Diferenças em Diferenças

Da intuição ao estimador de Callaway & Sant'Anna

Semana de Avaliação

Duração: 6 horas · Maio de 2026

Bloco de 4 horas · Parte teórica
Bloco de 2 horas · Parte prática

Sobre este curso

Conteúdo, objetivos e roteiro

Objetivos

- Compreender a lógica causal do DiD e quando usá-lo
- Identificar suposições (tendências paralelas, não antecipação, SUTVA)
- Reconhecer quando o TWFE tradicional falha em adoção escalonada
- Aplicar o estimador de Callaway & Sant'Anna (2021)
- Ler e interpretar gráficos de event study corretamente
- Comunicar resultados para tomada de decisão em política pública

Público e pré-requisitos

- Analistas, gestores e pesquisadores de políticas públicas
- Nível: intermediário — sem ênfase em código
- Pré-requisitos: noções de regressão linear, variáveis aleatórias e média
- Familiaridade com dados em painel (unidade × tempo) ajuda, mas não é essencial
- Referências principais: Cunningham (2021) e Callaway & Sant'Anna (2021)

Agenda — 4 horas

Quatro blocos de 50–55 minutos, com intervalos

Bloco 1 · 50 min

Fundamentos de DiD

Potencial contrafactual, ATT, intuição gráfica, equação e decomposição

Bloco 2 · 55 min

Identificação e suposições

Tendências paralelas, event study, não antecipação, SUTVA, covariáveis

Bloco 3 · 55 min

Adoção escalonada e TWFE

Por que o TWFE quebra; Goodman-Bacon; pesos negativos; viés dinâmico

Bloco 4 · 55 min

Callaway & Sant'Anna

ATT(g,t), grupos de comparação, DR, agregações, aplicação a PP

Bloco 1

Fundamentos de DiD

Avaliação causal, potencial contrafactual e a intuição da dupla diferença

Motivação — por que DiD?

Avaliar o efeito causal de uma política exige um contrafactual

A pergunta do analista de política pública

› **O impacto é causal?** Não basta comparar antes e depois — outros fatores mudam.

› **Comparável?** Entre unidades tratadas e não tratadas, há seleção.

› **Robusto?** Precisamos separar tendência subjacente do efeito da política.

› **Para quem?** Efeito médio, por subgrupo, dinâmico ou cumulativo?

Exemplo

Alguns municípios expandem sua rede do SUAS em 2018. Observamos a taxa de pobreza cair 5 pp de 2017 para 2019 nos municípios que receberam novas unidades do CRAS.

Mas a pobreza caiu no país todo. Quanto disso é efeito da política?

DiD isola o efeito comparando com municípios similares que não expandiram a rede.

Origens históricas — DiD nasceu na saúde pública

Antes da econometria formalizar o método, médicos já comparavam tratados e não tratados ao longo do tempo

Ignaz Semmelweis — Viena, 1846

Problema. No Hospital Geral de Viena, parturientes na ala dos médicos morriam de febre puerperal a 13–18%; na ala das parteiras, apenas 3%.

Tratamento. Em 1847 Semmelweis exige lavagem das mãos com cloro apenas na ala dos médicos — as parteiras seguem o protocolo antigo.

Resultado. A mortalidade na ala tratada cai para níveis próximos aos da ala de controle.

Desenho de DiD. Compara a mudança antes/depois entre dois grupos — o controle absorve tendências comuns (estação, ventilação, infraestrutura).

A evidência foi rejeitada na época: a teoria dos "miasmas" ainda dominava. Boa identificação não basta — também é preciso comunicar.

John Snow — Londres, 1854

Problema. Cólera em Londres — a teoria dominante era miasma. Snow suspeitava da água do Tamisa.

Experimento natural. A Lambeth Water Co. moveu a captação de água para montante do centro entre 1849 e 1854; a Southwark & Vauxhall não.

Desenho de DiD. Mortes por cólera por companhia em 1849 (pré) e 1854 (pós). Lambeth = tratada; S&V = controle.

Resultado. A mortalidade desaba na área servida pela Lambeth e não cai na S&V — uma das primeiras provas quantitativas de transmissão hídrica.

Snow também removeu a bomba da Broad Street — política pública embasada em evidência quase-experimental.

Aplicação — expansão do SUAS e dos CRAS

Um caso próximo do nosso dia a dia para fixar a intuição do contrafactual

O contexto

› **SUAS.** Sistema Único de Assistência Social, gerido pelo MDS em parceria com estados e municípios.

› **CRAS.** Centros de Referência de Assistência Social — porta de entrada do SUAS na proteção social básica (cadastro único, busca ativa, PAIF).

› **Expansão.** Em janelas de cofinanciamento federal, novos municípios passam a receber unidades de CRAS conforme criticidade socioeconômica.

› **Pergunta de política.** Municípios que ganharam novos CRAS reduziram pobreza, insegurança alimentar e cobertura do Cadastro Único mais do que reduziriam sem a expansão?

Como DiD ajuda

Tratado: municípios que receberam novos CRAS na janela t.

Controle: municípios eligíveis que ainda não receberam, com perfil socioeconômico semelhante.

Resultado (Y): cobertura do CadÚnico, BPC, beneficiados pelo Bolsa Família, pobreza monetária.

DiD: mudança antes/depois nos tratados, menos a mudança no controle — isola o que é efeito do CRAS do que era tendência geral do país.

Correlação não é causalidade

O analista precisa de um contrafactual

- **Seleção** Quem recebe a política difere de quem não recebe.
- **Variáveis omitidas** Choques econômicos, demográficos, culturais.
- **Causalidade reversa** Talvez o resultado influencie a adoção.
- **Tendências** Sem política, o indicador já estaria mudando?

A ideia do DiD

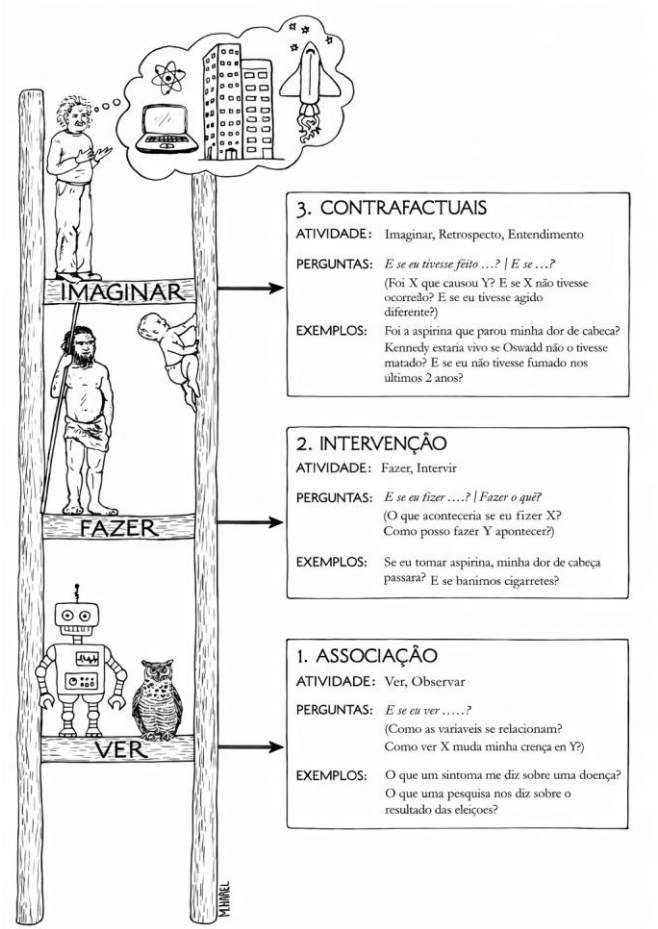
Em vez de comparar níveis, comparamos variações.

Diferença #1: antes vs. depois
Diferença #2: tratados vs. controle

A dupla diferença remove efeitos fixos de unidade e de tempo — sob a suposição de tendências paralelas.

A escada da causalidade — Judea Pearl

Três níveis de raciocínio causal: ver, fazer e imaginar



Três tipos de pergunta — três degraus

O cientista Judea Pearl organizou o raciocínio causal em uma escada de três degraus. Cada degrau responde a um tipo de pergunta — e a estatística tradicional alcança apenas o primeiro.

Ver — como X e Y se movem juntos (associação).

Fazer — o que acontece se eu intervir no sistema.

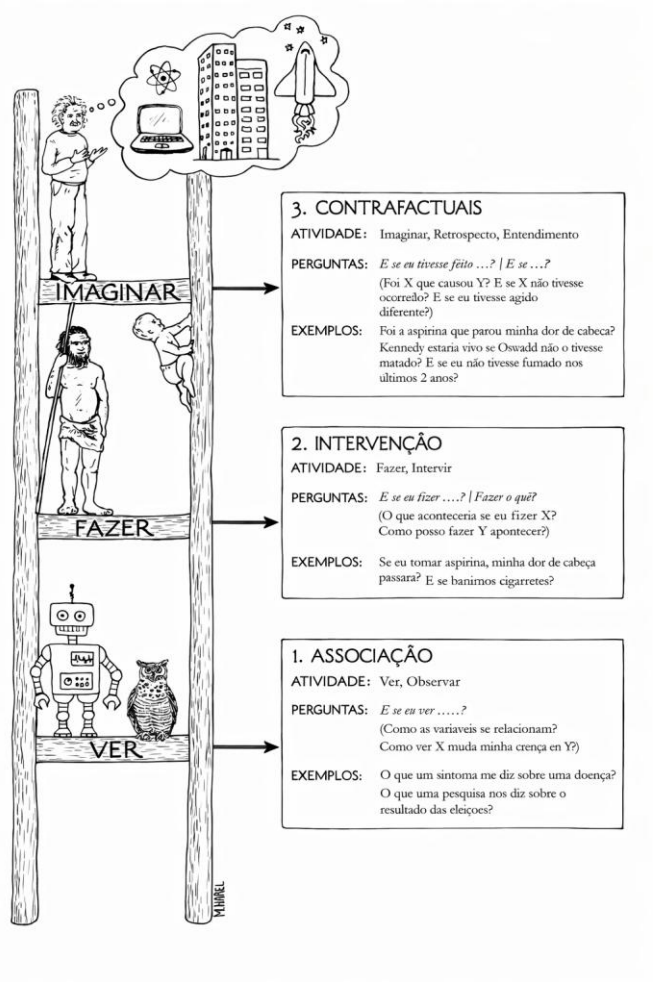
Imaginar — o que teria acontecido se a história fosse outra.

