportamento de forma artificial, comprometendo a validade dos resultados. Outro ponto crítico em experimentos é o chamado **"cumprimento parcial" da seleção** (*non-compliance*, em inglês). Isso ocorre quando indivíduos designados ao grupo de tratamento ou de controle não seguem a designação original — por exemplo, quando alguém do grupo de controle consegue acessar o benefício da política ou, inversamente, quando alguém do grupo de tratamento não o utiliza. Esse tipo de desvio pode enfraquecer a identificação do efeito causal e deve ser cuidadosamente monitorado e tratado na análise.

Por motivos como esses, muitas vezes as estratégias empíricas não experimentais acabam sendo as únicas viáveis. Em essência, esses métodos utilizam dados já disponíveis para tentar reproduzir, com algum grau de credibilidade, as condições necessárias para que uma abordagem experimental estime o parâmetro de interesse.

Nesses casos, os dados utilizados não são necessariamente produzidos com o objetivo de acompanhar a política ao longo de sua execução. Frequentemente, as avaliações não experimentais se valem de fontes de dados externas à política, mas que eventualmente captam seus efeitos de forma indireta. A principal limitação dos métodos não experimentais é que a validade dos resultados estimados depende do cumprimento de hipóteses estruturais dos modelos, muitas vezes fortes e, em grande parte dos casos, não verificáveis por meio de testes estatísticos.

Como discutido nesta seção, o método de aleatorização[11], quando corretamente aplicado, é bastante valorizado por garantir resultados estatísticos confiáveis. Porém, em diversos

[11] Uma discussão mais detalhada sobre o método, com maiores formalizações, pode ser vista no livro de Angrist e Pischke (2008). Um exemplo de aplicação do método é o trabalho de Barros et al. (2012) sobre o Programa Jovem de Futuro, no Brasil. O trabalho de Aker, Collier e Vicente (2017) utiliza a metodologia de aleatorização para estimar o impacto de educação e informação eleitoral sobre participação em eleições, para o caso de Moçambique. Também para Moçambique, temos os estudos de Batista e Vicente (2017), Hosono e Aoyagi (2018) e Custódio, Mendes e Metzger (2019). Para uma

contextos, sua implementação pode não ser viável. Nesses casos, recorremos aos métodos não experimentais ou quase-experimentais, que serão discutidos a seguir.

3.2 Desenho de Regressão Descontínua

O primeiro método quase-experimental a ser apresentado é o **Desenho de Regressão Descontínua** (RDD, na sigla em inglês para *Regression Discontinuity Design*). A adoção dessa metodologia cresceu significativamente a partir da virada do século XXI, tornando-se amplamente utilizada em avaliações de políticas públicas. Os contextos em que esse método pode ser aplicado são, em geral, bem definidos.

Frequentemente, a seleção de participantes de uma política pública é determinada por regras ou critérios baseados em alguma variável contínua, como renda, nota em exame de admissão ou idade. Nesses casos, o RDD é aplicável quando os indivíduos são elegíveis para participar da política por estarem acima (ou abaixo) de um valor de corte estabelecido nessa variável. Esse ponto de corte cria uma descontinuidade na probabilidade de tratamento, o que permite comparar indivíduos situados muito próximos a ele — pressupõe-se que esses indivíduos sejam semelhantes em todos os aspectos, exceto pelo tratamento, o que viabiliza a identificação do efeito causal da política.

Imagine que o governo de um país decida aumentar o investimento em segurança pública de algumas cidades, por meio da destinação de verbas aos municípios, específicas para este fim. Para decidir quais cidades irão receber o auxílio, o governo define que utilizará o tamanho da população de cada local: cidades com mais de 50.000 habitantes irão receber o auxílio, enquanto as que possuem menos do que 50.000 habitantes não irão recebê-lo. Neste caso, existe uma descontinuidade do tratamento para cidades com 50.000 habitantes, o que torna o método de RDD indicado para capturar os efeitos da intervenção sobre os resultados escolhidos.

A ideia central é que **a seleção sofre uma descontinuidade (Ver Box 3.2) em determinada variável** no ponto de corte: as unidades com valores acima (ou abaixo) desse valor de

aplicação do método experimental em contexto de política em Angola, temos o trabalho de Di Maro et al. (2020). Em Fazzio et al. (2020) os autores analisam o impacto de uma política educacional em Guiné-Bissau. Por fim, temos um estudo de avaliação de impacto que utiliza esse método e engloba diversos países africanos em Vicente (2014).

referência recebem o tratamento da política, enquanto aquelas com valores abaixo (ou acima) não recebem. Em contextos como esse, uma simples comparação entre as médias dos dois grupos pode ser inadequada, pois envolveria observações com características potencialmente muito distintas. O método de Regressão Discontínua (RDD) busca contornar esse problema ao considerar apenas as unidades situadas muito próximas ao ponto de corte — parte ligeiramente acima e parte ligeiramente abaixo do limiar. Parte-se do pressuposto de que essas unidades são semelhantes em praticamente todos os aspectos, exceto pela exposição ao tratamento, o que permite estimar com maior precisão o efeito causal da política.

Os casos de aplicação do RDD costumam ser bem definidos já na formulação da política, especialmente quando se estabelece a metodologia de seleção dos participantes. Sempre que essa seleção for baseada em alguma regra clara, relacionada a uma variável contínua, facilmente observável e não manipulável — como o tamanho da população de um município, a idade de indivíduos ou o número de funcionários de uma empresa —, o uso do RDD é recomendado para a avaliação do impacto da política.

Box 3.2. O conceito de descontinuidade

O conceito matemático de descontinuidade não será abordado aqui de maneira formal, mas sua intuição pode ser compreendida como uma espécie de "salto" na função: a variável dependente muda bruscamente de valor, de forma desproporcional ao comportamento anterior da função. Por exemplo, a **Figura 3.1** ilustra uma função descontínua no ponto em que a variável do eixo horizontal atinge o valor 30. Note que, ao alcançar esse valor, a função — que antes assumia valores próximos de 1.000 no eixo vertical — passa abruptamente a valores na casa de 5.000. Ou seja, há um salto repentino na função exatamente nesse ponto.
É justamente esse tipo de descontinuidade que o método de Regressão Descontínua (RDD) explora para estimar o impacto de uma política pública: ao comparar unidades situadas imediatamente antes e depois do ponto de corte, parte-se do pressuposto de que qualquer diferença observada no resultado pode ser atribuída ao tratamento.

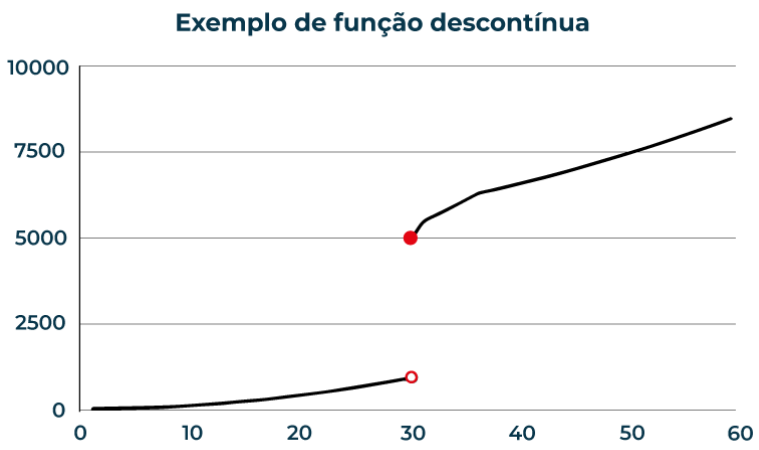


Figura 3.1. Exemplo de Função descontínua
Fonte: elaboração própria.

O método de regressão descontínua (RDD) consiste em explorar as diferenças no indicador de impacto entre observações situadas imediatamente acima e imediatamente abaixo do valor de referência. A lógica é que as unidades localizadas logo acima desse ponto têm maior probabilidade de receber o tratamento do que aquelas logo abaixo. Como essas observações são muito próximas entre si em termos da variável de corte, presume-se que sejam semelhantes em todos os demais aspectos — permitindo, assim, a identificação do efeito causal da política. Além disso, por serem observações cujos valores da característica de interesse são muito próximos do valor de referência, pode-se dizer que são, no geral, muito parecidas e, portanto, que **ao redor[12] do valor de referência o tratamento é como se fosse aleatório**.

Exemplo

Ainda considerando a política de segurança apresentada anteriormente, comparar algum indicador relacionado à ocorrência de crimes entre cidades tratadas e não tratadas pode ser inadequado, pois no grupo selecionado para receber a política, temos grandes cidades, com 1 milhão de habitantes ou mais, por exemplo, ao passo que no grupo que não é tratado, temos pequenas cidades com população próxima a 4.000 indivíduos. Assim sendo, a média do indicador de ocorrência de crimes com certeza não será comparável entre os grupos, pois em um deles, temos grandes metrópoles, enquanto no outro, temos cidades muito

[12] A ideia de estar "ao redor" do valor de referência é, de certa forma, arbitrária, e depende da unidade de medida da variável que determina o tratamento. Ainda assim, é possível estabelecer valores para servirem de distâncias que funcionem como as tolerâncias do que será definido como "ao redor" do valor de referência. Uma discussão mais aprofundada a respeito da escolha de qual o tamanho mais adequado para tais intervalos é abordada em Imbens e Kalyanaraman (2012).

pequenas. O arcabouço de Regressão Descontínua irá restringir a análise apenas para cidades com população ao redor do valor de referência de 50.000 habitantes. Por exemplo, é razoável imaginar que cidades com 51.000 habitantes se assemelhem às que possuam 49.000 habitantes. Como apenas as que possuem 51.000 habitantes estão no grupo de tratamento, o método explora a diferença do indicador de ocorrência criminal entre os dois grupos.

No arcabouço de RDD, teremos bastante facilidade para identificar as unidades que comporão cada um dos grupos de análise: o tratamento será formado por observações logo acima (ou abaixo) de algum valor de referência, enquanto o grupo de controle será formado por aquelas logo abaixo (ou acima) do mesmo valor de referência.

Dentro da metodologia de Regressão Descontínua (RDD), existem dois tipos principais: ***Sharp*** e ***Fuzzy***. A diferença entre eles está relacionada à forma como a variável contínua utilizada para a seleção das unidades influencia a probabilidade de receber o tratamento. No caso **Sharp**, o tratamento é determinado de forma totalmente mecânica por uma regra clara: todas as unidades que atendem ao critério (por exemplo, estarem acima ou abaixo de um valor de corte) recebem o tratamento, enquanto aquelas que não atendem compõem o grupo de controle. Nesse cenário, a descontinuidade na variável de corte gera uma mudança abrupta e completa na probabilidade de tratamento — que passa de 0 para 1 no ponto de corte.

Já no contexto da **RDD Fuzzy**, o valor da variável contínua define apenas a **elegibilidade** para o tratamento, mas não garante que todas as unidades elegíveis de fato o recebam — nem que todas as unidades tratadas estejam, de fato, dentro do critério de elegibilidade. Esse cenário é semelhante ao caso de *cumprimento parcial* (non-compliance) em experimentos aleatórios, e exige o uso de métodos de estimação específicos, como variáveis instrumentais, para identificar o efeito causal entre os elegíveis que foram efetivamente tratados.

3.2.1 RDD *Sharp*

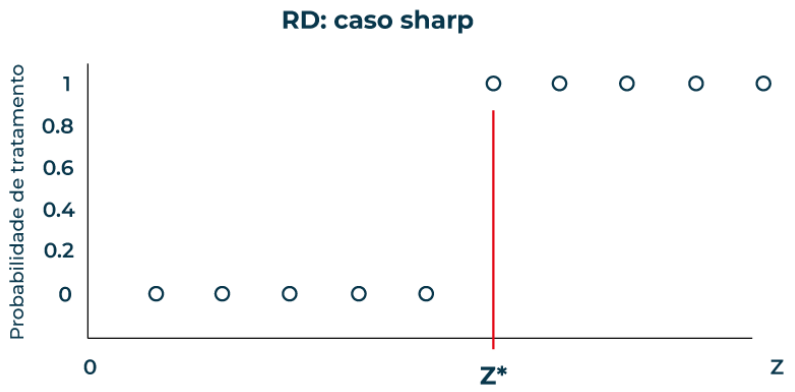
No caso do RDD *Sharp*, temos que o tratamento é determinado exclusivamente por uma variável que chamaremos de Z. Sendo assim, qualquer observação acima (ou abaixo) do valor de referência estará selecionada para participar da política, enquanto qualquer observação abaixo (ou acima) do valor de referência (Z*) estará no grupo de controle.

Para o exemplo da política de segurança, Z seria a população de cada município, enquanto Z* seria 50.000. Assim, todos os

municípios com população Z acima de 50.000 receberiam o auxílio, enquanto nenhum dos que estão abaixo deste valor receberiam.

A **Figura 3.2** ilustra a relação entre a variável que determina a probabilidade de estar no grupo de tratamento $\Pr[D = 1]$ e Z, indicando a mudança na primeira quando $Z \geq Z^*$. Note que, quando a variável Z atinge o valor Z*, há um "salto" na probabilidade de tratamento, que passa de 0 para 1. Isto é, o tratamento fica totalmente definido pela variável Z ser maior ou menor do que Z*.

Figura 3.2. RDD – caso sharp



Fonte: elaboração própria.

No caso do RDD *Sharp*, o parâmetro estimado corresponde a um LATE, pois o efeito identificado é válido apenas para as unidades localizadas próximas ao ponto de corte da política. Em outras palavras, trata-se de uma estimativa do impacto do tratamento sobre os indivíduos que estão na vizinhança imediata do valor de referência, e não para toda a população-alvo.
Esse caráter "local" do efeito é uma das principais características do RDD: a validade externa dos resultados é limitada ao entorno do ponto de corte, o que exige cautela na generalização dos achados para outros grupos.

3.2.2 RDD *Fuzzy*

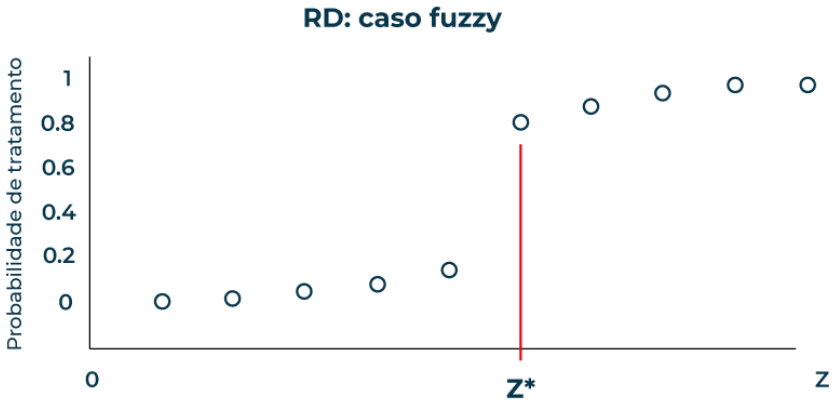
Nesse caso, o tratamento não é determinado exclusivamente apenas por uma variável Z, mas a probabilidade de ser selecionado para participação é afetada de maneira descontínua por tal variável. Sendo assim, nem todas as observações que estão acima (abaixo) do valor de

referência (Z *) são selecionadas para participar, ao passo que algumas que estão abaixo (acima) do valor de referência são selecionadas. A ideia é que existem fatores externos que influenciam na seleção, mas desde que a probabilidade de estar no grupo de tratamento $\Pr[D = 1]$ seja afetada de maneira descontínua por Z, o método pode ser aplicado.

Partindo do exemplo da política de segurança pública citado anteriormente, a ideia seria de que alguns municípios com mais de 50.000 habitantes não conseguiram o auxílio do governo por conta da necessidade de envio de um requerimento para ganhar o benefício. Como alguns municípios acabaram não enviando, mesmo sendo cidades acima do valor de referência, ficaram sem receber o benefício. Por outro lado, algumas cidades com menos de 50.000 habitantes podem acabar conseguindo o benefício por terem feito um requerimento de cadastro de reserva para o auxílio, em caso de sobra de equipamentos.

A **Figura 3.3** ilustra a relação entre a probabilidade de estar no grupo afetado pela política $\Pr[D = 1]$ e a variável Z, indicando a grande mudança na primeira quando $Z \geq Z^*$. Note, contudo, que a probabilidade de receber o tratamento já era maior do que 0 antes da variável Z atingir o Z^* de corte $Z \geq Z^*$. Similarmente, após observarmos $Z \geq Z^*$, não temos a garantia de que todas as unidades serão tratadas, já que o eixo vertical indica uma probabilidade menor do que 1 para esses casos.

Figura 3.3 RDD – caso fuzzy



Fonte: Elaboração própria.

No caso do RDD *Fuzzy*, assim como no caso do experimento com cumprimento parcial, será possível estimar dois parâmetros: o **ITT**, para a população ao redor do ponto de corte (Z*), e o **LATE**, para a população ao redor do ponto de corte que só recebeu o tratamento porque estava acima (ou abaixo) do ponto de corte. É importante notar aqui que, apesar de ambos os casos *Sharp* e *Fuzzy* estimarem um **LATE**, eles se referem a populações diferentes.

Deve-se ressaltar que, em contextos de aplicação de políticas, o caso do RDD *Fuzzy* será muito mais comum do que o *Sharp*. Ainda que a formulação da política preveja algo diferente, é muito improvável que tenhamos, na prática, todas as unidades designadas a receber a política efetivamente sendo beneficiadas. Dessa maneira, a etapa de implementação da política traz informações muito importantes para determinar a factibilidade do uso de RDD e seu tipo (*Sharp* ou *Fuzzy*), pois um monitoramento bem-feito poderá ajudar a garantir que a seleção dos participantes para receberem a política esteja acontecendo conforme o planejado.

Um dos motivos para se utilizar o RDD é que suas hipóteses são simples e não requerem condições pouco plausíveis, além de ser um método de fácil implementação. Entretanto, como a análise do método de RDD foca apenas em observações ao redor do valor de referência, **o impacto estimado pode ser interpretado apenas localmente**, ou seja, apenas para as unidades que apresentam valores próximos ao valor de referência. Essa é uma desvantagem do método de Regressão Descontínua.

Box 3.3. Aplicação do método de Regressão Descontínua

Distribuição de vacinas em diversos países do mundo: Aliança das Vacinas (GAVI)

a) Contexto: No início do século XXI, o projeto da Aliança das Vacinas (Gavi) distribuiu, com o objetivo de aumentar as taxas de imunização e reduzir a taxa de mortalidade infantil, um total de 13 bilhões de dólares em vacinas para 70 países ao redor do mundo. Aliança foi formada por uma parceria entre governos de diversos países, a Organização Mundial da Saúde (OMS), o Banco Mundial e outras instituições. Dykstra et al (2019) realizaram um estudo para estimar o impacto da distribuição de subsídios para vacinas realizada pela Gavi sobre indicadores de vacinação, testando se as vacinas compradas pelos doadores resultaram em um aumento no número de crianças vacinadas. Entre os países elegíveis para receber apoio do projeto Gavi na primeira etapa estavam Angola, Guiné-Bissau e Moçambique.

b) Método: O método de estimativa do impacto utilizado pelos autores parte da abordagem de Regressão Descontínua, aproveitando-se do fato de que a Aliança utilizou um critério de elegibilidade baseado na renda per capita (GNI per capita, em inglês). Inicialmente, países que, em 1998, excediam 1.000 dólares de renda per capita foram excluídos do auxílio. Dessa forma, foi possível explorar o fato de que países próximos a esse valor de referência eram muito parecidos, mas apenas os que estavam abaixo dos 1.000 dólares puderam receber a ajuda da Gavi.

c) Resultados: Os resultados encontrados não indicam um impacto significativo (no sentido estatístico) de receber as doações na cobertura das vacinas. Com exceção do caso da vacina

de Hepatite B, as outras quatro vacinas analisadas não exibiram resultados distintos entre os dois grupos de países analisados. No caso da vacina de Hepatite B, o impacto foi de aumento em 7 pontos percentuais na cobertura de vacinação por conta da elegibilidade para receber a ajuda. Ainda assim, os autores concluíram que não houve um desperdício de doações, pois os países puderam aplicar o dinheiro que usariam para as vacinas em outros investimentos.

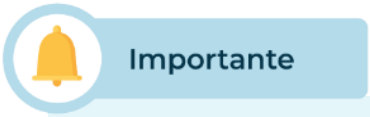
Referência. DYKSTRA, S., GLASSMAN, A., KENNY, C., & SANDEFUR, J. Regression Discontinuity Analysis Of Gavi's Impact On Vaccination Rates**. Journal of Development Economics**, 140, 12-25, 2019.

Em termos de implementação, intuitivamente, o método de RDD no caso *Sharp* poderia estimar um modelo conforme a **Equação 1**, com a diferença de que a amostra utilizada para isso ficaria restrita aos casos definidos como próximos do valor de referência. Daí vem a semelhança de arcabouço com o método experimental. Na prática, no caso *Sharp* e especialmente no caso *Fuzzy*, são feitos ajustes metodológicos para a implementação.[13]

3.3 Pareamento

Como dito anteriormente, a determinação para seleção na participação da política pode não ocorrer de maneira aleatória. Sendo assim, teremos muitos contextos em que o grupo de tratamento é definido de maneira menos rígida do que nos casos de RDD (em particular, sem um valor de referência bem definido), ou por um conjunto maior de variáveis observáveis. Quando esse for o caso, teremos a possibilidade de aplicar a metodologia de pareamento.

[13] Os detalhamentos destas metodologias são facilmente encontrados nas referências sugeridas a seguir e, em particular, no Capítulo 7 de Menezes Filho e Pinto (2017), em português. Algumas referências para maiores detalhamentos e formalizações a respeito da estimação pelo método de Regressão Descontínua são o livro de Angrist e Pischke (2008) ou o de Cunningham (2018), além dos trabalhos de Hahn, Todd e Van der Klaauw (2001) e o de Lee e Lemieux (2010). Uma avaliação do impacto via RD do Programa Bolsa Família, do Brasil, sobre indicadores educacionais pode ser encontrada em Romero e Hermeto (2009). Além disso, Anderson (2017) utiliza o mesmo método para avaliar o impacto de um sistema jurídico que utiliza direito comum sobre taxas femininas de HIV para alguns países da África subsaariana. Uma aplicação da metodologia de Regressão Descontínua para a análise do impacto sobre a idade penal e crimes no Brasil pode ser vista no artigo de Costa et al. (2018).



Características observáveis são aquelas que podem ser mensuradas diretamente e para as quais existem dados disponíveis para a realização de avaliações de impacto — como a idade de uma pessoa, o número de funcionários de uma empresa ou o PIB de um país. Por outro lado, há características não observáveis, cujos dados podem não estar disponíveis (como as habilidades cognitivas de uma pessoa) e/ou que não são diretamente mensuráveis.

A intuição desse método de estimativa de impacto é bem simples. Se a seleção para participar do grupo de tratamento for determinada apenas por um conjunto de **características observáveis**, então é possível utilizá-las para escolhermos observações não tratadas para compor o grupo de comparação que sejam parecidas com cada uma do grupo tratado. Dessa maneira, ao dizer que as observações são suficientemente parecidas, podemos comparar a diferença do(s) indicador(es) de impacto entre os dois grupos para mensurar o impacto da política. Portanto, a abordagem fica bem parecida com a de RDD: como estamos supondo que as observações do grupo de tratamento e do grupo de comparação são bastante parecidas (a partir de algumas variáveis observáveis), **é possível dizer que o tratamento tem um caráter aleatório após condicionarmos as variáveis que influenciam na seleção**.

Imagine uma política pública em uma cidade que busca implementar programas de qualificação e melhoria no trabalho para o desenvolvimento de carreiras, selecionando os participantes a partir de variáveis como sexo, idade e número de filhos. Suponha que uma mulher de 30 anos, com um filho, tenha sido selecionada para participar dos programas de qualificação. A ideia do método de pareamento consiste em encontrar uma pessoa semelhante a ela — por exemplo, outra mulher, também com 30 anos e um filho — que não tenha sido selecionada para participar. Como a seleção depende apenas dessas três variáveis e as duas pessoas são muito parecidas entre si, é possível calcular a diferença em algum indicador de impacto relacionado ao trabalho e atribuir essa diferença ao efeito do tratamento.

A composição do grupo de tratamento, nesse caso, é feita por aquelas unidades que receberam a política, enquanto o grupo de comparação é composto por unidades muito parecidas com as que participaram do tratamento, mas que não o receberam. Na prática, o método cria o grupo de controle escolhendo para cada unidade tratada o seu par não tratado mais semelhante possível.

Como mencionado anteriormente, o conceito de "muito parecidas" é definido exclusivamente com base em um conjunto de características observáveis. Dessa forma, o grupo de comparação pode incluir pessoas que atendem aos critérios para participar da política, mas que, por algum motivo, não foram selecionadas, ou ainda cidades com características similares às que receberam o tratamento, mas localizadas em regiões diferentes daquelas onde a política foi implementada.

Após o exercício de encontrar um par para cada unidade do grupo tratado, é preciso obter a média condicional da(s) variável(is) de interesse a certas características para cada um dos grupos (tratamento e pares) e calcular a diferença entre elas, obtendo o parâmetro de **efeito médio do tratamento sobre os tratados** (ATT). A intuição é de que **os pares representam bem o cenário contrafactual das observações do grupo de tratamento**.

Existem algumas dificuldades decorrentes da implementação do método de pareamento[14]. A principal delas diz respeito à dimensionalidade de variáveis que determinam o pareamento, isto é, a quantidade de categorias criadas na medida em que mais características são necessárias para a seleção de atendimento pela política. A ideia é que, conforme mais variáveis determinam o tratamento, fica cada vez mais difícil encontrar pares exatos para as observações tratadas e, em alguns casos, os pares nem existirão. Além disso, o método requer que variáveis contínuas (como o rendimento, por exemplo) tenham que ser **discretizadas** ou transformadas em intervalos. Diante disso, temos a possibilidade de utilização do pareamento via escore de propensão (EP).