Onde o DiD se encaixa

Avaliar o efeito de uma política é uma pergunta de terceiro degrau: exige imaginar o contrafactual — o que teria acontecido sem ela.

Os três degraus e a econometria

Cada degrau da escada corresponde a um conjunto de ferramentas



1 · Ver — Associação

Covariância e probabilidade condicional

Descrever como duas variáveis se movem juntas. É o terreno da estatística descritiva — e onde mora a armadilha "correlação não é causalidade".

2 · Fazer — Intervenção

Intervenção e experimento

O que acontece quando se age sobre o sistema. É o terreno do experimento controlado e dos ensaios aleatorizados.

3 · Imaginar — Contrafactuais

Contrafactual e inferência

O que teria acontecido sem a política. É exatamente neste degrau que o DiD opera.

No terceiro degrau — o Homem-Leão

A capacidade humana de imaginar o que nunca existiu

O Homem-Leão

Estátua pré-histórica · Alemanha · 35 a 41 mil anos

Combinar um leão e um homem numa figura que nunca existiu é um salto cognitivo: imaginar algo contrafactual.

Para Judea Pearl, é essa capacidade — ausente em qualquer processo puramente associativo — que define o raciocínio do terceiro degrau.

Toda avaliação de impacto mobiliza essa mesma imaginação contrafactual: perguntar o que teria acontecido sem a política.

Resultado potencial

O problema fundamental da inferência causal

Para cada unidade i, dois resultados potenciais existem em cada período t:

$Y_{it}(1)$: resultado se i for tratada | $Y_{it}(0)$: resultado se i NÃO for tratada

Observamos apenas um deles. O outro é o contrafactual.

O que vemos

- Y(1) para unidades tratadas
- Y(0) para unidades não tratadas
- Um painel de unidade × tempo

O que precisamos imaginar

- Y(0) para os tratados: o que teria acontecido sem a política
- DiD produz uma estimativa plausível desse contrafactual
- Sob suposições explícitas e verificáveis (parcialmente)

Notação formal — D, Y(0), Y(1), δ_i

Da pergunta de pesquisa à pergunta causal — com vocabulário preciso

(1) Indicador de tratamento — $D_{it} \in \{0, 1\}$

$$D_{it} = 1 \text{ se a unidade } i \text{ recebeu o tratamento no período } t, \ 0 \text{ caso contrário}$$

(2) Resultados potenciais — o que aconteceria em cada estado do mundo

$$Y_{it}(1) \quad \text{resultado da unidade } i \text{ no tempo } t \text{ SE for tratada}$$
$$Y_{it}(0) \quad \text{resultado da unidade } i \text{ no tempo } t \text{ SE não for tratada}$$

(3) Equação de comutação — o Y observado é um dos dois resultados potenciais

$$Y_{it} = D_{it} \cdot Y_{it}(1) + (1 - D_{it}) \cdot Y_{it}(0)$$

(4) Efeito individual e o problema de dados ausentes

$$\delta_i = Y_i(1) - Y_i(0) \quad \text{efeito individual do tratamento}$$

Para a mesma unidade i, observamos **apenas um** dos dois resultados potenciais — o outro é contrafactual. Daí o **problema fundamental da inferência causal**: δ_i nunca é observável diretamente. O alvo possível é uma **média** sobre uma população — o ATT.

ATT — o parâmetro de interesse

Average Treatment effect on the Treated

Definição

$$ATT = E[\, Y(1) - Y(0) \mid D = 1 \,]$$

efeito médio do tratamento sobre quem foi tratado

› **Por que o ATT?** É o que interessa em avaliação: qual o efeito sobre quem recebeu a política.

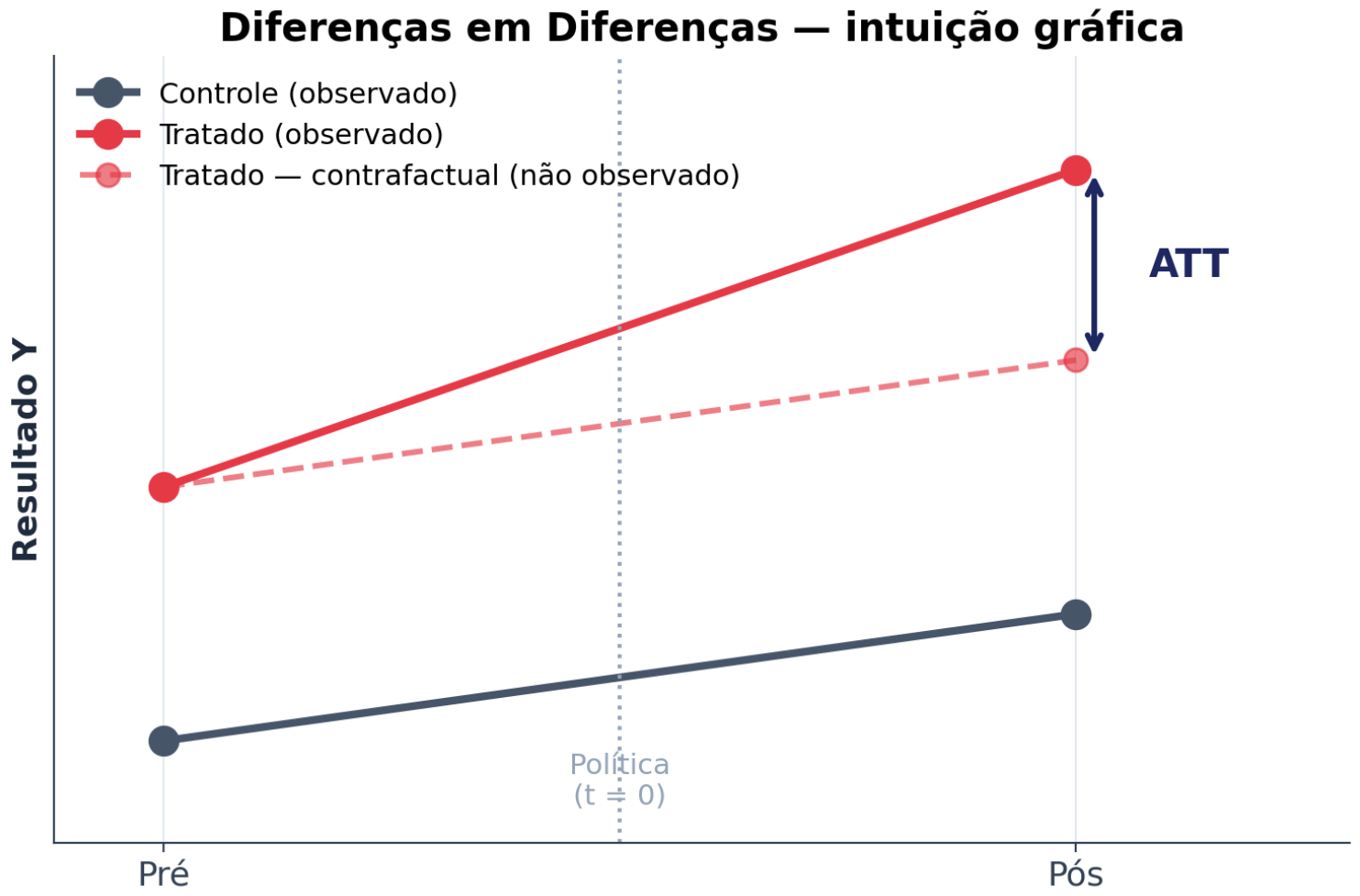
› **ATT vs. ATE** ATE = E[Y(1) − Y(0)] na população inteira; ATT = média só nos tratados. Ex.: 100 municípios elegíveis, 25 recebem CRAS — ATE mede o efeito médio nos 100, ATT só nos 25.

› **Por que DiD identifica só o ATT** Precisamos do contrafactual Y(0) apenas para os tratados; o controle informa só a tendência. Para o ATE, a tendência paralela teria de valer também para quem nunca seria tratado — suposição mais forte (e raramente plausível).

› **Em políticas públicas** Geralmente o ATT é o alvo natural (beneficiários do programa).

A intuição gráfica do DiD

Dupla diferença: antes × depois e tratados × controle



› **Linha cinza** Trajetória do grupo de controle.

› **Linha vermelha contínua** Trajetória observada dos tratados.

› **Linha tracejada** Contrafactual: o que teria ocorrido sem a política, assumindo tendências paralelas.

› **Seta ATT** Efeito do tratamento = diferença entre observado e contrafactual no período pós.

4 médias, 3 subtrações. *Cada extremidade da figura é uma média (tratado pré, tratado pós, controle pré, controle pós). $\delta = (T_{pós} - T_{pré}) - (C_{pós} - C_{pré})$ — três subtrações sobre quatro médias.*

A equação da dupla diferença

"Quatro médias, três subtrações" (Ashenfelter)

Diferença de diferenças de médias

$$\hat{\delta} = (\bar{Y}^T_{pós} - \bar{Y}^T_{pré}) - (\bar{Y}^c_{pós} - \bar{Y}^c_{pré})$$

Lendo a fórmula

- $\bar{Y}^T_{pós} - \bar{Y}^T_{pré}$ variação observada no grupo tratado entre antes e depois.
- $\bar{Y}^c_{pós} - \bar{Y}^c_{pré}$ variação no grupo de controle — proxy da tendência que o tratado teria sem a política.
- δ diferença entre as duas variações — é o efeito atribuído à política sob tendências paralelas.