A implementação do método de pareamento apresenta algumas dificuldades. A principal delas diz respeito à dimensionalidade das variáveis utilizadas no pareamento, ou seja, ao aumento no número de categorias à medida que mais características são consideradas na seleção dos participantes da política. Quanto maior o número de variáveis que determinam o tratamento, mais difícil se torna encontrar pares exatos para as observações tratadas — e, em alguns casos, esses pares podem nem existir. Além disso, o método exige que variáveis contínuas (como o rendimento) sejam **discretizadas** ou agrupadas em intervalos, o que pode levar à perda de

[14] Uma referência para maiores detalhamentos e formalizações do método de pareamento é o livro de Angrist e Pischke (2008) e o trabalho de Abadie e Cattaneo (2018). Uma aplicação que utiliza a técnica de pareamento para o caso de uma política de incentivo a atividades tecnológicas no Brasil, denominado Programa de Desenvolvimento Tecnológico Industrial pode ser encontrada em Avellar e Alves (2008). O artigo de Resende e Oliveira (2008) avalia o impacto da política de transferência de renda Bolsa-Escola, no Brasil, sobre consumo de famílias. Casini, Riera e Santos Monteiro (2014) avaliam uma política de acesso a crédito em Cabo Verde a partir do método de pareamento via escore de propensão. Em Djimeu (2014) o autor utiliza o mesmo método para estimar o impacto de uma política de saúde infantil para Angola. Cunguara e Darnhoffer (2011) e Nkala, Mango e Zikhali (2011) avaliam o impacto de políticas relacionadas à agricultura em Moçambique também usando esse método.

informação. Diante dessas limitações, uma alternativa é a utilização do **pareamento por escore de propensão (EP)**.

Importante

Ao falar em **discretização** de uma variável contínua, referimo-nos ao processo de agrupar variáveis que assumem infinitos valores possíveis dentro de um intervalo — como os rendimentos salariais — em faixas ou categorias. Por exemplo, no caso dos salários, é possível classificá-los com base no salário-mínimo (s.m.), criando faixas como: menos de 1 s.m., entre 1 e 2 s.m., mais de 2 s.m., e assim por diante.

Esse tipo de classificação frequentemente utiliza **características observáveis**, ou seja, aquelas que podem ser mensuradas diretamente e para as quais existem dados disponíveis, como a idade de uma pessoa, o número de funcionários de uma empresa ou o PIB de um país. Por outro lado, há **características não observáveis**, cujos dados podem não estar disponíveis (como habilidades cognitivas) e/ou que não são diretamente mensuráveis.

A metodologia de pareamento via EP é parecida com a de pareamento, com a vantagem de ser unidimensional com relação às características que determinam a seleção para o tratamento. A ideia dessa abordagem é estimar a probabilidade de tratamento $\Pr[D = 1]$ a partir das variáveis X que determinam a participação no grupo que recebe a política. Essa estimativa vai calcular[15], portanto, um valor de probabilidade de recebimento do tratamento $p(X)$ a partir das variáveis observadas. Perceba que, independentemente da quantidade de variáveis em X, $p(X)$ é unidimensional, isto é, funciona como apenas uma única variável e, portanto, facilita a obtenção de um par. Além disso, como $p(X)$ é uma probabilidade, os valores ficam restritos ao intervalo entre zero e um. Intuitivamente, iremos observar o valor de $p(X)$ para cada observação, tanto do grupo tratado quanto do grupo de comparação[16]. Depois, para cada observação do grupo tratado, iremos buscar, no grupo de comparação, a observação com o valor $p(X)$ mais próximo. Feito isso, comparam-se os indicadores dos dois grupos para estimativa do impacto via efeito médio do tratamento sobre os tratados (ATT). Em termos de hipóteses, elas estão descritas no **apêndice deste guia**.

[15] Normalmente, essa estimação é feita por métodos não-lineares como o Logit ou Probit, também chamados de "modelos de escolha binária" ou "modelos de escolha discreta". Para mais detalhes, ver Wooldridge (2010b, cap. 17) ou Wooldridge (2010a, cap. 15).

[16] Uma dúvida frequente é achar que somente para os tratados o $p(X)$ será estimado, pois sabe-se que os não tratados não foram contemplados de fato pela política. Entretanto, a ideia do método de pareamento é buscar casos de observações que não participaram da política, mas possuíam uma alta probabilidade de participar, por exemplo, que sirvam para representar a situação contrafactual daqueles que foram tratados e tinham um alto escore de propensão.

Uma última discussão relevante relacionada aos métodos de pareamento, discutida no Box 3.4, diz respeito ao método de seleção de pares para composição do grupo de comparação. Para comparar observações com probabilidade próxima de receber o tratamento, são escolhidas unidades parecidas — sendo uma tratada e a outra não. No entanto, existem diferentes estratégias de pareamento: é possível optar por selecionar apenas um par para cada unidade tratada ou, alternativamente, adotar uma abordagem mais flexível, escolhendo vários pares para cada unidade tratada. A discussão a seguir aborda brevemente essas possibilidades e suas implicações.

Box 3.4. Decisões sobre viés e variância: métodos de seleção do número de pares

A seleção da quantidade de pares para cada observação do grupo tratado envolve uma troca entre viés e variância na estimativa do impacto da política. Por um lado, métodos como o do vizinho mais próximo (*Nearest Neighbor*, em inglês), selecionam apenas um par para cada unidade tratada, priorizando a similaridade máxima entre os pares — o que tende a reduzir o viés, mas pode aumentar a variância da estimativa devido ao uso de menos informações. Por outro lado, métodos como o de **n-vizinhos**, **Kernel** ou **pareamento por raio** utilizam mais de um par para cada unidade tratada. Esses métodos ampliam a base de comparação, o que pode reduzir a variância da estimativa, mas ao custo de incluir observações menos semelhantes às tratadas — o que pode introduzir viés. Essa discussão não possui uma resposta definitiva e é explorada em profundidade por Caliendo e Kopeinig (2008).

Box 3.5. Aplicação do método de pareamento para a avaliação de uma política pública

Política agroflorestal de pagamento por serviços ambientais de sequestro de carbono em Moçambique: Projeto Nhambita

a) **Contexto**: Projetos de pagamento por serviços ambientais (PES, na sigla em inglês) baseiam-se na ideia de que os fornecedores de serviços ambientais devem ser compensados (normalmente, em dinheiro), enquanto aqueles que utilizam os benefícios devem arcar com os custos.

Recentemente, a utilização de PES cresceu em diversos países. Um projeto de pequena escala visando o sequestro de carbono, por exemplo, foi implementado em uma região de Moçambique a partir de 2002. Os participantes eram agricultores de pequenas propriedades que assinavam um contrato de maneira voluntária, sob o qual deveriam plantar árvores em suas fazendas, mantendo-as por 25 anos para ganhar pagamentos em dinheiro. Um sistema

denominado Plano Vivo (Plan Vivo system) garantia o monitoramento do sequestro de carbono. Já a redução de carbono era vendida no mercado voluntário internacional.

O estudo conduzido por Hedge e Bull (2011) buscou avaliar os impactos da política agroflorestal de sequestro de carbono sobre indicadores familiares como a renda e consumo, além de variáveis relacionadas à produção agrícola. O estudo foi feito com uma amostra aleatória de 290 famílias em uma região da província de Sofala, Moçambique, que compreende cinco vilarejos. Um questionário identificava algumas características das famílias, tanto que participavam do projeto quanto as que não participavam.

b) Método: A avaliação utilizou a metodologia de pareamento via escore de propensão para estimar o impacto da política, selecionando pares a partir de características disponíveis no questionário. A seleção dos pares se deu por quatro diferentes técnicas (vizinho mais próximo, raio, kernel e estratificação).

c) Resultados: Os autores encontraram efeitos positivos da política sobre a renda familiar e o consumo. Além disso, observaram uma redução na produção líquida das colheitas, o que é compatível com a diminuição da área disponível em razão do plantio de árvores. Já no caso do uso de produtos derivados da floresta, também houve uma redução; no entanto, diferentemente dos demais impactos, esse efeito não apresentou significância estatística.

Referência Bibliográfica. HEDGE, R. e BULL, G.Q. Performance Of An Agro-forestry Based Payments-for-Environmental-Services Project In Mozambique: a household level analysis. **Ecological Economics,** 71, pp.122-130, 2011.

3.4 Diferença-em-Diferenças

A aplicação do método de pareamento depende da hipótese de que a seleção é determinada por variáveis observáveis, o que pode não ser muito factível em determinados contextos de políticas públicas. Muitas vezes, o que leva à participação ou não no tratamento é definido por **características não observáveis** das unidades de análise da política, como práticas de governança em cidades e países ou a motivação de indivíduos. Nesses casos, não será possível aplicar o ferramental de pareamento, já que esse método pressupõe que as variáveis que determinam a participação sejam observáveis. É nesse contexto que se utiliza o método de diferença-em-diferenças (*Difference-in-Differences – DiD*), que permite estimar efeitos causais mesmo quando há fatores não observáveis influenciando a seleção.

Intuitivamente, o método de diferença-em-diferenças elimina o efeito de qualquer característica — observável ou não — que influencie a participação no tratamento, desde que essa característica permaneça constante ao longo do tempo em torno do período de

implementação da política. Para isso, é preciso que seja definido um grupo de comparação similar (com relação ao contexto e comportamento da(s) variável(is) de interesse) ao tratado, com a diferença de que apenas os últimos recebem o tratamento. Além disso, precisamos observar cada um dos dois grupos em pelo menos dois momentos distintos no tempo: um (ou mais) anterior à ocorrência da política e um (ou mais) posterior à sua implementação.

A principal hipótese do método é a de tendências paralelas entre os dois grupos, isto é, que a trajetória do(s) indicador(es) do grupo de controle represente bem o comportamento contrafactual do grupo tratado. Portanto, estamos dizendo que, se não ocorresse a política, o comportamento do indicador do grupo que participa da política ao longo do tempo seria similar ao que de fato se observa do grupo de comparação. O método consiste em comparar como cada um dos grupos evoluiu ao longo do tempo, confrontando os cenários anterior e posterior à política. Dessa maneira, recupera-se o impacto médio do tratamento pelo parâmetro de **efeito médio do tratamento sobre os tratados** (ATT). Para uma discussão mais aprofundada sobre a utilização do método da diferença-em-diferenças, incluindo novos avanços da literatura, recomendamos a leitura do guia "Diferença-em-Diferenças"[17].

A **Figura 3.4.** ilustra a ideia básica do método para realizar a estimativa de impacto. Nela, temos a trajetória do indicador do grupo tratado dada pela curva de cor preta, enquanto a curva cinza indica a trajetória do indicador para o grupo de controle. A trajetória de cor preta pontilhada representa o cenário contrafactual para o grupo tratado, isto é, mostra a trajetória que o indicador teria, para o grupo tratado, na ausência da política. Supondo que o período "0" seja bem perto da implementação da política e o período "1" seja posterior à política ser colocada em prática, teremos o impacto dado pela diferença entre os pontos indicados. Note que o ponto vermelho não pode ser observado, já que é um cenário contrafactual. Dessa maneira, utiliza-se a trajetória do grupo de controle para estimar o ponto vermelho que **representa o valor do indicador para o grupo tratado na ausência da política**.

[17] Disponível em www.fgvclear.org/biblioteca/

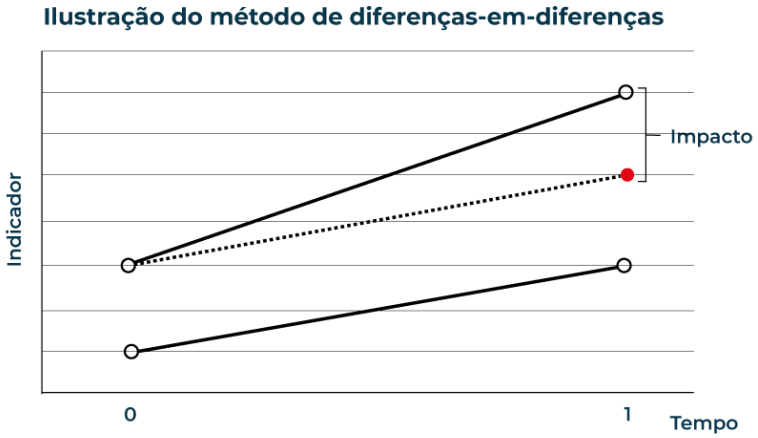


Figura 3.4. Ilustração do método de diferenças-em-diferenças
Fonte: elaboração própria.

Vamos imaginar um caso em que, além de características observáveis como sexo e idade, características individuais não observáveis, como motivação ou habilidades socioemocionais, também influenciam a seleção para o tratamento — por exemplo, quando essas características são avaliadas qualitativamente durante entrevistas no processo seletivo de participação. Nesse cenário, é possível que o grupo de tratamento seja composto por pessoas mais motivadas e habilidosas do que aquelas do grupo de comparação. Essa diferença pode se refletir no indicador de impacto, fazendo com que o grupo tratado apresente, em média, um desempenho mais elevado não apenas em razão da política, mas também porque seus integrantes são, por natureza, mais esforçados ou habilidosos. Esse tipo de viés compromete a validade da estimativa do impacto, podendo levar a resultados enganosos sobre a real efetividade da política.

Por outro lado, **se formos capazes de observar os indivíduos em um momento anterior à implementação da política e outro posterior, podemos eliminar o problema causado por características que não variam ao longo do tempo**.

Continuando o exemplo da seção 3.3, vamos supor que, além de idade, sexo e número de filhos, características não observáveis dos indivíduos sejam fatores determinantes para a participação no programa. Neste caso, o método de DiD seria indicado, uma vez que este seria capaz de capturar o efeito destas características na estimação do impacto da política.

Supondo que características como esforço e habilidade não variem ao longo do tempo, podemos calcular a média do indicador de impacto para cada grupo (tratado e de comparação) em dois momentos distintos: antes e depois da implementação da política. Em seguida, tomamos a diferença entre os dois períodos para cada grupo, capturando a evolução temporal de cada um.

Como os fatores não observáveis em questão (como esforço e habilidade) são constantes ao longo do tempo, seu efeito é eliminado nesse cálculo — ou seja, eles são "diferenciados para fora". Com isso, não há mais influência dessas características na estimativa do impacto.
O passo final consiste em tomar a diferença entre essas duas diferenças, processo que dá nome ao método de diferença-em-diferenças (DiD). Essa operação permite isolar o efeito da política, controlando para características fixas e não observáveis que poderiam enviesar a análise.

Suponha que estejamos medindo o impacto de uma política pública sobre os salários, com o objetivo de verificar se a qualificação promovida por ela resultou em melhores remunerações para as mulheres do grupo tratado. Nas duas primeiras colunas da Tabela 3.1, são apresentadas as médias salariais de cada grupo (listados nas linhas) nos períodos anterior e posterior à implementação da política. A última coluna mostra a diferença temporal dos salários dentro de cada grupo, ou seja, a evolução dos salários ao longo do tempo analisado.
É importante observar que, se utilizássemos apenas a variação salarial do grupo tratado (um aumento de 28 dólares) para medir o impacto da política, estaríamos superestimando seu efeito. Isso porque, ao analisarmos a evolução dos salários no grupo de controle no mesmo período, também identificamos um aumento — o que indica que os salários já estavam em trajetória de crescimento, independentemente da política.
Portanto, para estimar corretamente o impacto da política sobre o desempenho no mercado de trabalho, devemos calcular a diferença entre as evoluções dos dois grupos — procedimento conhecido como diferença-em-diferenças. Segundo o valor apresentado na última linha da tabela, o impacto real da política sobre os salários foi um aumento médio de 13 dólares, valor significativamente menor do que o estimado por uma simples comparação antes e depois apenas dentro do grupo tratado.

Tabela 3.1 - Exemplo de aplicação do método de diferença-em-diferenças

Variável: salários (média), em dólares	Pré implementação	Pós implementação	Diferença (pós - pré)
Mulheres que participaram (tratadas)	103	131	28
Mulheres que não participaram (comparação)	97	112	15
Diferença-em-diferenças	-	-	**13*** **

Fonte: Elaboração própria.

Nota: Os valores são fictícios e não representam nenhum caso real de política.

*** Indica que a diferença de médias é estatisticamente significativa ao nível de 10%

Assim como no método de pareamento e de aleatorização, o grupo de tratamento no arcabouço de DiD é composto pelas observações selecionadas para receber o tratamento. O grupo de controle, por sua vez, é composto de maneira similar ao do pareamento, ou seja, a partir de observações similares às do grupo de tratamento, mas que não são selecionadas para participar da política.

A diferença aqui é que precisamos que o grupo de controle tenha um **comportamento similar** ao grupo tratado com relação aos indicadores de impacto da análise, pois desejamos ter a melhor representação possível do cenário contrafactual das unidades tratadas. Essa ideia de "comportamento similar" ficará mais clara no **apêndice deste guia**, onde são detalhadas as hipóteses do método.

O cálculo matemático para a obtenção do impacto é bastante simples e se assimila ao da aleatorização, apenas com um passo a mais. Primeiramente, basta calcular, para cada um dos grupos de análise (tratamento e comparação), a média da(s) variável(is) de interesse para os dois (ou mais) momentos de tempo (sendo, pelo menos, um anterior e outro posterior à implementação da política). Em seguida, com essas diferenças em mãos, é preciso apenas realizar a diferença entre as evoluções (diferenças) delas ao longo do tempo – daí o nome do método. Esse procedimento é exatamente igual à estimativa do seguinte modelo econométrico:

Equação 2 – Modelo Econométrico para o Método de Diferença-em-diferenças

$$y = \alpha + \gamma D + \delta T + \beta (D \cdot T) + u,$$

sendo y o indicador de impacto, D uma variável binária que indica se a observação pertence ao grupo de tratamento (igual a 1, nesse caso) ou ao grupo de controle (igual a 0, nesse caso), T uma variável binária que indica se o momento de tempo é posterior à implementação da política (igual a 1, nesse caso) ou anterior (igual a 0, nesse caso), α o coeficiente de intercepto, γ e δ são parâmetros de inclinação e u o termo de erro aleatório. O parâmetro β é o que temos interesse em estimar nesse caso, associado à interação entre as duas variáveis binárias, indicando valor 1 para observações do grupo de tratamento e que estejam em um momento do tempo posterior ao início da política, mensurando o efeito médio do tratamento sobre os tratados (ATT) da política.

Box 3.6. Aplicação do método de diferença-em-diferenças para a avaliação de uma política pública

Tratamentos para combate à malária em Angola a partir da introdução de Terapias Combinadas à base de Artemísia (TCA): Programa Nacional de Combate à Malária (PNCM).

a) Contexto: A malária tem configurado como um problema de saúde recorrente no continente africano. Nesse contexto, muitos países da África Subsaariana passaram a subsidiar Terapias Combinadas à base de Artemísia (TCA), que se tornaram a principal ferramenta no enfrentamento da malária. Outro método amplamente utilizado na prevenção da doença é o uso de mosquiteiros, que funcionam como redes de proteção contra os mosquitos transmissores. Em muitos casos, esses mosquiteiros são instalados sobre as camas — especialmente as de crianças pequenas, grupo considerado mais vulnerável aos efeitos da malária.

No caso de Angola, as TCAs foram introduzidas a partir de 2006, por meio do Programa Nacional de Combate à Malária (PNCM), em parceria com a UNICEF e a Organização Mundial da Saúde (OMS). O estudo de Berthélemy et al (2015) avalia o impacto da introdução do tratamento contra malária sobre comportamentos relacionados à prevenção, a partir de pesquisas realizadas em 2006 e 2011 pelo *Malaria Indicators Survey* (MIS). As unidades de observação avaliadas são crianças com menos de 5 anos, pois, para esse caso, os tratamentos existiam desde o início da política.

b) Método: A estimativa do impacto parte do método de diferença-em-diferenças, utilizando o fato de que a intensidade do tratamento foi diferente de acordo com recortes geográficos, fazendo com que a exposição de indivíduos aos TCA tenha sido distinta. Dessa maneira, províncias que possuem maior acesso ao tratamento à doença a partir de TCAs puderam ser comparadas com as que possuem menos.

c) Resultados: Os impactos estimados sugerem que o aumento no acesso ao tratamento com base na Artemísia pode ter causado um efeito negativo sobre o uso de mosquiteiros (ITN). Em termos de magnitude, em dada província, o efeito de um aumento de 1 ponto percentual no uso de TCAs sobre o uso de ITNs foi estimado em uma redução de cerca de 5,2 pontos percentuais. Sendo assim, o aumento de tratamento contra a doença estaria negativamente correlacionado com a utilização de métodos preventivos, o que mostra que a sinergia entre os dois objetivos da política pode estar intrincada, já que o aumento de um indicador fez com que o outro tivesse redução.

Referência Bibliográfica. BERTHÉLEMY, J. C., DOUBLIEZ, V., & THUILLIEZ, J. Prevention Or Treatment? The introduction of a new antimalarial drug in Angola. ⟨halshs-01244406⟩, 2015.

Uma grande vantagem desse método é que ele elimina efeitos de características observáveis e não-observáveis invariantes no tempo das unidades de análise da política, garantindo uma estimativa de impacto que não sofra problemas de viés relacionados com essas características. Além disso, como precisamos apenas que o comportamento do(s) indicador(es), ao longo do tempo, seja parecido entre os dois grupos de análise, podemos usar dois grupos distintos no que diz respeito ao nível absoluto desse(s) indicador(es). Por outro lado, ao longo do tempo, a mudança na composição dos grupos pode comprometer a estimativa de impacto que utiliza esse método.

O **apêndice deste guia** traz algumas discussões mais avançadas relacionadas ao método, tais como a combinação do mesmo com o método de pareamento, a implementação do arcabouço de DiD com mais de um período pós-tratamento e o conceito de dados em painel.[18]

3.5 Controle Sintético

Uma alternativa que vem sendo utilizada mais recentemente na literatura de análise de impactos é a do Controle Sintético (SC, do inglês, *Synthetic Control*). Este método permite estudar o impacto de políticas em contextos em que existe apenas uma unidade sendo afetada pela política (grupo tratado)[19] e diversas unidades não afetadas (grupo de tratamento). O SC realiza uma combinação das unidades do grupo de controle para tentar criar artificialmente uma unidade que represente bem o que ocorreria com a unidade tratada caso ela não tivesse sido afetada pela política. Assim como no arcabouço de diferença-em-diferenças, o uso do SC requer dados observados antes e depois da implementação da política, permitindo a comparação da evolução de uma ou mais variáveis de interesse ao longo do tempo.

[18] Maiores detalhamentos sobre a metodologia de diferença-em-diferenças podem ser encontrados em Bertrand, Duflo e Mullainathan (2004), Angrist e Pischke (2008) e Cunningham (2018). Um exemplo de aplicação do método para avaliação de política trata do caso da Lei Maria da Penha, no Brasil, que pode ser visto em Cerqueira et al. (2015). O trabalho de Pereira Filho, Sousa e Alves (2018) avalia o impacto de guardas municipais em municípios do Brasil utilizando a metodologia de DD. Vicente (2010) utiliza indivíduos de Cabo Verde como grupo de controle para os de São Tomé e Príncipe para avaliar o impacto de recursos naturais sobre a percepção de corrupção por parte da população, a partir da metodologia de diferença-em-diferenças. Jindal, Kerr e Carter (2012) avaliam o impacto de políticas agrícolas de sequestro de carbono em Moçambique a partir da mesma metodologia. Alguns trabalhos de avaliação utilizam diversos países ao usar o método de DD, como em Jakubowski et al. (2017), Greßer e Stadelmann (2019) e Jaupart, Dipple e Dercon (2019).

[19] O método de controle sintético também pode ser usado quando, ao invés de apenas uma unidade tratada, houver uma quantidade pequena de unidades tratadas pela política/intervenção conforme explicado em Abadie, Diamond e Hainmueller (2015).

Ao "criar" um grupo de controle, a lógica do método de Controle Sintético (SC) se aproxima da combinação entre os métodos de pareamento e diferença-em-diferenças (DiD). Diferentemente do pareamento tradicional, que seleciona unidades específicas para compor o grupo de comparação com base em alguma métrica de similaridade, o SC utiliza todas as unidades potenciais do grupo de controle para construir um controle artificial por meio de médias ponderadas das variáveis de interesse. Essas ponderações serão obtidas utilizando-se toda a informação relevante, **anterior à política**, que tivermos disponível na amostra de dados[20]. Como só possuímos uma unidade de observação no grupo tratado, **iremos compor um grupo de controle que contenha também apenas uma unidade, denominando-o de "controle sintético"**. Nesse contexto, ao estimarmos o que teria ocorrido com a unidade tratada na ausência da política, estamos calculando o **efeito médio sobre os tratados (ATT)**.

O método de SC é muito comum em aplicações de avaliações de políticas em unidades de observação mais agregadas, como cidades, estados ou países. Além disso, há uma crescente literatura a respeito, dada a sua relativamente recente aplicação no campo da avaliação de impacto.[21]

Imagine que um determinado estado tenha sofrido diversas enchentes devido às fortes chuvas durante o inverno. Para definir o volume de recursos destinados à região, o gestor público deseja avaliar os impactos sobre a produção agrícola daquele estado. Nesse cenário existe apenas uma unidade tratada (o próprio estado), e é pouco provável que qualquer outro estado, individualmente, funcione como um bom controle. Por isso, o método do Controle Sintético se apresenta como uma boa alternativa para estimar estes impactos, ao permitir a construção de um grupo de controle artificial que combine informações de vários estados não afetados para representar o que teria ocorrido na ausência das enchentes.

Na implementação do método de controle sintético, alguns pontos são importantes para a estimativa do impacto:

[20] Os artigos de Doudchenko e Imbens (2016) e de Ferman, Pinto e Possebom (2020) discutem a respeito da construção dos pesos e cuidados na escolha das variáveis para seu cálculo. Kaul et al. (2016) também discutem a respeito da utilização da variável dependente defasada para a composição da unidade sintética.

[21] Algumas referências que abordam detalhadamente o uso do método de controle sintético são Abadie, Diamond e Hainmueller (2015), Athey e Imbens (2017b) e Cunningham (2018). O trabalho de Cerqueira et al. (2020) avalia o impacto do Estado Presente em Defesa da Vida, no Brasil, que tinha como objetivo reduzir problemas relacionados à violência. Ebeke, Mansour e Rota-Graziosi (2016) e Kassa e Coulibaly (2019) são dois estudos que analisam impactos em diferentes indicadores para alguns países africanos. Algumas outras referências para ver aplicações práticas de CS são Abadie, Diamond e Hainmueller (2010) e Montalvo (2011).

- A amostra de dados deve **conter informações para as observações em diversos períodos anteriores à implementação da política**. Dessa maneira, poderemos construir a unidade sintética com mais similaridade possível à unidade tratada, já que teremos em mãos vários períodos.
- Precisamos que o grupo de controle contenha apenas unidades não afetadas em sua composição. Portanto, unidades que também são afetadas por políticas similares ou pela própria política (às vezes por um efeito de transbordamento, por exemplo) não podem fazer parte do grupo de unidades que irá compor a unidade sintética.
- Da mesma forma, qualquer unidade cujo(s) indicador(es) de interesse tenha(m) sido afetado(s) por grandes choques ocorridos no mesmo período da política deve ser excluída do conjunto de unidades utilizadas para compor o controle sintético — especialmente se tais choques não afetariam a unidade tratada em um cenário contrafactual.
- Além disso, é importante garantir que as unidades incluídas no grupo de controle apresentem algum grau mínimo de similaridade com a unidade tratada. Isso evita o problema conhecido como overfitting, que ocorre quando o controle sintético replica quase perfeitamente os valores da unidade tratada nos períodos anteriores à política — não porque a composição seja apropriada, mas porque o grande número de unidades disponíveis permite que choques aleatórios nos períodos pré-intervenção expliquem artificialmente a similaridade nos indicadores.