DiD como regressão OLS — equivalência com 4 médias

O coeficiente da interação é numericamente igual à conta de Ashenfelter (Cunningham, 01-Basics)

Especificação (dois grupos, dois períodos)

$$Y_{it} = \alpha_0 + \alpha_1 \cdot Treat_i + \alpha_2 \cdot Post_t + \delta \cdot (Treat_i \times Post_t) + \varepsilon_{it}$$

Lendo cada coeficiente

- α_0 média do grupo de controle no pré
- α_1 diferença de nível entre tratado e controle no pré
- α_2 variação temporal comum a tratado e controle (do pré para o pós)
- δ **coeficiente da interação — o ATT sob tendências paralelas, em 2 grupos e 2 períodos**

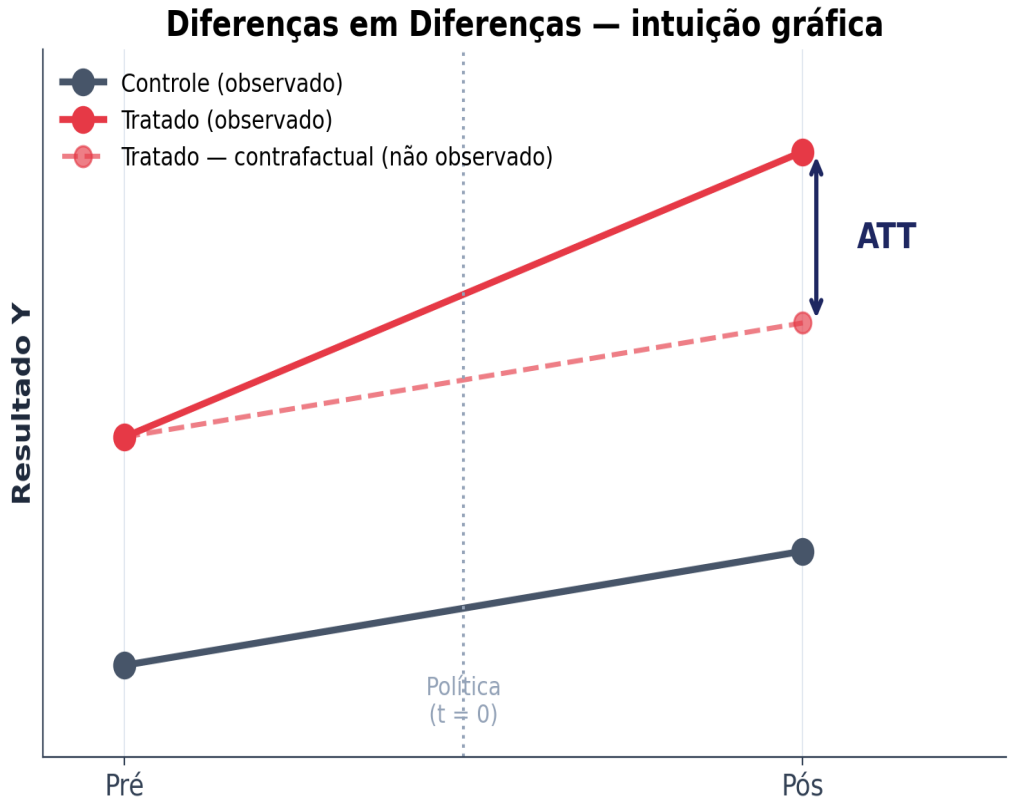
Equivalência (Ashenfelter): o δ do OLS é **numericamente igual** ao δ calculado por "4 médias e 3 subtrações"

Por que usar regressão? • erros padrão legíveis • acomoda covariáveis pré-tratamento • escala para event study, FE de unidade/tempo e DiD escalonado

Da intuição gráfica à equação OLS

Cada coeficiente da regressão tem um nome no gráfico — o OLS apenas formaliza o que se vê

$$Y_{it} = \alpha_0 + \alpha_1 \cdot Treat_i + \alpha_2 \cdot Post_t + \delta \cdot (Treat \times Post) + \varepsilon_{it}$$



α_0 Intercepto — média do controle no pré (onde a linha cinza começa no gráfico).

α_1 Diferença de nível tratado vs controle no pré — o "gap" vertical entre as duas curvas no início.

α_2 Variação temporal comum — a inclinação compartilhada (linha cinza de pré→pós e a tracejada do contrafactual).

δ ATT

Abertura **vertical extra** entre a linha tratada observada e a tracejada (contrafactual sob PTA) no pós — a "seta ATT" do gráfico.

Cada coeficiente do OLS é uma das "quatro médias" do gráfico — o estimador apenas formaliza a leitura visual.

Exemplo clássico — Card & Krueger (1994)

Salário mínimo e emprego: NJ vs. PA

O contexto

› **Tratamento** Nova Jersey subiu o salário mínimo em abril de 1992.

› **Controle** Pensilvânia não alterou — condições econômicas similares.

› **Resultado** Emprego em fast-food em NJ e PA, antes e depois.

› **Abordagem** DiD de duas diferenças — nem natural experiment puro, nem simples regressão.

O resultado e o debate

› **Efeito estimado** Aumento do salário mínimo NÃO reduziu emprego — até subiu levemente.

› **Contra a intuição clássica** Literatura prévia previa queda.

› **Debate metodológico** Seleção de controle, medição, validade externa — todos questionados.

› **Legado** O paper tornou DiD quasi-experimental mainstream na economia aplicada.

Outro clássico — Mariel Boatlift (Card, 1990)

Choque migratório em Miami: efeito sobre o mercado de trabalho local

O contexto

› **Choque.** Em 1980, Fidel Castro abre o porto de Mariel; ~125 mil cubanos chegam a Miami entre abril e outubro.

› **Tratamento.** Aumento súbito de ~7% na força de trabalho de Miami, concentrado em baixa qualificação.

› **Controle.** Atlanta, Houston, Los Angeles, Tampa-St. Petersburg — porte e composição setorial próximos.

› **Pergunta.** Como um choque migratório inesperado afeta salários e desemprego dos nativos menos qualificados?

O resultado e o debate

› **Efeito estimado.** Card (1990) encontra pouco ou nenhum efeito sobre salários e emprego, mesmo entre cubanos já residentes e negros.

› **Contra a intuição clássica.** Modelos competitivos previam queda salarial — ajustes ocorreram via produção e composição setorial.

› **Debate.** Borjas (2017) reanalisa com subgrupos e acha efeitos negativos sobre não-graduados — mostra sensibilidade à definição de tratado/controle.

› **Legado.** Junto com Card & Krueger, consolidou DiD como ferramenta-padrão em economia do trabalho.

Aplicações em políticas públicas (Brasil)

Onde o DiD aparece na avaliação de PP brasileiras

Bolsa Família

Pobreza e frequência escolar

Municípios com maior cobertura × menor cobertura

Mais Médicos

Mortalidade infantil e cobertura APS

Municípios que receberam médicos × similares sem

Programa Minha Casa

Formalização e preço do aluguel

Regiões com forte expansão × sem obras

Lei Seca 2008

Acidentes de trânsito noturnos

Séries temporais de municípios antes/depois

Fundeb e ENEM

Desempenho em redes

Redes com ganho maior de repasse × menor

IR - Lei Kandir

Arrecadação e exportações

Estados exportadores × não exportadores

Bloco 2

Identificação e suposições

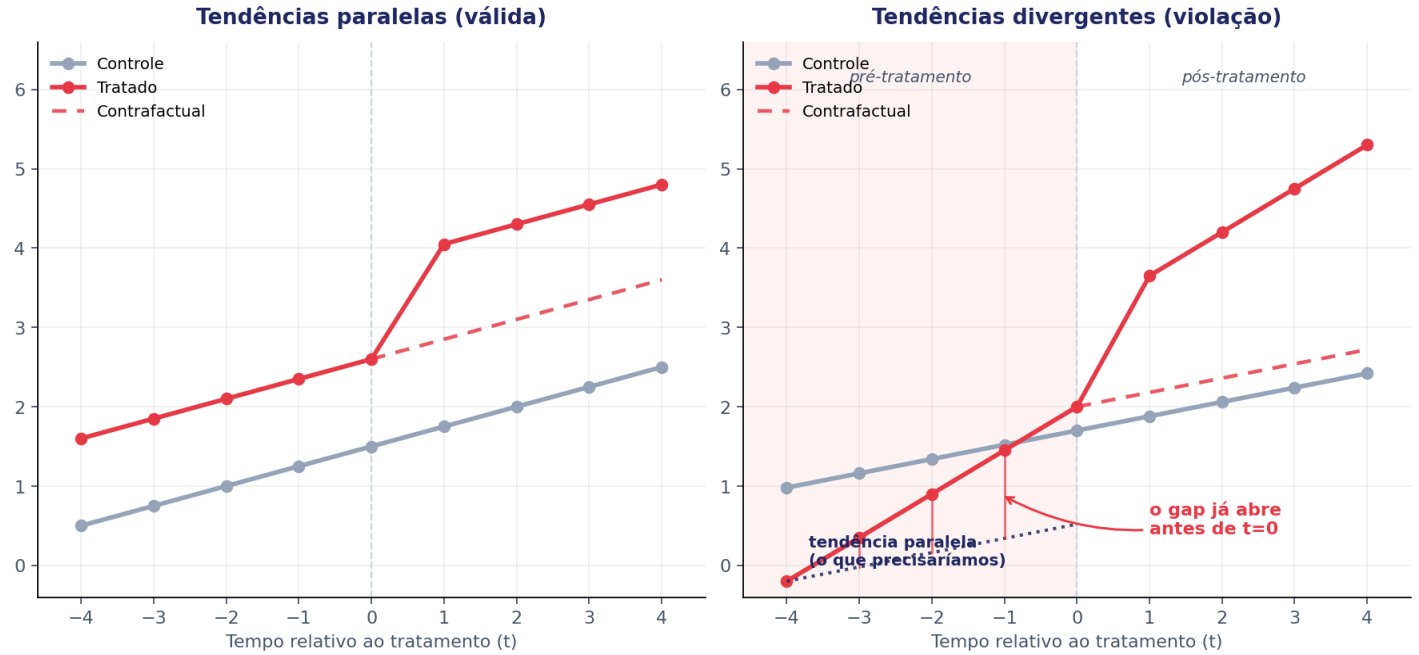
Quando o DiD identifica o ATT — e como diagnosticamos

Suposição 1 — Tendências paralelas

A mais importante (e mais escrutinada)

Parallel Trends Assumption (PTA)

$$E[\,\Delta Y(0) \mid D = 1\,] \;=\; E[\,\Delta Y(0) \mid D = 0\,]$$



› **Painel esquerdo** Trajetórias caminham paralelas antes de t=0.

› **Painel direito** Tratados já sobem mais rápido — PTA suspeita.