Uma maneira simples e intuitiva de testar a qualidade da estimativa via SC é por meio de um experimento de "placebo". Nesse procedimento, ao calcular o impacto da política, realizamos uma troca: substituímos a unidade tratada por uma das unidades de controle e repetimos a estimativa.

Se o impacto da política for real, ele deverá aparecer apenas na unidade tratada. Nos testes placebo, não se espera observar diferenças significativas entre os indicadores da unidade "falsamente tratada" e seu controle sintético nos períodos posteriores à implementação da política. Esse exercício pode ser repetido sucessivamente, permutando a unidade tratada com cada uma das unidades de controle, para verificar se os efeitos estimados nos placebos são consistentemente menores do que aquele obtido na especificação correta — ou seja, quando a unidade realmente afetada pela política é tratada como tal, e as demais compõem o controle sintético.

Box 3.7. **Aplicação do método de controle sintético em uma avaliação de impacto**

Entrada de Guiné-Bissau na zona do Franco CFA: impactos sobre indicadores econômicos

a) Contexto: A União Econômica e Monetária do Oeste Africano (UEMOA) foi criada em 1994 por países majoritariamente com passado colonial francês, com o objetivo de aprimorar acordos anteriores e intensificar a integração econômica entre os membros. O bloco permitia a livre circulação de pessoas, serviços e capital monetário entre os países participantes. Em 1997, a Guiné-Bissau passou a integrar a UEMOA, adotando o Franco CFA como moeda. A economia do país é fortemente dependente de atividades agropecuárias no litoral, o que torna produtos como arroz e pescados particularmente relevantes para a economia local. O estudo conduzido por Ndiyae (2020) avaliou o impacto da adesão ao Franco CFA sobre o crescimento econômico da Guiné-Bissau, com base em indicadores como PIB per capita, produção de arroz e volume de pesca.

b) Método: Como a política foi implementada, naquele momento, em apenas uma unidade de observação — a Guiné-Bissau —, o método escolhido para a avaliação de impacto foi o de Controle Sintético. Para compor a unidade de controle sintético, foram selecionados 39 países (ou 8, em uma especificação mais restrita) que nunca integraram a união monetária do Franco CFA, mas que apresentavam semelhanças econômicas com a Guiné-Bissau. A seleção foi baseada em variáveis como indicadores de conflitos e índices de exportação e importação no período de 1985 a 1997. Entre os países candidatos a compor o controle sintético, destacam-se, por exemplo, o Brasil e Moçambique.

c) Resultados: Os resultados sugeriram que a entrada na zona da união do CFA contribuiu para a queda no PIB per capita de Guiné-Bissau, além de impactos também negativos no volume de pesca e de produção de arroz. O gráfico abaixo ilustra o resultado para o caso do indicador de PIB per capita, sendo a curva sólida os valores para Guiné-Bissau, enquanto a curva tracejada são os valores para a unidade de controle sintético.

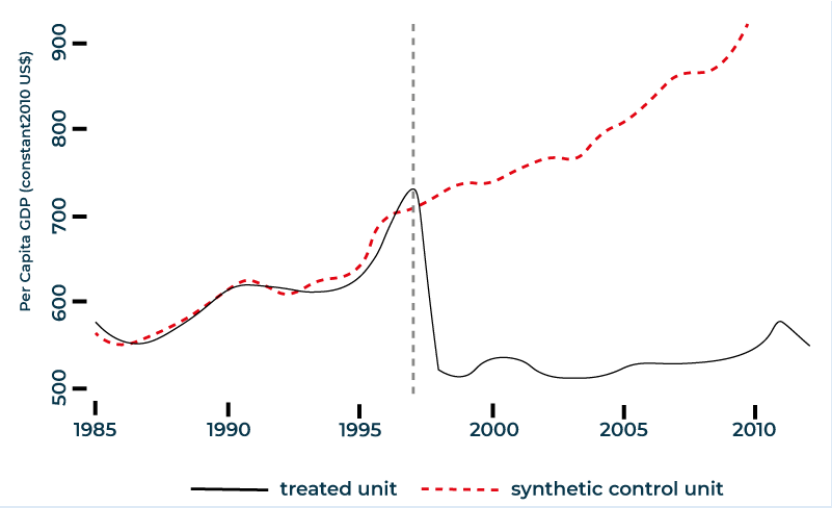


Figura 7.5. Efeito da entrada de Guiné-Bissau na zona do Franco CFA sobre o PIB per capita
Fonte: Figura 3A de Ndiaye (2020).

O eixo vertical indica o PIB per capita em dólares de 2010, enquanto o eixo horizontal indica os anos. A linha vertical tracejada demarca o ano de 1997, quando Guiné-Bissau entra na zona do Franco CFA. A curva sólida representa o valor do indicador para a unidade tratada, Guiné-Bissau, enquanto a linha tracejada é o valor do mesmo indicador para a unidade sintética, obtida a partir de 39 países.

Referência. NDIAYE, Mouhamadou Fallilou. Persistent Inequality In Guinea-Bissau: the role of France, the CFA Franc, and long-term currency imperialism. African Review of Economics and Finance, v. 12, n. 1, p. 123-151, 2020.

3.6 Validade externa dos métodos apresentados

Para finalizar esta parte do guia, propomos uma discussão sobre a validade externa de cada um dos métodos apresentados. Como discutido na Seção 2.3, a validade externa refere-se à medida em que os resultados encontrados em um caso específico podem ser generalizados para a população em geral. Um resumo dos principais pontos abordados nesta seção pode ser consultado na Tabela 3.2.

A randomização, ou ensaio controlado randomizado, é frequentemente considerada o "padrão-ouro" para a identificação de efeitos causais, uma vez que a atribuição aleatória ao tratamento reduz vieses de seleção. No entanto, os resultados obtidos em experimentos nem sempre são diretamente aplicáveis a outros contextos — especialmente quando a amostra não representa adequadamente a população-alvo ou quando as condições do estudo diferem daquelas enfrentadas na implementação real da política.

Em outras palavras, a capacidade de generalização dos achados (validade externa) depende tanto do desenho do estudo quanto da semelhança entre o contexto experimental e o contexto real onde se pretende aplicar os resultados. No desenho de descontinuidade de regressão (RDD), a identificação causal se baseia em uma mudança súbita no tratamento em torno de um ponto de corte específico. Embora a validade interna possa ser alta para as unidades próximas a esse limite, a validade externa tende a ser mais restrita. Isso ocorre porque os resultados são aplicáveis, principalmente, aos indivíduos localizados nas imediações do ponto de corte, o que dificulta sua extrapolação para indivíduos distantes desse limite ou para contextos distintos, nos quais o mesmo critério de elegibilidade não é utilizado.

Além disso, alterações no contexto ou na definição do ponto de corte podem limitar ainda mais a capacidade de generalização. No caso do modelo *fuzzy*, a validade externa é ainda mais restrita, uma vez que os efeitos estimados dizem respeito a uma população mais específica, composta apenas por indivíduos cuja probabilidade de tratamento foi afetada pela regra, mas que não foram necessariamente tratados conforme essa regra.

No caso do pareamento, busca-se criar grupos comparáveis ao combinar unidades tratadas e de controle com características observáveis semelhantes. Embora o método possa melhorar a validade interna ao reduzir vieses associados a variáveis observáveis, a validade externa pode ser limitada caso a amostra não seja representativa ou se variáveis relevantes não forem totalmente observadas. Em contextos nos quais a base de dados inclui indivíduos ou empresas muito específicos — por exemplo, de um único setor econômico ou região geográfica —, a generalização dos resultados para outros grupos pode ser comprometida. Assim, a validade externa dependerá da amplitude dos dados disponíveis e da proximidade entre as características da amostra analisada e da população-alvo.
No método de diferenças em diferenças (DiD), a identificação causal baseia-se na suposição de que, na ausência do tratamento, as tendências de evolução das unidades tratadas e não tratadas seriam paralelas. A validade externa, nesse caso, dependerá do quão representativas são as unidades tratadas e da robustez da suposição de tendências paralelas em outros contextos ou períodos. Por exemplo, uma política pública avaliada por meio de diferenças em diferenças em determinada região pode não gerar os mesmos efeitos em regiões com dinâmicas socioeconômicas distintas. Além disso, choques estruturais específicos a uma localidade podem dificultar a extrapolação dos resultados para outros contextos.

O método de Controle Sintético combina unidades de controle para criar uma "unidade sintética" que sirva de contrafactual próximo ao caso tratado. Embora possa oferecer estimativas robustas em termos de validade interna quando há boa similaridade entre a unidade tratada e as disponíveis para compor o contrafactual, a validade externa pode ser

limitada se não houver um conjunto de possíveis "doadores" (unidades de controle) que se assemelhe à realidade de outras populações ou contextos. Em outras palavras, se a unidade tratada for muito distinta das unidades disponíveis para a construção do controle sintético, a extrapolação para contextos diversos torna-se imprecisa.

Tabela 3.2 – Validade Externa dos Métodos apresentados

Método	Considerações sobre validade externa
Randomização	Alta validade interna, mas a generalização depende da representatividade da amostra e das condições do experimento
Desenho de Regressão Descontínua (RDD)	Focaliza-se nos indivíduos próximos ao ponto de corte; extrapolação limitada para outros perfis e contextos
Pareamento	Depende da disponibilidade de dados representativos e de variáveis relevantes; pode não se generalizar se a amostra for muito específica ou não abrange todas as variáveis importantes
Diferença em Diferenças (DiD)	Exige suposição de tendências paralelas; resultados podem não se estender a populações ou regiões cujas dinâmicas sejam muito diferentes
Controle Sintético (SC)	Requer conjunto amplo de unidades "doadoras" semelhantes; quanto mais distinta a unidade tratada, mais restrita a possibilidade de generalização

Fonte: Elaboração própria.

3.7 Cuidados ao comparar os impactos de políticas similares

Muitas vezes, além de compreender os efeitos gerados por uma política específica, é de interesse dos gestores comparar os resultados de sua avaliação de impacto com os de outras políticas similares. No entanto, para que essa análise comparativa seja informativa e confiável, alguns cuidados importantes devem ser observados:

- **Parâmetro estimado:** É fundamental verificar qual parâmetro está sendo estimado em cada avaliação. Não é adequado, por exemplo, comparar diretamente o efeito médio sobre os tratados (ATT) de uma política com o efeito médio local do tratamento (LATE) de outra. Além disso, é necessário certificar-se de que o parâmetro que acreditamos estar estimando é, de fato, aquele que está sendo calculado. Em algumas situações, pode-se presumir que o efeito estimado está sendo captado via ATE, por exemplo, mas na prática poucos dos selecionados para receber o tratamento estão participando da política. Essa situação ocorre quando não há um cumprimento muito assíduo do grupo tratado em relação à participação (como, por exemplo, uma política de capacitação profissional em que o tratamento exija algum deslocamento dos indivíduos para

participar de treinamentos) e, então, poderemos estar estimando, na prática, o efeito médio associado à intenção de tratar (ITT).

- **Período analisado:** Outro fator relevante é o horizonte temporal das políticas comparadas. Mesmo que duas políticas sejam similares em desenho e público-alvo, o contexto socioeconômico, político e institucional pode ter se alterado substancialmente ao longo do tempo. Assim, comparar uma política implementada recentemente com outra que ocorreu há 30 anos pode gerar conclusões equivocadas, uma vez que os fatores que afetam os desfechos de interesse podem ter mudado significativamente.
- **Unidade de análise:** É igualmente importante checar se as unidades de análise utilizadas nas avaliações são compatíveis. Comparar o impacto estimado a partir de dados individuais com aquele obtido com base em dados agregados de municípios ou estados pode ser enganoso, já que a interpretação dos resultados varia conforme o nível de agregação.
- **Aspectos operacionais e custo:** Ainda que os impactos estimados de duas políticas sejam semelhantes em magnitude e ambos se refiram a parâmetros comparáveis, vale considerar diferenças no custo de implementação e no planejamento envolvido em cada política — como desenho, monitoramento e adequação a princípios éticos. A depender do contexto, uma política que apresenta impacto similar à outra, mas com menor custo ou menor complexidade operacional, pode ser considerada mais vantajosa.

CAPÍTULO 4.

Conclusão

A avaliação de impacto é uma ferramenta essencial para o aprimoramento das políticas públicas, permitindo que gestores e formuladores tomem decisões informadas por evidências. Ao buscar mensurar os efeitos causais de uma intervenção, esse tipo de avaliação contribui para identificar o que realmente funciona, em quais contextos e para quais públicos.

Apesar dos desafios metodológicos envolvidos — como a construção de um contrafactual válido, o enfrentamento de vieses de seleção e a garantia de validade interna e externa —, os benefícios de avaliações bem conduzidas são significativos. Elas fortalecem a prestação de contas e a transparência na gestão pública, além de fornecerem subsídios valiosos para a replicação e o aprimoramento de políticas bem-sucedidas.

Este guia teve como propósito introduzir os principais conceitos e métodos relacionados à avaliação de impacto, oferecendo uma base para que gestores, técnicos e pesquisadores possam se familiarizar com os fundamentos dessa abordagem. Para aprofundamentos teóricos e aplicados, recomenda-se a consulta aos demais materiais produzidos pelo FGV CLEAR, que exploram com maior detalhamento os temas aqui apresentados. Alguns exemplos os guias: *Diferença-em-Diferenças*, *Modelagem de Microssimulação* e *Cálculo de Poder Estatístico*[22].

[22] Disponíveis em www.fgvclear.org/biblioteca/

Referências Bibliográficas

ABADIE, A.; CATTANEO, M. D. Econometric Methods For Program Evaluation. **Annual Review of Economics,** v. 10, p. 465-503, 2018.

ABADIE, A.; DIAMOND, A.; HAINMUELLER, J. Comparative Politics And The Synthetic Control Method. **American Journal of Political Science**, v. 59, n. 2, p. 495-510, 2015.

AKER, J. C., COLLIER, P.; VICENTE, P. C. **Is Information Power? Using mobile phones and free newspapers during an election in Mozambique**, 2017.

ANDERSON, S. Legal Origins And Female HIV. **American Economic Review**, 108(6), 1407-39, 2018.

ANDREWS, I.; OSTER, E. A Simple Approximation For Evaluating External Validity Bias. **Economics Letters,** 178, 58-62, 2019.

ANGRIST, J.D.; PISCHKE, J.S. **Mostly Harmless Econometrics: an empiricist's companion**. Princeton University Press, 2008.

ATHEY, S.; IMBENS, G. W. **The Econometrics Of Randomized Experiments**. *In*: Handbook Of Economic Field Experiments (Vol. 1, pp. 73-140). North-Holland, 2017a.

ATHEY, S.; IMBENS, G. W. The State Of Applied Econometrics: causality and policy evaluation. **Journal of Economic Perspectives**, v. 31, n. 2, p. 3-32, 2017b.

AVELLAR, A. P., & ALVES, P. F. Avaliação de impacto de programas de incentivos fiscais à inovação: um estudo sobre os efeitos do PDTI no Brasil. **Economia**, 9(1), 143-164, 2008.

BARROS, R., DE CARVALHO, M.; FRANCO, S.; ROSALÉM, A. Impacto Do Projeto Jovem de Futuro. **Estudos em Avaliação Educacional***,* 23(51), 214-226, 2012.

BATISTA, C., & VICENTE, P. C. **Improving Access To Savings Through Mobile Money: experimental evidence from smallholder farmers in Mozambique** (No. novaf: wp1705). Universidade Nova de Lisboa, Faculdade de Economia, NOVAFRICA, 2017.

BEIN, P.; MILLER, T.; WATERS, W. British Columbia Road-user Unit Costs. **Proceedings of the Canadian Transportation Research Forum Annual Meeting**, University of Saskatchewan, 714–727, 1994.

BERTANHA, M.; IMBENS, G. W. External Validity In Fuzzy Regression Discontinuity Designs. **Journal of Business & Economic Statistics,** 38(3), 593-612, 2020.

BERTRAND, M.; DUFLO, E.; MULLAINATHAN, S. How Much Should We Trust Differences-In-Differences Estimates?. **The Quarterly Journal of Economics**, 119(1), 249-275, 2004.

BETTER EVALUATION. **IMPACT EVALUATION.** Disponível em: https://www.betterevaluation.org/themes/impact_evaluation. Acesso em: 10 set. 2020.

BHALOTRA, Sonia R.; ROCHA, R.; SOARES, R. R. **Does Universalization of Health Work? Evidence from health systems restructuring and expansion in Brazil,** 2019.

BLUNDELL, R.; DIAS, M. C. Alternative Approaches To Evaluation In Empirical Microeconomics. **Journal of Human Resources,** v. 44, n. 3, p. 565-640, 2009.

BRASIL. Casa Civil da Presidência da República *et al.* Casa Civil. **Avaliação De Políticas Públicas: guia prático de análise *ex post***, v. 2. Brasília, 2018.

BRUNIE, A. *et al.* Can Village Savings And Loan Groups Be A Potential Tool In The Malnutrition Fight? Mixed method findings from Mozambique. **Children and Youth Services Review**, 47, 113-120, 2014.

BUSSAB, W. O.; MORETTIN, P. **Estatística Básica.** Saraiva Educação SA, 2017.

CALIENDO, M.; KOPEINIG, S. Some Practical Guidance For The Implementation Of Propensity Score Matching. **Journal of Economic Surveys**, v. 22, n. 1, p. 31-72, 2008.

CARTER, M.; LAAJAJ, R.; YANG, D. Subsidies And The African Green Revolution: direct effects and social network spillovers of randomized input subsidies in Mozambique. **National Bureau of Economic Research**, 2019.

CASINI, P.; RIERA, O.; SANTOS MONTEIRO, P. **Labor Market Effects Of Improved Access To Credit Among The Poor: evidence from Cape Verde.** Available at SSRN 2562844, 2014.

CASTRO, M. F. **Defining And Using Performance Indicators and Targets in Government M&E Systems.** *In*: Lopez-Acevedo, Gladys, et al., eds. Building Better Policies: the nuts and bolts of monitoring and evaluation systems. Washington: World Bank, 2012.

CERQUEIRA, D. *et al*. **Avaliando A Efetividade Da Lei Maria da Penha**, 2015. Disponível em: http://repositorio.ipea.gov.br/handle/11058/3538. Acesso em: 06 set 2020.

CERQUEIRA, D. R. C. *et al.,* 2020. Uma Avaliação De Impacto De Política De Segurança Pública: o Programa Estado Presente do Espírito Santo. **Texto para Discussão** (TD) 2543. Disponível em: http://repositorio.ipea.gov.br/handle/11058/9704. Acesso em: 06 set 2020.

CHEN, G.; WARBURTON, R. N. Do Speed Cameras Produce Net Benefits? Evidence from British Columbia, Canada. Journal of Policy Analysis and Management: **The Journal of the Association for Public Policy Analysis and Management,** v. 25, n. 3, p. 661-678, 2006.

COSTA, F. *et al.* Homicides And The Age of Criminal Responsibility in Brazil: a density discontinuity approach. Economia, **the journal of LACEA**, 2018.

CUNGUARA, B.; DARNHOFER, I. Assessing The Impact Of Improved Agricultural Technologies On Household Income In Rural Mozambique. **Food Policy**, v. 36, n. 3, p. 378-390, 2011.

CUSTÓDIO, C.; MENDES, D.; METZGER, D. **The Impact Of Financial Literacy On Medium And Large Enterprises – evidence from a randomized controlled trial in Mozambique**, 2019.

DHALIWAL, I., DUFLO, E., GLENNERSTER, R.; TULLOCH, C. Comparative Cost Effectiveness Analysis To Inform Policy In Developing Countries: a general framework with applications for education. **Education Policy in Developing Countries**, pp.285-338, 2013.

DI MARO, V.; LEEFFERS, S.; SERRA, D.; VICENTE, P. C. **Mobilizing Parents at Home and at School: an experiment on primary education in Angola** (No. wp2002). Universidade Nova de Lisboa, Faculdade de Economia, NOVAFRICA, 2020.

DJIMEU, E. W. The Impact Of Social Action Funds On Child Health In A Conflict Affected Country: evidence from Angola. **Social Science & Medicine**, 106, 35-42, 2014.

DOUDCHENKO, N.; IMBENS, G. W. Balancing, Regression, Difference-In-Differences And Synthetic Control Methods: a synthesis. **National Bureau of Economic Research**, 2016.

EBEKE, C.; MANSOUR, M.; ROTA-GRAZIOSI, G.. **The Power to Tax In Sub-Saharan Africa: LTUs, VATs, and SARAs**, 2016.

FAZZIO, I. *et al*. Large Learning Gains in Pockets of Extreme Poverty: experimental evidence from Guinea Bissau. **National Bureau of Economic Research Working Paper** No. 27799, 2020.

FERMAN, B.; PINTO, C.; POSSEBOM, V. Cherry Picking With Synthetic Controls. **Journal of Policy Analysis and Management**, v. 39, n. 2, p. 510-532, 2020.

GREßER, C.; STADELMANN, D. Evaluating Water And Health Related Development Projects: a cross-project and micro-based approach (No. 2019-06). **CREMA Working Paper**, 2019.

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica.** Amgh Editora, 2011.

HAHN, J.; TODD, P.; VAN DER KLAAUW, W. Identification And Estimation Of Treatment Effects With A Regression-Discontinuity Design. **Econometrica,** 69(1), pp.201-209, 2001.

HEDGE, R.; BULL, G.Q. Performance Of An Agro-Forestry Based Payments-for-Environmental-Services Project In Mozambique: a household level analysis. **Ecological Economics**, 71, pp.122-130, 2011.

HOSONO, T.; AOYAGI, K. Effectiveness Of Interventions To Induce Waste Segregation By Households: evidence from a randomized controlled trial in Mozambique. **The Journal of Material Cycles and Waste Management,** 20(2), 1143-1153, 2018.

INSTITUTO JONES DOS SANTOS NEVES. **Guia Para Avaliar Políticas Públicas** - Vol 4: e quando a política está em andamento? Avaliação ex post. Vitória, ES, 2018.

JAKUBOWSKI, A. *et al.* The US President's Malaria Initiative and under-5 child mortality in sub-Saharan Africa: A difference-in-differences analysis. **PLoS medicine**, 14(6), e1002319, 2017.

JAUPART, P.; DIPPLE, L.; DERCON, S. Has Gavi Lived Up To Its Promise? Quasi-experimental evidence on country immunization rates and child mortality. **BMJ global health**, 4(6), 2019.

JINDAL, R.; KERR, J. M.; CARTER, S. **Reducing Poverty Through Carbon Forestry? Impacts of the N'hambita community carbon project in Mozambique**, 2012.

Jovem de Futuro – Instituto Unibanco. Disponível em: https://www.institutounibanco.org.br/iniciativas/jovem-de-futuro/. Acesso em: 14 set. 2020.

KASSA, W.; COULIBALY, S. Revisiting The Trade Impact Of The African Growth And Opportunity Act: a synthetic control approach. Econometric Modeling: **International Economics Journal**, 2019.

KAUL, A. *et al*. **Synthetic Control Methods: never use all pre-intervention outcomes as economic predictors.** University of Hohenheim, Department of Economics, 2016.

KEUZENKAMP, H. A.; MAGNUS, J. R. On Tests And Significance In Econometrics. **Journal of Econometrics**, v. 67, n. 1, p. 5-24, 1995.

KIRIGIA, J. M. *et al.* A Performance Assessment Method For Hospitals: the case of municipal hospitals in Angola. **Journal of medical systems**, v. 32, n. 6, p. 509-519, 2008.

KITAGAWA, T. A Test For Instrument Validity. **Econometrica**, v. 83, n. 5, p. 2043-2063, 2015.

LEE, David S.; LEMIEUX, T. Regression Discontinuity Designs In Economics. **Journal of Economic Literature**, v. 48, n. 2, p. 281-355, 2010.

MANSKI, C. F. **Public Policy In An Uncertain World: analysis and decisions.** Harvard University Press, 2013.

MASTONSHOEVA, S.; MYRTTINEN, H.; CHIRWA, E.; SHONASIMOVA, S.; GULYAMOVA, P.; Shai, N. & JEWKES, R. Evaluation of Zindagii Shoista (Living with Dignity), an intervention to prevent violence against women in Tajikistan: impact after 30 months**. What Works to Prevent Violence against Women and Girls,** 2020. Disponível em: https://www.whatworks.co.za/resources/reports/item/687-evaluation-of-zindagii-shoista-living-with-dignity-an-intervention-to-prevent-violence-against-women-in-tajikistan-impact-after-30-months. Acesso em: 09 ago. 2020.

MCCRARY, J. Manipulation Of The Running Variable In The Regression Discontinuity Design: a density test. **Journal of Econometrics**, v. 142, n. 2, p. 698-714, 2008.

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

MEYER, P. L**. Probabilidade: aplicações à estatística**. 2ed. Livros Técnicos e Científicos Editora SA, 1983.

MOGUES, T.; MUELLER, V.; KONDYLIS, F.. **Cost-effectiveness Of Community-based Gendered Advisory Services To Farmers: analysis in Mozambique and Tanzania,** v. 14, n. 3, 2019.

MONTALVO, J. G. Voting After The Bombings: a natural experiment on the effect of terrorist attacks on democratic elections**. Review of Economics and Statistics**, v. 93, n. 4, p. 1146-1154, 2011.

NDIAYE, M. F. Persistent Inequality in Guinea-Bissau: The role of France, the CFA Franc, and long-term currency imperialism. **African Review of Economics and Finance**, v. 12, n. 1, p. 123-151, 2020.

NEWCOMER, K.E.; HATRY, H.P.; WHOLEY, J.S. **Handbook Of Practical Program Evaluation**. John Wiley & Sons, 2015.

NKALA, P.; MANGO, N.; ZIKHALI, P. Conservation Agriculture And Livelihoods Of Smallholder Farmers In Central Mozambique. **Journal of Sustainable Agriculture**, 35(7), 757-779, 2011.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO - OCDE. **Programme For International Student Assessment (PISA)** - FAQ: http://www.oecd.org/pisa/pisafaq/. Acesso em: 14 set. 2020.

PEIXOTO, B. T. *et al.* **Prevenção e controle de homicídios: uma avaliação de impacto no Brasil.** Belo Horizonte: UFMG, Cedeplar, 2008.

PEREDA, P.C.; ALVES, D.C. **Econometria Aplicada.** Elsevier Brasil, 2018.