› **Implicação** Se PTA falha, o DiD atribui à política algo que seria tendência.

Derivação: por que δ = ATT + viés

Quatro médias → resultados potenciais → somando um zero → ATT + viés (Cunningham)

(1) Estimador amostral — quatro médias, três subtrações

$$\hat{\delta} = (\bar{Y}^{T}_{pós} - \bar{Y}^{T}_{pré}) - (\bar{Y}^{C}_{pós} - \bar{Y}^{C}_{pré})$$

(2) Resultados potenciais — substituir Y pelo resultado potencial correspondente

$$\delta = (E[Y(1)|D=1,pós] - E[Y(0)|D=1,pré]) - (E[Y(0)|D=0,pós] - E[Y(0)|D=0,pré])$$

(3) Reagrupar — ATT + viés de tendências não-paralelas

$$\{E[Y(1)|D=1,pós] - E[Y(0)|D=1,pós]\} \leftarrow \text{ATT}$$
$$+ \{(E[Y(0)|D=1,pós] - E[Y(0)|D=1,pré]) - (E[Y(0)|D=0,pós] - E[Y(0)|D=0,pré])\} \leftarrow \text{viés (zero sob PTA)}$$

Reagrupando por cores — δ = ATT + viés

Cada termo já aparece pintado com a cor do grupo onde vai parar

ATT

viés

Expressão expandida (resultados potenciais + zero adicionado)

$$\delta = +E[Y(1)|D=1,pós] \; -E[Y(0)|D=1,pré] \; -E[Y(0)|D=0,pós] \; +E[Y(0)|D=0,pré]$$

+ zero adicionado: $+E[Y(0)|D=1,pós] \; -E[Y(0)|D=1,pós]$

Reagrupando os termos pela cor

$$ATT = +E[Y(1)|D=1,pós] \; -E[Y(0)|D=1,pós]$$

$$viés = +E[Y(0)|D=1,pós] \; -E[Y(0)|D=1,pré] \; -E[Y(0)|D=0,pós] \; +E[Y(0)|D=0,pré]$$

$$\delta = ATT + viés$$

O par +E[Y(0)|D=1,pós] / −E[Y(0)|D=1,pós] aparece em cores opostas e se cancela — metade vai para ATT, metade para o viés.

O que o DiD estima — decomposição

Quando δ é ATT? Quando carrega viés?

$$\delta = ATT + \{ E[\Delta Y(0) \mid D=1] - E[\Delta Y(0) \mid D=0] \}$$

Interpretação

- **O termo entre chaves é o viés** Diferença entre as trajetórias potenciais sem tratamento.
- **Zero sob tendências paralelas** Se, na ausência da política, tratados e controle teriam a mesma variação média.
- **Mensagem-chave** DiD não exige níveis iguais entre tratados e controle — exige trajetórias paralelas no cenário sem política.
- **Consequência prática** Podemos comparar grupos com níveis bem diferentes, desde que a dinâmica subjacente seja similar.

Suposição 2 — Não antecipação

Unidades não reagem antes do tratamento ocorrer

Antes do tratamento, os resultados potenciais Y(0) e Y(1) coincidem.

O que pode violar

› **Anúncios prévios** Política é anunciada com meses de antecedência.

› **Expectativas** Empresas / domicílios se preparam para o tratamento (investem, contratam, mudam hábitos).

› **Exemplo** Anúncio de revisão do Cadastro Único: famílias correm para se cadastrar antes da vigência, e a cobertura sobe no "pré-tratamento" — o Y já reflete a política antes dela entrar formalmente.

› **Como mitigar** Incluir janela de antecipação nos ATT(g,t); inspecionar pré-trends próximos a t = 0.

Quando a não antecipação falha — violação e viés

Se o pré já estava sob tratamento, o DiD 2×2 identifica três termos — não o ATT

(1) Setup — o tratado já era tratado no pré

$$\hat{\delta} = \left(E[Y|D=1,\text{pós}] - E[Y|D=1,\text{pré}] \right) - \left(E[Y|D=0,\text{pós}] - E[Y|D=0,\text{pré}] \right)$$

(2) Substituir pelos resultados potenciais — tratado tem Y(1) no pré e no pós

$$\delta = \left(E[Y(1)|D=1,\text{pós}] - E[Y(1)|D=1,\text{pré}] \right) - \left(E[Y(0)|D=0,\text{pós}] - E[Y(0)|D=0,\text{pré}] \right)$$

(3) Reagrupar — três termos identificados pelo DiD 2×2 sob violação de NA

$$\delta = ATT(\text{pós}) + \text{viés de tendências não-paralelas} - ATT(\text{pré})$$

Consequência — se efeitos são **constantes**, ATT(pós) = ATT(pré) → o DiD pode dar **zero** mesmo havendo efeito real. **Defesa:** redefinir o pré para período claramente anterior · event study com leads · excluir últimos períodos pré.

Reagrupando por cores — violação de NA

Três cores, três grupos — ATT(pós), viés e −ATT(pré)

ATT(pós) | viés de tendências não-paralelas | −ATT(pré)

Expressão expandida (resultados potenciais + dois zeros adicionados)

$$\delta = +E[Y(1)|D=1,\text{pós}] - E[Y(1)|D=1,\text{pré}] - E[Y(0)|D=0,\text{pós}] + E[Y(0)|D=0,\text{pré}]$$

$$\text{+ dois zeros: } +E[Y(0)|D=1,\text{pós}] - E[Y(0)|D=1,\text{pós}] \;;\; +E[Y(0)|D=1,\text{pré}] - E[Y(0)|D=1,\text{pré}]$$

Reagrupando os termos pela cor

$$\text{ATT(pós)} = +E[Y(1)|D=1,\text{pós}] - E[Y(0)|D=1,\text{pós}]$$

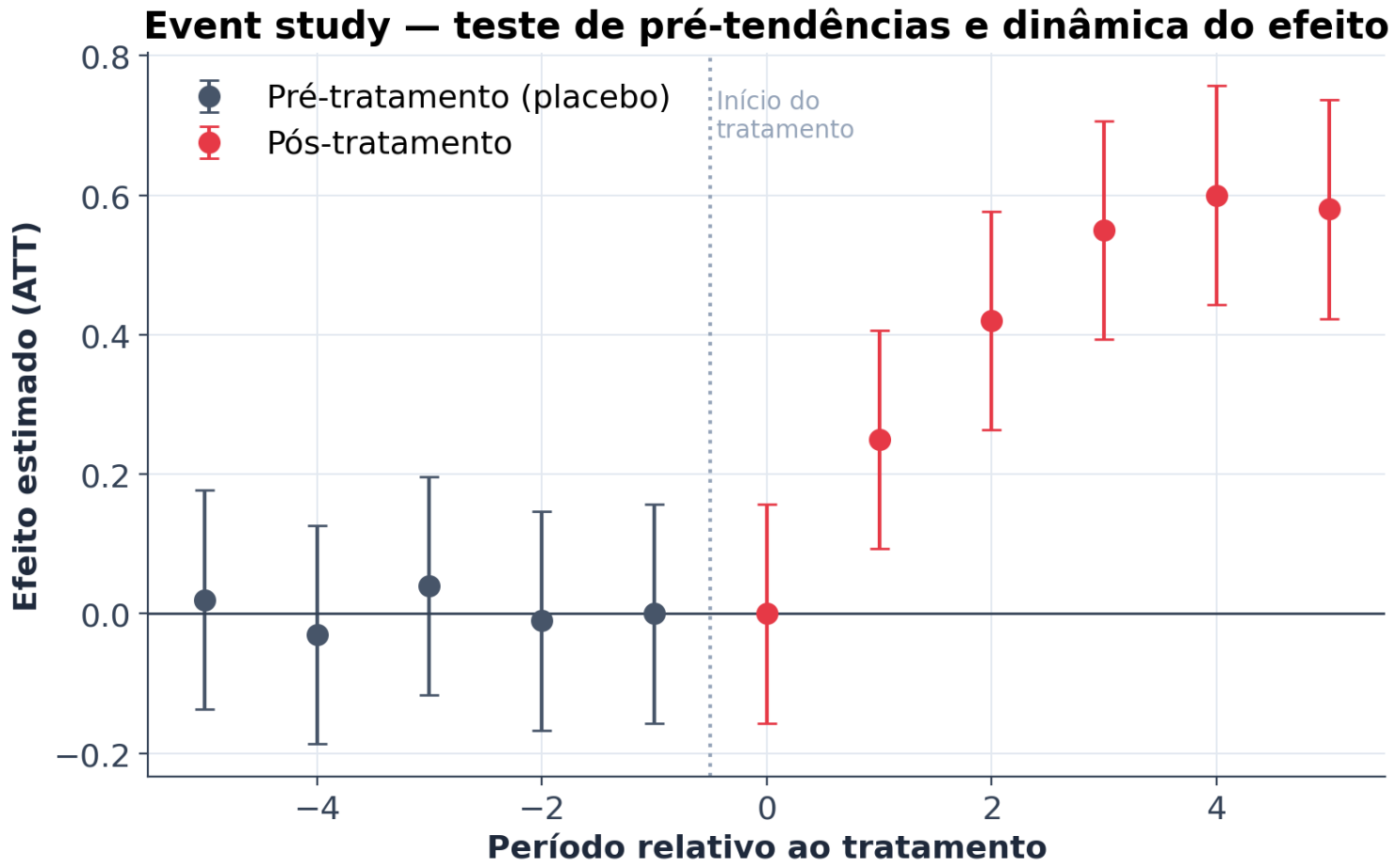
$$\text{viés} = +E[Y(0)|D=1,\text{pós}] - E[Y(0)|D=1,\text{pré}] - E[Y(0)|D=0,\text{pós}] + E[Y(0)|D=0,\text{pré}]$$

$$-\text{ATT(pré)} = -E[Y(1)|D=1,\text{pré}] + E[Y(0)|D=1,\text{pré}]$$

$$\delta = \text{ATT(pós)} + \text{viés} - \text{ATT(pré)}$$

Diagnóstico — event study

Testando pré-tendências e revelando a dinâmica do efeito



› **$\hat{\mu}_\tau$ ($\tau < -1$): pré-tratamento** Diferença média tratado − controle no período τ relativo, com $\tau = -1$ omitido como baseline. Sob PTA + NA, devem ser ~0.

› **$\tau = -1$: baseline (omitido)** Todos os outros coeficientes são lidos em relação a este período — por isso o ponto sobre $\tau = -1$ fica em zero.

› **δ_τ ($\tau \geq 0$): pós-tratamento** ATT estimado τ períodos após o tratamento. δ_0 = efeito imediato, δ_1, δ_2, ... = efeitos defasados; mostram persistência ou decaimento.

› **Atenção** $\hat{\mu}_\tau \approx 0$ é evidência compatível com PTA, não prova; e o efeito médio do DiD 2×2 corresponde, grosso modo, à média dos $\delta_\tau \geq 0$.

Suposição 3 — SUTVA e spillovers

Stable Unit Treatment Value Assumption

› **Consistência** Quando i é tratado, o Y observado é exatamente o resultado potencial Y(1) — não há "versões" diferentes do tratamento (ex.: CRAS aberto em todo lugar tem a mesma estrutura, equipe e oferta de serviços).

› **No interference** O tratamento de j não afeta o resultado de i.

› **Isto é violado quando...** Há transbordamento geográfico, concorrência entre mercados, difusão de políticas.

› **Exemplos em PP** CRAS atrai famílias do município vizinho; migração entre municípios; difusão de boas práticas entre prefeituras vizinhas.

› **Diagnóstico** Desenhos com buffers, fronteiras, testes de sensibilidade por distância.

Sinal de alerta

Se o 'controle' sofre influência da política (spillovers), ele deixa de ser um bom proxy de Y(0): se o efeito sobre o controle é positivo, o DiD SUBESTIMA o efeito; se é negativo, SUPERESTIMA.

Ex.: CRAS atrai famílias do município vizinho — o "controle" também melhora seus indicadores, e o DiD subestima o ganho real.

Quando a SUTVA falha — spillover e viés

Se o "controle" recebe parte do tratamento (spillover), o DiD 2×2 também identifica três termos

(1) Setup — o controle não é puro: recebe o tratamento por spillover ou já era tratado

$$\hat{\delta} = \left(E[Y|D=1,\text{pós}] - E[Y|D=1,\text{pré}] \right) - \left(E[Y|D=0,\text{pós}] - E[Y|D=0,\text{pré}] \right)$$

(2) Substituir pelos resultados potenciais — o controle exibe Y(1) por contaminação

$$\delta = \left(E[Y(1)|D=1,\text{pós}] - E[Y(0)|D=1,\text{pré}] \right) - \left(E[Y(1)|D=0,\text{pós}] - E[Y(1)|D=0,\text{pré}] \right)$$

(3) Reagrupar — três termos identificados pelo DiD 2×2 sob violação de SUTVA

$$\delta = \text{ATT(tratado, pós)} + \text{viés de tendências não-paralelas} - \Delta\text{ATT(controle)}$$

Consequência — a magnitude do ΔATT(controle) determina o tamanho do viés. Se o "controle" tem efeito **parecido com o tratado** (spillover forte) → o DiD vai para **zero** mesmo com efeito real positivo (subestimação). **Defesa:** excluir municípios contíguos a tratados · usar controle geograficamente isolado · medir intensidade do spillover (raio, fluxo cadastral).

Reagrupando por cores — violação de SUTVA

Três cores, três grupos — ATT(tratado, pós), viés e −ΔATT(controle)

ATT(tratado, pós) **viés** **−ΔATT(controle)**

Expressão expandida (resultados potenciais + três zeros adicionados)

$$\delta = +E[Y(1)|D=1,pós] -E[Y(0)|D=1,pré] -E[Y(1)|D=0,pós] +E[Y(1)|D=0,pré]$$

+ três zeros: $+E[Y(0)|D=1,pós] -E[Y(0)|D=1,pós]$; $+E[Y(0)|D=0,pós] -E[Y(0)|D=0,pós]$; $+E[Y(0)|D=0,pré] -E[Y(0)|D=0,pré]$

Reagrupando os termos pela cor

$$ATT(tratado, pós) = +E[Y(1)|D=1,pós] -E[Y(0)|D=1,pós]$$

$$viés = +E[Y(0)|D=1,pós] -E[Y(0)|D=1,pré] -E[Y(0)|D=0,pós] +E[Y(0)|D=0,pré]$$

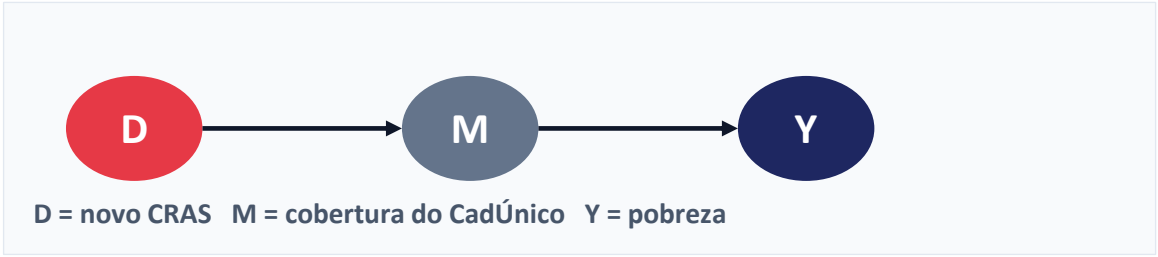
$$-\Delta ATT(controle) = -E[Y(1)|D=0,pós] +E[Y(1)|D=0,pré] +E[Y(0)|D=0,pós] -E[Y(0)|D=0,pré]$$

$$\delta = ATT(tratado, pós) + viés - \Delta ATT(controle)$$

Mediador × colisor — aplicado à expansão do SUAS

Nem toda variável relacionada a D e Y deve entrar como covariável

Mediador (não condicionar)

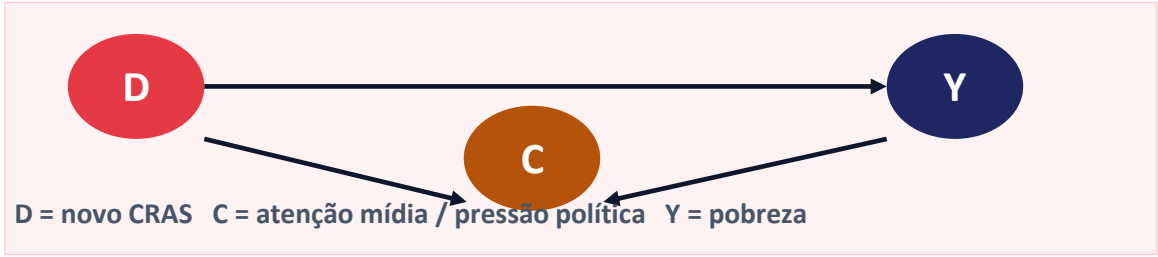


O CRAS aumenta cobertura do CadÚnico, que por sua vez reduz pobreza. M está **no caminho causal entre D e Y**.

NÃO condicionar em M. Se controlamos por M, bloqueamos parte do efeito que queremos medir — o ATT cai e **subestimamos** o efeito do CRAS sobre pobreza.

Regra: covariáveis devem ser medidas em t = 0 (pré-tratamento), nunca pós.

Colisor (não condicionar)



Tanto a abertura do CRAS quanto a redução de pobreza geram cobertura noticiosa C: C **recebe setas de D e Y** — C é um colisor.

NÃO condicionar em C. Filtrar a amostra por atenção política/mídia abre um caminho espúrio entre D e Y — introduzimos correlação onde não havia, podendo **superestimar** ou subestimar.

Mediador: filho de D, não controlar. Colisor: filho de D e Y, não controlar. Confundidor: pai comum de D e Y, controlar.

Quando a PTA costuma falhar

Padrões recorrentes a procurar antes de publicar resultados

Grupos heterogêneos — *Urbano vs. rural, capital vs. interior, Norte vs. Sudeste*

Trajetórias estruturalmente diferentes antes da política.

Mean reversion — *Municípios que adotam a política quando o resultado está pior*

Regressão à média cria falso efeito positivo.

Choques específicos — *Crise afeta mais um setor; política focaliza outro*

Tratados e controle enfrentam trajetórias diferentes por razões externas.

Composição mudando — *Mudança no rol de respondentes / migração / fechamento*

Compara-se maçãs com laranjas ao longo do tempo.

Efeitos de longo prazo — *Tratamentos com resposta lenta*

Janela de análise curta pode não capturar efeito— e longa sofre com choques.

Tendências paralelas condicionais

Quando PTA incondicional não se sustenta

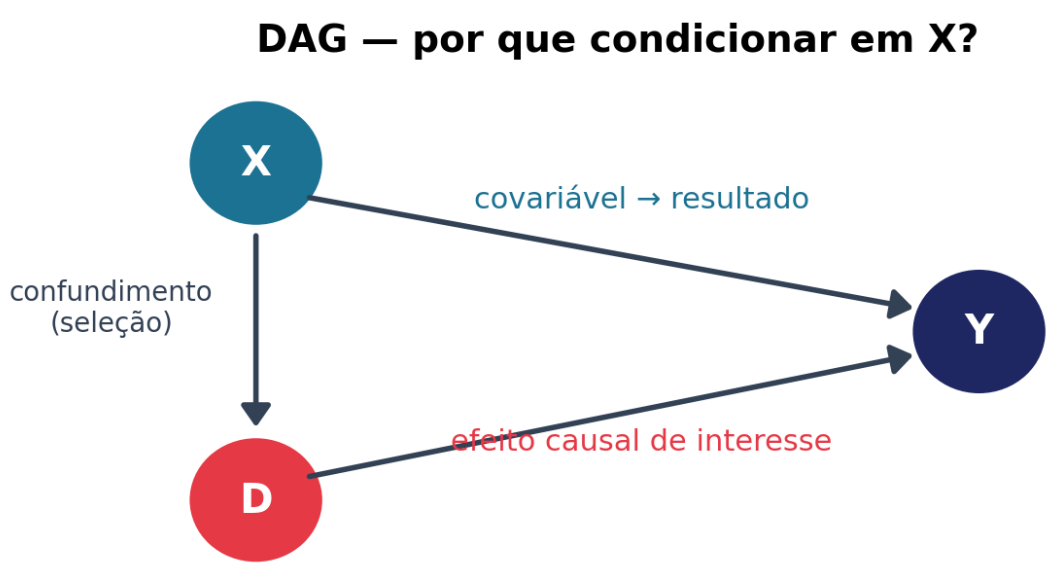
Conditional PTA

$$E[\,\Delta Y(0) \mid X, D = 1\,] = E[\,\Delta Y(0) \mid X, D = 0\,]$$

- **Ideia** Trajetórias paralelas apenas dentro de subgrupos definidos por X.
- **Por que ajuda** Recupera comparabilidade ao condicionar em covariáveis pré-tratamento.
- **X é pré-tratamento** Nunca use variáveis afetadas pelo tratamento — causa viés (mediador).
- **Estimadores de CPT** Outcome regression, IPW e Doubly Robust (abordado adiante).