PEREIRA FILHO, O. A.; SOUSA, M. D. C. S. D.; ALVES, P. F. Avaliação De Impacto Das Guardas Municipais Sobre A Criminalidade Com O Uso De Tratamentos Binários, Multivalorados E Contínuos. **Revista Brasileira de Economia**, 72(4), 515-544, 2018.

POMPEO, J. N. U.; HAZZAN, S. **Matemática Financeira.** Saraiva Educação SA, 2017.

Portal do INEP – SAEB: **Sistema de Avaliação da Educação Básica do Brasil.** Disponível em: http://portal.inep.gov.br/educacao-basica/saeb. Acesso em: 14 set. 2020.

POSSE, M. E. et al. The Effect Of Food Assistance On Adherence To Antiretroviral Therapy Among HIV/AIDS Patients In Sofala Province, In Mozambique: a retrospective study. **J AIDS Clin Res 4**: 198, 2013. doi:10.4172/2155-6113.1000198

RESENDE, A. C. C.; OLIVEIRA, A. M. H. C. D. Avaliando Resultados De Um Programa De Transferência De Renda: o impacto do Bolsa-Escola sobre os gastos das famílias brasileiras. **Estudos Econômicos** (Sao Paulo), 38(2), 235-265, 2008.

ROMERO, J. A. R.; HERMETO, A. M. Avaliação De Impacto Do Programa Bolsa Família Sobre Indicadores Educacionais: uma abordagem de regressão descontínua. **Encontro Nacional de Economia**, 37, 2009.

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. **Experimental And Quasi-experimental Designs For Generalized Causal Inference.** Boston: Houghton Mifflin, 2002.

SOARES, S. S. D. **O Conhecimento Paga Bem? Habilidades cognitivas e rendimentos do trabalho no Brasil (e no Chile).** 2011. 146 f., il. Tese (Doutorado em Ciências Econômicas) - Universidade de Brasília, Brasília, 2011.

UNICEF, World Bank. **Abolishing School Fees in Africa: lessons from Ethiopia, Ghana, Kenya, Malawi, and Mozambique.** Washington, DC: The World Bank, 2009.

USD/BRL - **Dólar Americano Real Brasileiro.** Disponível em: https://br.investing.com/currencies/usd-brl-converter. Acesso em: 14 set. 2020.

USD/CAD - **Dólar Americano Dólar Canadense**. Disponível em: https://br.investing.com/currencies/usd-cad-converter. Acesso em: 14 set. 2020.

USD/MZN - **Dólar Americano Metical de Moçambique.** Disponível em: https://br.investing.com/currencies/usd-mzn-converter. Acesso em: 14 set. 2020.

VICENTE, P. C. Does Oil Corrupt? Evidence from a natural experiment in West Africa. **Journal of development Economics,** 92(1), 28-38, 2010.

VICENTE, P. C. Is Vote Buying Effective? Evidence from a field experiment in West Africa. **The Economic Journal,** 124(574), F356-F387, 2014.

WOOLDRIDGE, J. M. **Introdução à Econometria: uma abordagem moderna**. São Paulo: Cengage Learning, 2010a.

WOOLDRIDGE, J. M. **Econometric Analysis Of Cross Section And Panel Data**. MIT press, 2010b.

APÊNDICE TÉCNICO – HIPÓTESES DOS MÉTODOS

Esse apêndice serve como apoio e complemento para os métodos *ex post* de avaliação de impacto de políticas, apresentados ao longo deste guia. Aqui, estão descritas as hipóteses assumidas em cada método exposto, além de outros tópicos a eles relacionados.

Box A.1. SUTVA

Para utilizar qualquer um dos métodos de avaliação de impacto apresentados neste guia, é necessário que a hipótese de SUTVA (do inglês, Stable Unit Treatment Value Assumption) seja válida. Em inferência causal, essa suposição envolve dois pontos principais:

1) **Ausência de interferência entre unidades**: O tratamento recebido por uma unidade não afeta o resultado de outra unidade; e

2) **Ausência de variações ocultas do tratamento**: Cada nível de tratamento é aplicado de forma consistente e tem o mesmo significado para todas as unidades.

Juntas, essas condições garantem que os resultados potenciais de uma unidade dependam apenas do seu próprio tratamento, sem influências externas ou variações não observadas.

Box A.2. Método Experimental

Hipótese 1 – Aleatorização

A seleção dos participantes é feita de maneira aleatória e não depende de nenhuma característica das observações. Essa hipótese implica em dizer que não há nenhuma variável relacionada às observações que seja capaz de alterar a probabilidade de receber o tratamento. Portanto, todas as unidades selecionadas para participarem do sorteio devem ter chances iguais de receber a política.

Box A.3. Regressão Descontínua

Hipótese 1 – RD

Precisamos que haja **continuidade** entre a variável (Z) que afeta o tratamento e quaisquer outras variáveis que determinam o indicador de impacto. Isso é importante, pois se ocorrer outro tipo de descontinuidade gerada por conta de algum valor de referência na variável Z ou em outra variável que não seja o indicador de impacto e que o afete, então, os efeitos dessa descontinuidade e da descontinuidade gerada na probabilidade de recebimento do tratamento se misturariam, inviabilizando o uso de RD.

Por exemplo, suponha que uma regra para participar em uma política que determina que serão selecionados para o grupo tratado os indivíduos que possuírem renda familiar abaixo de determinado valor de referência. Sendo assim, a hipótese de continuidade diz que, na ausência da política, a variável de interesse é uma função contínua da renda. Se o mesmo valor de referência (ou algum bem próximo) é utilizado para definir alguma outra política, essa hipótese de continuidade dificilmente se sustentará.

Uma maneira simples e já conhecida para checar a validade dessa primeira hipótese é a partir de uma análise do balanceamento da amostra, conforme sugerido, também, no contexto de aleatorização. Como se espera que nenhuma outra variável cause algum tipo de descontinuidade na probabilidade de recebimento da política, não devem existir muitas diferenças entre a composição dos grupos, tratado e controle, quando analisadas outras variáveis que não a(s) de interesse. Outra maneira de checar isso seria mudando arbitrariamente o valor de referência, $Z = Z^*$, que determina a seleção para a política, para um valor qualquer $Z' \neq Z^*$. Esse é um procedimento comumente chamado de **teste de placebo** dentro da literatura de avaliação de impacto. Como a probabilidade de receber a política só deve ser afetada quando $Z \geq Z^*$, então, espera-se um efeito nulo do impacto da política para valores $Z' \neq Z^*$.

Hipótese 2 – RD

É necessário que exista **controle impreciso** sobre a variável (Z) que afeta a seleção para participação. Isso significa dizermos que as observações não podem controlar perfeitamente o valor que terão com relação a essa variável, pois caso o pudessem, poderiam escolher um valor um pouco acima (ou abaixo) do ponto de referência (Z^*) para garantir a seleção na participação da política, por exemplo.

Um exemplo de variável clássica em que há controle impreciso sobre é a idade de pessoas adultas, já que não existe possibilidade de manipularmos a data de nascimento delas. Portanto, regras decisórias que partam da idade das pessoas são seguras quanto a aplicação dessa hipótese.

Um exemplo onde pode existir um tipo de manipulação das observações sobre a variável de seleção é a quantidade de alunos em escolas. Suponha que o governo irá determinar a participação de escolas em uma política de subsídios a partir da quantidade de alunos matriculados, sendo as participantes as que possuírem, no máximo, uma certa quantidade de referência. Sendo assim, a ideia da estimativa de impacto via RD seria comparar escolas com a quantidade de alunos logo acima com as que possuem a quantidade de alunos logo abaixo do valor de referência. Entretanto, se as escolas souberem disso, podem escolher matricular alunos de forma a não atingir o valor de referência, permitindo a qualificação para serem selecionadas a participar da política, indicando certo controle na variável de número de matrículas.

Uma maneira muito comum, na literatura de avaliação de impacto, que utiliza o método de RD para testar a validade dessa hipótese é o Teste McCrary, introduzido originalmente por McCrary (2008). A ideia é bem simples: checar a distribuição de frequências da variável de seleção, Z, e investigar se próximo do ponto $Z = Z^*$ há alguma concentração de valores maior do que o normal.

Hipótese 3 – RD (apenas para o caso Fuzzy)

Caso estejamos em um arcabouço *Fuzzy* do RD, é preciso que haja **monotonicidade**, ou seja, estar acima (ou abaixo) do valor de referência (Z^*) não faça com que a probabilidade de receber o tratamento $\Pr[D = 1]$ se altere de maneira contrária em relação ao esperado. Se imaginamos que estar acima do valor de referência ($Z \geq Z^*$) faz com que a probabilidade de ser selecionado aumente, estamos assumindo que para qualquer observação que tenha $Z \geq Z^*$ e não participa (isto é, $D = 0$), ela já faria isso de qualquer forma. A hipótese garante que não existe um caso, nesse exemplo, de uma observação em que $Z \geq Z^*$ e, por conta disso, $D = 0$.

Imagine o exemplo de uma política de aposentadoria que seleciona os participantes a partir de algum valor de idade (pelo menos, 65 anos). Sendo assim, pessoas com 65 ou mais anos de idade poderão receber o benefício da política e, caso tenham menos de 65 anos, não poderão. Assim, poderá ocorrer de alguns indivíduos com mais de 65 anos não receberem e, algumas

pessoas com menos de 65 receberem (trata-se de um caso RD *Fuzzy*). A hipótese garante que, os casos de pessoas com, pelo menos, 65 anos que não recebem a política são casos em que esses indivíduos já iriam escolher não participar da política, independentemente de ter ou não a possibilidade de participar.

O trabalho de Kitagawa (2015) formaliza essa hipótese mais detalhadamente e propõe uma maneira de testá-la.

Box A.4. Pareamento

Hipótese 1 – Pareamento

A primeira hipótese do método de pareamento é a de **seleção em observáveis**, na qual após condicionarmos o modelo às características observáveis (X) podemos dizer que é como se a seleção para participação tivesse sido aleatória. Portanto, a ideia da hipótese é de que se temos em mãos algumas características específicas sobre a observação, não há mais nada que esteja fora do que é observado que determine o tratamento.

Um exemplo dessa hipótese seria uma política de auxílio de renda para famílias em situação de extrema pobreza. Supondo que as famílias que podem receber esse auxílio devem comprovar algumas informações relacionadas a número de filhos, renda, região de moradia e idade dos filhos, então tais variáveis precisam estar disponíveis na estimativa do impacto da política, já que influenciam na possibilidade de tratamento. Portanto, a ideia seria que se tivéssemos disponíveis tais variáveis (nº de filhos, renda, região de moradia e idade dos filhos), é como se, condicional a elas, tivesse havido um sorteio entre quais famílias vão receber o auxílio.

Box A.5. Pareamento via Escore de Propensão

Hipótese 1 – Pareamento via Escore de Propensão

Após **condicionarmos o modelo à probabilidade** de receber o tratamento, $p(X)$, podemos dizer que é como se a seleção para participação tivesse sido aleatória. Portanto, a ideia da hipótese é a mesma da Hipótese 1 do Pareamento: se temos em mãos algumas características específicas sobre a observação, não há mais nada que esteja fora do que é observado que determine o tratamento.

Vamos utilizar o mesmo exemplo da hipótese análoga no caso do pareamento, onde temos uma política de auxílio de renda para famílias em situação de extrema pobreza. Como as famílias devem comprovar algumas informações relacionadas a número de filhos, renda, região de moradia e idade dos filhos, para estarem aptas a receber o auxílio, o procedimento é estimar

a probabilidade de seleção para a política, $p(X)$ com tais variáveis. Feito isso, poderemos ver o tratamento como se fosse aleatório após condicionarmos a $p(X)$.

Hipótese 2 – Pareamento via Escore de Propensão

Precisamos que exista **suporte comum** entre os grupos de tratamento e controle no que diz respeito à probabilidade de receber o tratamento estimado via escore de propensão. Isso significa dizer que nenhuma das variáveis contidas em X poderá ser capaz de determinar perfeitamente se as observações serão ou não selecionadas para receber o tratamento, pois caso isso fosse verdade, não seria possível construir um grupo de controle parecido com o grupo tratado, pois $p(X)$ seria sempre igual a zero ou um. Em termos matemáticos, precisamos que $0 < p(X) < 1$.

Box A.6. Diferença-em-diferenças

Hipótese 1 – DD

A principal hipótese do método é a de **tendências paralelas** entre os dois grupos, isto é, que a trajetória do(s) indicador(es) do grupo de controle represente bem o comportamento contrafactual do grupo tratado. Portanto, estamos dizendo que se não ocorresse a política, o comportamento do indicador do grupo que participa da política seria similar ao que, de fato, é observado do grupo de comparação.

Um exemplo para visualizarmos quando não há o cumprimento dessa hipótese seria o caso em que existe uma política de combate à violência em determinadas cidades de uma região do país, enquanto outras cidades de outras regiões do país não recebem a política. Suponha que algum indicador de violência possa ser observado para os dois grupos de municípios em diversos momentos anteriores à política vigorar. Inspecionando os dados, vemos que, para o grupo tratado, o indicador está cada vez maior (crescente), enquanto, para o grupo controle, o indicador está cada vez menor (decrescente) conforme o período fica mais próximo da implementação da política. Dessa maneira, não poderemos utilizar tais cidades na composição do grupo de controle, já que a tendência do indicador antes da política era muito diferente.