Escolher covariáveis — pensando com DAGs

Não inclua de tudo: pense na estrutura causal



› **Incluir** Covariáveis pré-tratamento que afetam Y e potencialmente a adoção.

› **Evitar** Mediadores (afetados pelo tratamento) — introduzem viés.

› **Evitar** Colliders — condicionar neles abre caminhos espúrios.

› **Dica prática** Sempre medidas em período anterior à política. Nunca atualizadas dinamicamente.

Estimadores com covariáveis

Três formas de operacionalizar CPT

Outcome Regression (OR)

Heckman, Ichimura & Todd (1997): regrida ΔY em X só nos controles e use o modelo ajustado para imputar Y(0) dos tratados.

› Especificação simples: $\Delta Y_i = \beta_0 + X_i'\beta + u_i$ estimado no controle
› Sensível a extrapolação fora do suporte comum de X.
› Exige um modelo correto da relação de X e Y

Inverse Prob. Weighting (IPW)

Estime propensity score e(X) = P(D=1 | X); pondere observações

› Balanceia covariáveis sem modelar o resultado
› Sensível a scores extremos (0 ou 1) — aparamento ajuda
› Exige modelo correto da adoção

Doubly Robust (DR)

Combina OR e IPW; consistente se apenas um dos dois estiver certo

› Recomendado por Callaway & Sant'Anna (2021)
› Eficiência assintótica melhor
› Mais robusto a má especificação

Como escolher covariáveis

Foco em variáveis que prevejam o Y(0) faltante — não apenas o tratamento

O que buscar — princípios

› **Preditoras de Y(0).** Falta o Y(0) dos tratados — escolhemos X que ajude a aproximar essa trajetória ausente (renda, IDH-M, cobertura prévia do CadÚnico).

› **Confundidores.** Variáveis que afetam tanto D quanto Y — bom senso, DAGs e conhecimento de domínio ajudam a identificá-las.

› **O que NÃO incluir.** Mediadores (filhos de D) e colisores (filhos de D e Y) — controlar nesses casos introduz viés (ver slide anterior).

› **Sempre em t = baseline.** Medir X no período pré-tratamento (b) — Abadie joga fora valores pós-tratamento porque a meta é restabelecer paralelismo no baseline.

3 abordagens data-driven (Cunningham)

(1) Só controles. Descarte os tratados pós-tratamento. Todos os Y restantes são Y(0). Regrida Y(0) nas X candidatas e selecione as significativas.

(2) Só tratados no pré. Use só os tratados, no baseline (ainda Y(0)). Regrida Y(0) em X de baseline e selecione.

(3) ΔY(0) dos tratados. Calcule a variação $E[\Delta Y(0)|D=1]$ entre t−2 e t−1; regrida em X medido em t−1 e selecione — escolhe X que prevê a tendência ausente.

Procedimento orientado a dados — ajuda a identificar candidatos a confundidores, mas não checa balanceamento (próximo slide).

Regras de bolso

› Comece com lista pequena baseada em conhecimento de domínio (renda, IDH-M, população, IGD-M, cobertura prévia). Documente as escolhas antes de olhar resultados.

› Use os 3 procedimentos data-driven como check, não como receita única. Compare as listas: covariáveis robustas aparecem em mais de uma.

Abadie (2005) — IPW para corrigir o DiD

Tornando controles comparáveis aos tratados por repesagem em X

O problema com o DiD tradicional

O DiD 2×2 supõe tendências paralelas **incondicionais**. Se tratados e controles diferem em X observáveis que afetam Y(0), o controle não representa o contrafactual e o ATT fica viesado.

Ideia de Abadie: reponderar o grupo controle para que sua composição em X "pareça" com a dos tratados — daí o DiD passa a valer condicionalmente em X.

1. Propensity score p(X)

$$p(X) = P(D = 1 \mid X)$$

› Probabilidade de receber tratamento dado X, estimada via logit ou probit.

› Reduz a dimensionalidade de X a um único escalar (Rosenbaum & Rubin, 1983).

› Estimado em t = baseline (pré-tratamento), só com X. Abadie descarta valores pós-tratamento.

2. Pesos IPW

$$w(X) = p(X) / (1 - p(X))$$

› Peso aplicado apenas ao grupo controle — o objetivo é "puxar" sua composição para parecer com a dos tratados.

› Controle com X "parecido com tratado" → p(X) alto → peso grande.

› Controle muito diferente → p(X) baixo → peso pequeno. Pesos extremos pedem aparamento.

Abadie (2005) — receita em 4 passos e hipóteses

Do propensity score ao ATT em quatro etapas — e o que precisa valer para que funcione

1

Estimar p̂(X)

Rodar logit ou probit de D em X medido em t = baseline.

Cada unidade ganha probabilidade entre 0 e 1 de ter sido tratada.

2

Construir pesos IPW

`w_i = p(X_i)/(1-p(X_i))`

Aplicado só aos controles. Faça aparamento nos extremos.

3

Calcular ΔY = Y_pós − Y_pré

Para cada unidade, a primeira diferença sobre o tempo (o "antes/depois" do DiD).

4

Comparar ΔY ponderados

Média de ΔY entre tratados menos média ponderada de ΔY entre controles → ATT consistente.

Hipóteses-chave + Legado

› **Hipóteses:** (i) tendências paralelas condicionais em X; (ii) não antecipação; (iii) modelo do propensity score corretamente especificado; (iv) suporte comum (0 < p(X) < 1).

› **Legado:** precursor dos doubly robust modernos; é o "motor" usado dentro de Callaway & Sant'Anna (2021); muito utilizado em avaliação de políticas públicas.

Diagnóstico — balanceamento de covariáveis

Antes de estimar efeitos, verifique a comparabilidade

› **Diferenças padronizadas** |SMD| < 0,1 é bom; > 0,25 é preocupante (Stuart, 2010).

$$\text{Norm.Diff}_\omega = (\bar{X}_T - \bar{X}_C) / \sqrt{(S^2_T + S^2_C) / 2}$$

Imbens & Rubin (2015): aceitável se < 0,25; ideal < 0,10. Compara médias entre tratado e controle padronizando pela variância combinada.

› **Overlap / suporte comum** A distribuição do propensity score deve ter massa nas duas caudas.

› **Rauch ou aparamento** Excluir observações com e(X) extremo (ex.: < 0,05 ou > 0,95).

› **Sensibilidade** Rode o modelo com e sem aparamento, com e sem controles.

› **Variáveis 'descartadas'** Documente quais covariáveis excluiu e por quê — evite pesca.

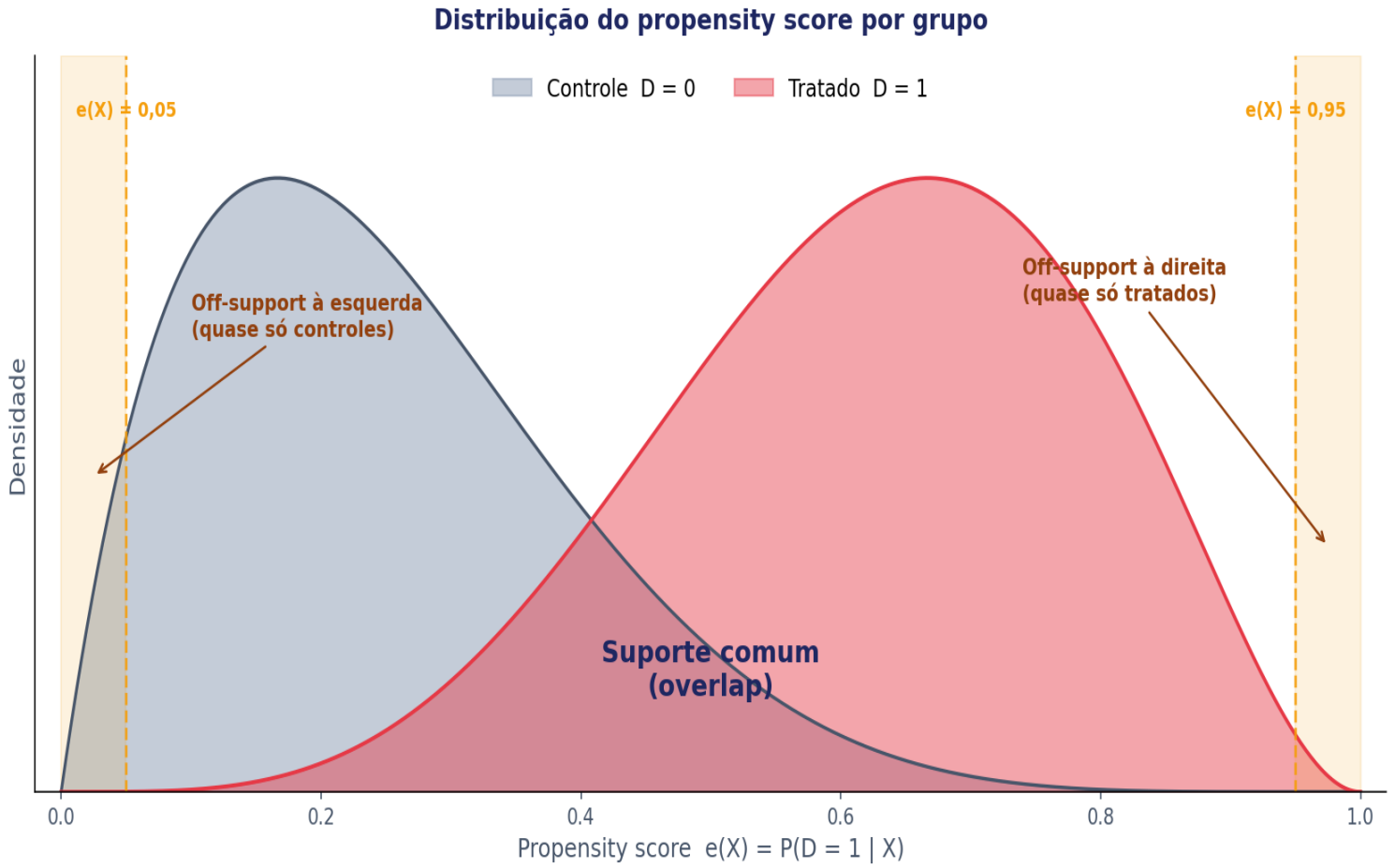
Regra de bolso

Se o balanceamento está ruim mesmo condicionando em X, o DiD está pedindo demais dos dados. Considere:

- Outra unidade de análise
- Redefinir tratados / controle
- Matching + DiD
- Outro método (synthetic control, RDD)

Suporte comum — visualizando o overlap

Onde tratados e controles têm propensity score parecido — sem isso o ATT não está identificado



› **Definição** Existe sobreposição em e(X) entre D = 1 e D = 0; só conseguimos comparar tratados e controles "parecidos".

› **Off-support** Caudas onde só um grupo aparece (ex.: e(X) ≈ 0 ou ≈ 1). Aí o DiD extrapola — risco alto.

› **O que fazer** Aparar observações com e(X) extremo (ex.: < 0,05 ou > 0,95) e reporte o ATT na amostra aparada. O alvo deixa de ser todos os tratados, mas vira identificável.

› **Lembrete** Suporte comum é condição necessária; balanceamento em X (slide anterior) é checagem complementar.

Propensity score — receita prática

Aplicando o IPW de Abadie à expansão dos CRAS — em quatro passos

(1) Escolha X pré-tratamento. Renda, população, IDH-M, cobertura prévia do CadÚnico, IGD-M — sempre medidos antes da expansão do CRAS.

(2) Estime ê(X). Rode probit ou logit de D em X. Cada município passa a ter uma probabilidade entre 0 e 1 de ter recebido CRAS.

(3) Checagens. Verifique suporte comum (sobreposição das distribuições de ê(X) entre D=0 e D=1) e faça aparamento se houver scores extremos (ex.: < 0,05 ou > 0,95).

(4) Estime o ATT. Aplique IPW: cada controle entra na diferença antes/depois com peso ê(X)/(1−ê(X)). DiD identifica o ATT sob tendências paralelas condicionais + suporte comum + modelo correto do propensity score.

Outcome Regression — Heckman, Ichimura & Todd (1997)

Em vez de repesar o controle, modelamos diretamente o resultado e imputamos Y(0) dos tratados

Problema fundamental da avaliação

Queremos **ATT = E[Y(1) − Y(0) | D = 1]**, mas para os tratados Y(0) é contrafactual — não observado.

Ideia de HIT (1997): ajustar uma regressão do resultado nos controles e usá-la para prever Y(0) de cada tratado.

O método em fórmula

$$m_0(X) = E[\, Y \mid D = 0, X \,] \qquad \hat{Y}^0_i = \hat{m_0}(X_i)$$

$$\hat{ATT} = (1/n\_T) \cdot \Sigma \,[\, Y_i - \hat{m_0}(X_i) \,]$$

- **$m_0(X)$** é a função de resultado esperada para o controle, dado X — ajustada **somente com os controles**.
- Para cada tratado i, imputamos seu Y(0) usando **$\hat{Y}^0_i = \hat{m}_0(X_i)$** — "rodamos o tratado pela equação do controle".
- O ATT é a média, entre os tratados, da diferença entre o observado e o imputado.

Outcome Regression — passos, limitações e contraste com IPW

Receita em três passos, onde OR funciona melhor e onde quebra

1

Ajustar $\hat{m}_0(X)$

Regredir Y em X usando só o grupo controle (D = 0). OLS, polinômios, spline ou ML, conforme a complexidade.

2

Imputar $\hat{Y}^0_i$

Para cada tratado i, calcular $\hat{Y}^0_i = \hat{m}_0(X_i)$ — o município "como teria se comportado" sem a política

3

Estimar o ATT

$\hat{ATT}$ = média entre os tratados de $Y_i - \hat{Y}^0_i$ — diferença entre observado e contrafactual imputado.

Limitações

- Muito sensível à especificação de $\hat{m}_0$; má forma funcional → viés.
- Extrapolação fora do suporte comum de X piora a imputação.
- Por isso inspirou os estimadores doubly robust (Sant'Anna & Zhao, 2020).

OR vs. IPW — intuição

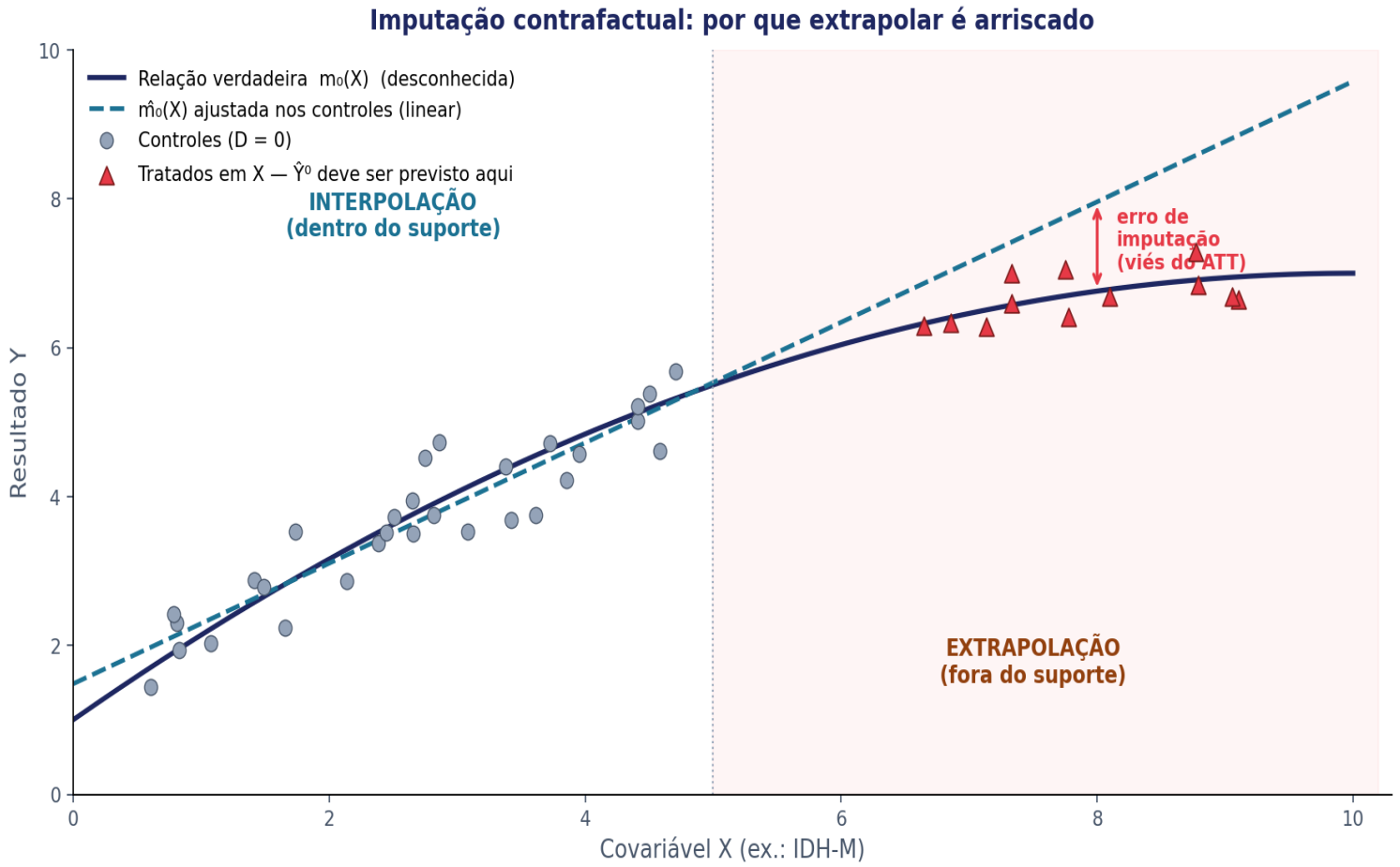
IPW (Abadie): repesa observações via propensity score — depende de modelar D corretamente.

OR (HIT): prevê diretamente o contrafactual — depende de modelar Y(0) corretamente.

DR: combina os dois — basta que um deles esteja certo.

Interpolação vs. extrapolação — por que falta de overlap quebra o OR

Quando os tratados ficam fora do suporte dos controles, $\hat{m}_0(X)$ só pode extrapolar — e o viés do ATT entra silenciosamente



› **Interpolação (zona segura)** Dentro do suporte comum, $\hat{m}_0(X)$ foi calibrada por dados próximos; a previsão de $\hat{Y}^0_i$ é razoável.

› **Extrapolação (zona arriscada)** Fora do suporte não há controle parecido. $\hat{m}_0(X)$ extrapola — a previsão depende inteiramente da forma funcional escolhida (linear, polinômio, ML…).

› **Por que importa** O erro de imputação entra direto no ATT ($Y_i - \hat{Y}^0_i$). E ele é silencioso: o ajuste in-sample do modelo continua bonito.

› **Como mitigar** Verifique o overlap; faça aparamento (perde validade externa, ganha interna); use Doubly Robust — basta que OR ou IPW estejam corretos.

Doubly Robust DiD — Sant'Anna & Zhao (2020)

Duas chances de acertar — se OR ou IPW estiver bem especificado, o ATT é consistente

Por que combinar OR e IPW?

OR precisa de $\hat{m}_0(X)$ bem especificada; IPW precisa de $\hat{p}(X)$ bem especificado. Cada um é inconsistente quando seu modelo falha. **DR combina os dois — assim ganhamos "duas chances de estar certo, ao invés de uma".**

O que ganhamos

› Consistente se **pelo menos um** dos modelos ($\hat{p}$ ou $\hat{\mu}_0,\Delta$) estiver corretamente especificado.