Note que não é possível verificar a validade dessa hipótese, já que ela trata de um cenário contrafactual e que, por isso, não irá acontecer na prática. Entretanto, é possível observar algum indicativo da plausibilidade dessa hipótese a partir de inspeções visuais do comportamento da(s) variável(is) de interesse no(s) momento(s) anterior(es) à política para cada um dos grupos de análise. Caso tenham comportamentos parecidos, existe uma evidência favorável à hipótese de tendências paralelas. Note que não é preciso que o(s) valor(es) do(s) indicador(es) seja(m) parecido(s) entre os grupos, mas apenas a tendência do(s) mesmo(s). Se a amostra de dados dispõe de informações sobre o indicador em diversos momentos pré-

implementação da política, teremos como investigar com maior facilidade a plausibilidade de tal hipótese.

A **Figura 1.** representa dois casos distintos em relação à evidência que podemos ter sobre a hipótese de tendências paralelas. As curvas pretas representam a trajetória do indicador para o grupo tratado pela política, enquanto a curva cinza representa a trajetória do mesmo indicador para o grupo de comparação. A linha tracejada indica o cenário contrafactual que está sendo estimado, em cada um dos casos, a partir do comportamento do grupo de controle após a implementação da política. Os períodos "-2", "-1" e "0" indicam momentos no tempo anteriores à política ser implementada, enquanto o período "1" indica um momento após ela ter sido iniciada.

No "Caso 1", o comportamento dos indicadores parece ser similar entre os dois grupos até o momento da política ser implementada, enquanto o "Caso 2" parece tratar de grupos com comportamentos muito distintos quanto à trajetória do indicador em períodos pré-intervenção. Veja que, ao prosseguirmos com a estimativa do impacto em um cenário como o "Caso 2", teremos um viés nos resultados. Ao pressupormos que a trajetória do grupo de controle é uma boa representação do cenário contrafactual para os tratados, estamos considerando que a evolução do indicador para o grupo tratado seria muito menor do que ela aparenta que seria na ausência da política. Basta notarmos que, para o grupo tratado, a tendência já era de uma crescente no indicador, mesmo antes da política ser implementada (períodos "-2", "-1" e "0"), enquanto o cenário do grupo de controle era de aumento, mas com uma tendência de estagnação.

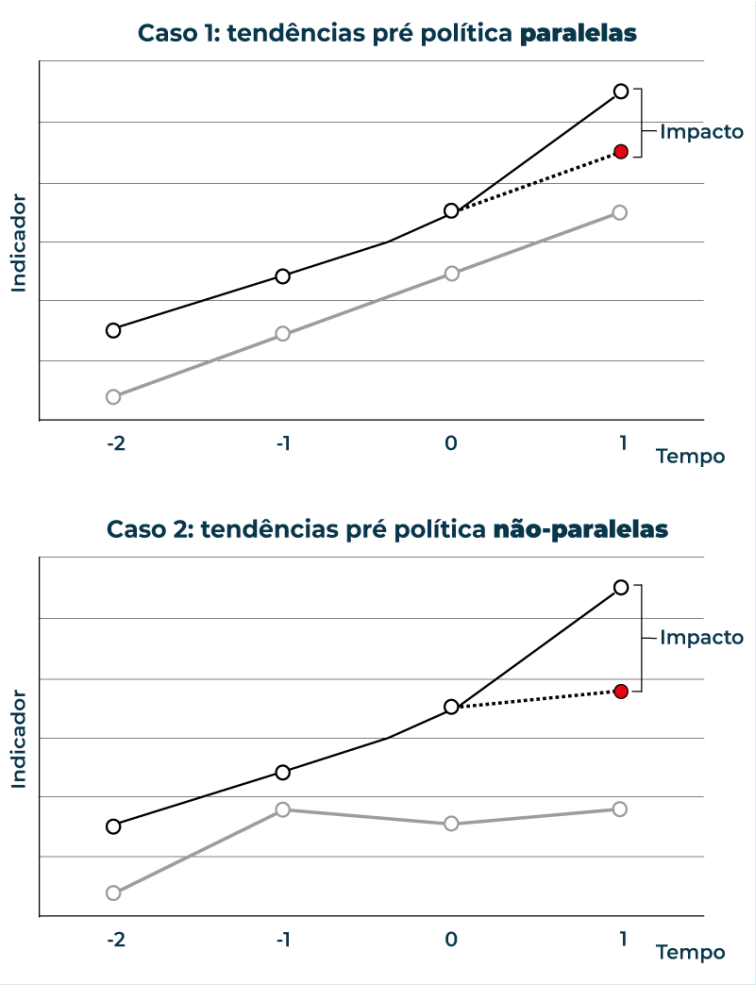


Figura 1. Tendências paralelas e não-paralelas

Fonte: elaboração própria.

Nota: Os períodos de tempo -2, -1 e 0 representam momentos prévios à implementação da política, enquanto o período 1 representa um momento após ela ser implementada. A trajetória da curva preta ilustra o indicador para o grupo tratado, enquanto a trajetória em cinza representa o comportamento do indicador para o grupo de controle. A trajetória tracejada é a representação do cenário contrafactual do grupo tratado.

Hipótese 2 – DD

Não podemos ter nenhuma mudança muito significativa com relação à **composição de cada um dos grupos**. Isso significa dizer que não podemos ter observações que entram ou saem da composição dos grupos ao longo do horizonte temporal analisado, principalmente entre os períodos temporais anteriores e posteriores à política ser implementada. Uma grande mudança de composição dos grupos seria um cenário em que as unidades tratadas pela política (ou as de controle) fossem se alterando com o passar do tempo. Se muitas mudanças ocorrerem entre as observações que fazem parte do grupo tratado e do grupo controle, as diferenças calculadas pelo método de DD podem ser advindas dessas mudanças de composição dos grupos e não pela política em si.

Por exemplo, suponha uma política que vise melhorar indicadores educacionais em uma cidade e que nossa análise de DD utilize o desempenho médio dos alunos de uma mesma coorte nas escolas tratadas e nas de controle entre dois anos, um anterior à política ser implementada e outro após. Caso exista algum movimento grande de saída de alunos na coorte dentre as escolas do grupo tratado (ou de controle), pode ser que os alunos que estão sendo avaliados, no momento posterior à política iniciar, não sejam todos que, de fato, foram afetados pela mesma e/ou tenham características diferentes daquelas observadas para o grupo em sua formação original, tornando as análises viesadas.

Hipótese 3 – DD

Os dois grupos de análise da política **não podem ser afetados de maneiras diferentes** após a implementação da política. Isso significa dizer que não existe mais nada ocorrendo após a implementação da política que possa afetar o(s) indicador(es) dos dois grupos de maneiras distintas. Se fosse o caso, poderíamos acabar estimando um impacto que ocorre por meio de mudanças em outros aspectos que não tenham a ver com a política.

Um exemplo para o caso em que se viola essa hipótese, seria quando uma outra política pública, diferente daquela em análise, entra em vigor após a implementação da política em questão e afeta o grupo de comparação. Sendo assim, teríamos um dos grupos tendo seu indicador de impacto sendo afetado por essa segunda política, viesando os resultados estimados.

Essa última hipótese, apesar de não poder ser testada na prática, tem reduzido o risco de ser violada se o momento de tempo posterior à ocorrência da política pública em que os grupos são observados for relativamente próximo do momento em que ela foi implementada. Dessa maneira, evita-se que outros aspectos relevantes que afetam diferentemente os grupos de tratamento e comparação possam mudar no período entre o início da política e a observação dos indicadores de impacto no pós-intervenção.

Uma possibilidade de implementação do desenho de DD é combinando esse método com o de pareamento. A ideia é que as observações que irão compor o grupo de controle serão escolhidas a partir da implementação de alguma das metodologias de pareamento. Com essa combinação de métodos, temos o benefício de uma flexibilização das hipóteses necessárias para a estimativa.

Box A.7. Combinação do método de Pareamento com diferença-em-diferenças

Para construir um grupo de controle na aplicação do método de diferença-em-diferenças, poderemos recorrer ao arcabouço do pareamento. Se tivermos disponibilidade de dados sobre observações que não foram tratadas em momentos anteriores e posteriores à política,

podemos definir as que irão compor o grupo de controle a partir do pareamento em variáveis observáveis. Dessa maneira, conseguimos selecionar um conjunto de observações que não foram afetadas pela política, mas que eram parecidas com as que foram afetadas, até o momento da implementação da política.

São necessárias duas hipóteses para a implementação desse método: (i) caso não ocorra tratamento (contrafactual), a evolução do indicador de impacto (antes e depois da implementação da política), condicionado à características observáveis, deve ser igual entre controles e tratados; e (ii) suporte comum (assim como na segunda hipótese de PSM) para antes e depois da implementação da política. Em resumo, a primeira hipótese diz que se não ocorrer a política, os indicadores de ambos os grupos (tratado e comparação) teriam comportamento semelhante (bem como na hipótese de tendências paralelas de DD), ou seja, as diferenças no comportamento dos indicadores podem ser interpretadas como efeitos da política. A segunda hipótese diz que toda unidade tratada possui ao menos uma observação para ser utilizada como controle no grupo não afetado, antes e depois da política ser implementada.

Maiores detalhamentos a respeito da combinação dos métodos de pareamento e diferença-em-diferenças podem ser encontrados em Blundell e Dias (2009). Um exemplo de aplicação dessa combinação de métodos é a avaliação de impacto de uma política de prevenção ao suicídio no Brasil, feita em Peixoto, Andrade e Azevedo (2008). Em Posse et al. (2013), os autores avaliam o impacto de uma política em Moçambique para incentivo à aderência aos tratamentos para combate ao HIV. Também para Moçambique, Brunie et al. (2014) estimam o impacto de uma política para melhoria de indicadores nutricionais.

Em muitos casos de implementação de políticas públicas, estaremos interessados em observar sua evolução ao longo do tempo e em diversos momentos após sua implementação ocorrer. Isso pode ser de interesse para saber se a política promoveu uma evolução em algum indicador de impacto, ou teve impacto no curto prazo, mas posteriormente não foi mais efetiva, por exemplo. Para conseguir explorar esse tipo de análise, podemos utilizar a abordagem de diferença-em-diferenças, que permite uma análise de heterogeneidade do impacto da política ao longo do tempo.

Box A.8. Metodologia de *Event Study*

Quando possuirmos, na amostra de dados, mais de um período posterior ao da implementação da política, é possível alterar um pouco a estrutura da Equação 2 a fim de analisar os (possíveis) efeitos em diferentes períodos. A ideia é que seja incluída, para cada período disponível – inclusive os anteriores à implementação – um termo como o $\beta(D \cdot T)$ da equação em questão. Dessa maneira, avaliaremos um possível impacto em cada momento de tempo analisado e não

apenas realizando uma comparação "antes e depois" da implementação da política. Além disso, ainda que não seja um resultado conclusivo, os coeficientes associados a períodos pré-implementação da política fornecem indicativos sobre a plausibilidade da hipótese de tendências paralelas, pois comparam o indicador entre os dois grupos – tratado e controle – em momentos do tempo nos quais a política ainda não havia sido implementada. Sendo assim, espera-se que tais coeficientes, nesses períodos, tenham magnitude sempre próxima de zero. A aplicação dessa metodologia pode ser muito útil para o acompanhamento da política e seus efeitos, ajudando no monitoramento da mesma e indicando possíveis caminhos a serem tomados, a depender dos resultados.

Um exemplo de utilização dessa metodologia pode ser encontrada em Bhalotra, Rocha e Soares (2019), que estimam o impacto do Programa Saúde da Família, do Brasil.

Como dito anteriormente, a estrutura de dados que precisamos ao trabalharmos com a metodologia de diferença-em-diferenças requer que as observações afetadas e não afetadas pela política sejam observadas em ao menos dois momentos do tempo: antes e depois da implementação da política. Na prática, isso significa que possuímos **dados em painel**, que nada mais são do que diferentes recortes temporais sobre características de um mesmo conjunto de observações[23]. Por exemplo, se possuímos uma amostra de dados anuais sobre variáveis econômicas de países subdesenvolvidos, para o período de 1990 a 2020, temos uma estrutura de dados em painel, onde as unidades de análise são os países e as observações referem-se a pares de país e ano. Quando nossos dados estão dispostos no formato de painel, é possível implementar diferentes metodologias para lidar com problemas de viés, conforme será discutido a seguir. Deve-se deixar claro que a estrutura de dados em painel requer que as mesmas unidades (ou, pelo menos, a grande maioria delas) possam ser observadas nos diferentes períodos de tempo.

Box A.9. Metodologia de Dados em Painel

Quando um mesmo conjunto de unidades puder ser analisado em diferentes recortes temporais, estaremos diante de uma estrutura de dados em painel. Conforme discutido ao longo da metodologia de diferença-em-diferenças, a grande vantagem dessa é que podemos eliminar características não observadas que sejam constantes ao longo do tempo. Sendo assim, qualquer viés que tais variáveis poderiam trazer é eliminado. Como o desenho de DD é um caso particular de dados em painel, as outras metodologias mencionadas a seguir também irão eliminar problemas de viés que as características não-observáveis e constantes no tempo iriam trazer. Os métodos mais comuns quando se lida com um painel de dados são os de

[23] Um conjunto de observações em determinado momento do tempo é muitas vezes denominado de *cross-section*.

Efeitos Aleatórios (EA), **Efeitos Fixos** (EF) e **Primeiras-Diferenças** (PD), sendo que cada um deles possui diferentes hipóteses para uma estimativa adequada de impactos de uma política. Uma referência que discute cada um dos métodos de forma bem detalhada é o livro de Wooldridge (2010a). Um exemplo de avaliação de impacto, a partir de dados em estrutura de painel, é o artigo de Kirigia et al. (2008) sobre efetividade de políticas de saúde pública em Angola.