› Não existe estimador consistente do ATT com menor variância — DR é o teto teórico de precisão (Hahn, 1998).

› É o "motor" do estimador de Callaway & Sant'Anna (2021) que aparece no próximo bloco.

Como o DR funciona

1. Reponderar cada unidade pelo propensity score (parte IPW).

2. Para cada unidade, tomar a diferença entre a mudança observada (ΔY) e a mudança que o modelo de OR previu para um controle parecido; em seguida, ponderar essa diferença pelo propensity score.

3. Se o OR estiver correto, essa diferença tem média zero — qualquer ponderação dá o ATT certo. Se o IPW estiver correto, os pesos balanceiam os grupos e cancelam o erro do OR. Daí "duas chances".

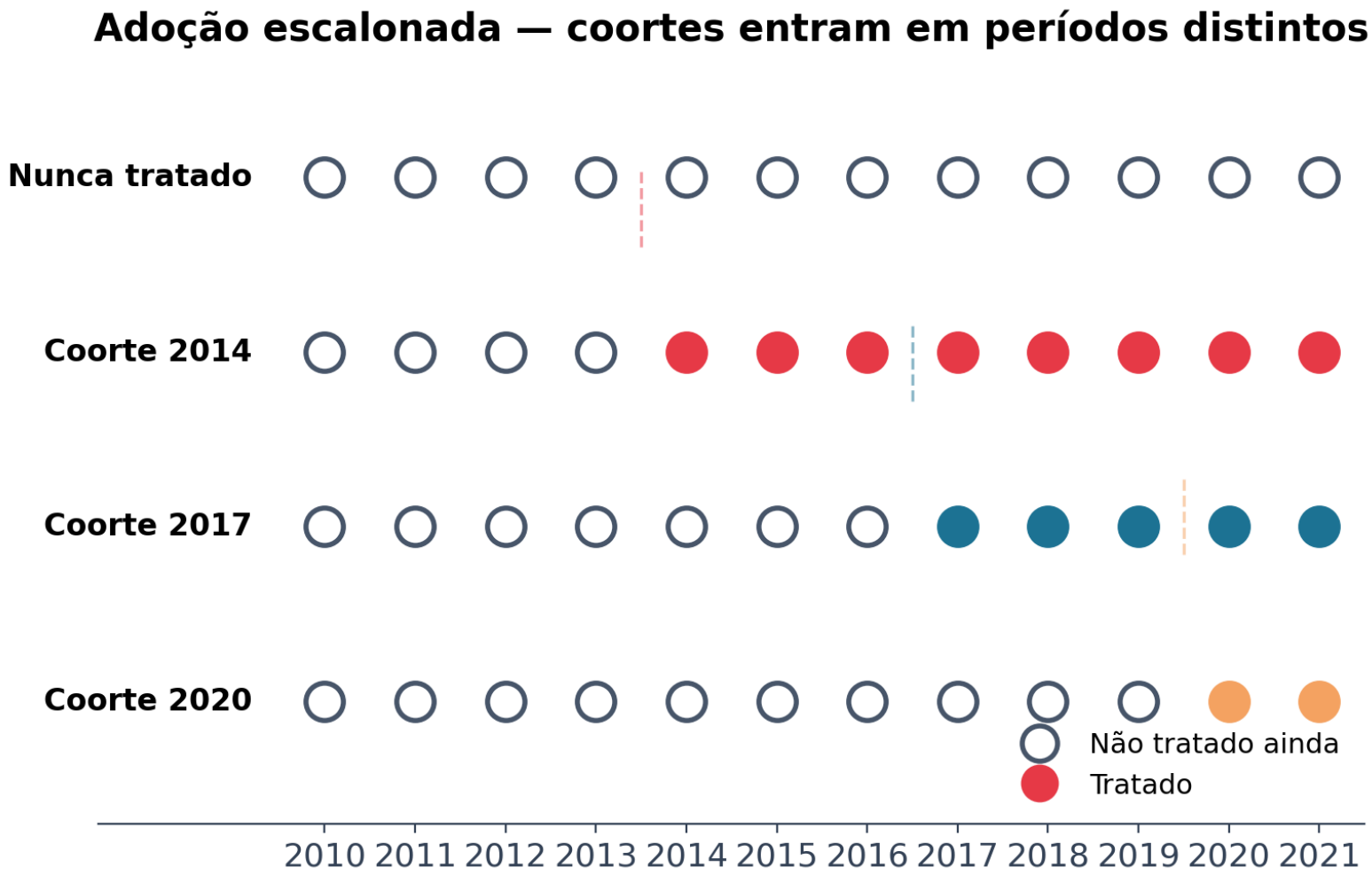
Bloco 3

Adoção escalonada e o problema do TWFE

Quando coortes entram em períodos distintos, o estimador tradicional quebra

Adoção escalonada — o que é

Coortes de unidades entram no tratamento em tempos distintos



› **Comum em PP** Leis adotadas por UF ou município em momentos diferentes.

› **Absorvente** Uma vez tratada, sempre tratada (hipótese usual de CS).

› **Não absorvente** Tratamento pode 'ligar e desligar' — exige outros estimadores (ex.: de Chaisemartin & D'Haultfœuille).

› **Atenção** Comum confundir com múltiplos períodos de tratamento simples.

O estimador TWFE — dois fixos

A receita padrão dos últimos 30 anos

$$Y_{it} = \alpha_i + \lambda_t + \delta \cdot D_{it} + \varepsilon_{it}$$

› **O atrativo** Simples, disponível em qualquer software, parece resolver fixos de unidade e tempo.

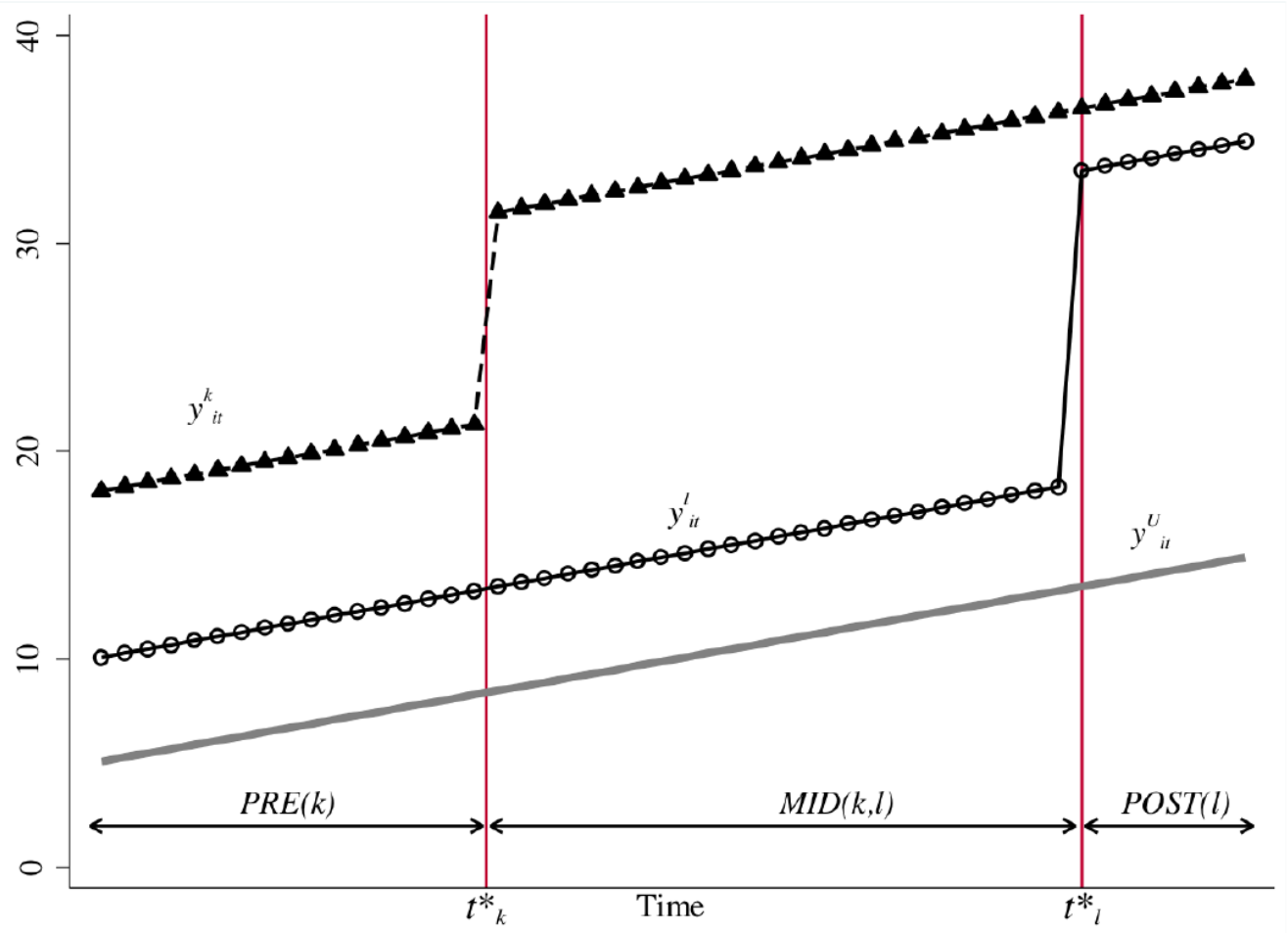
› **Quando δ = ATT?** Em um mundo com dois períodos, dois grupos — OU com efeitos homogêneos.

› **O problema** Em adoção escalonada com efeitos heterogêneos, δ NÃO é a média dos ATT(g,t).

› **Descoberta central** Goodman-Bacon (2021), de Chaisemartin & D'Haultfœuille (2020), Callaway & Sant'Anna (2021), Sun & Abraham (2021).

Goodman-Bacon — o diagnóstico

TWFE é uma média ponderada de comparações 2×2



› **2×2 OK** Tratado vs. nunca tratado; tratado cedo vs. tarde ANTES dos tarde entrarem.

› **2×2 problemático** Tratado tarde vs. cedo DEPOIS dos cedo entrarem — usa tratados como 'controle'.

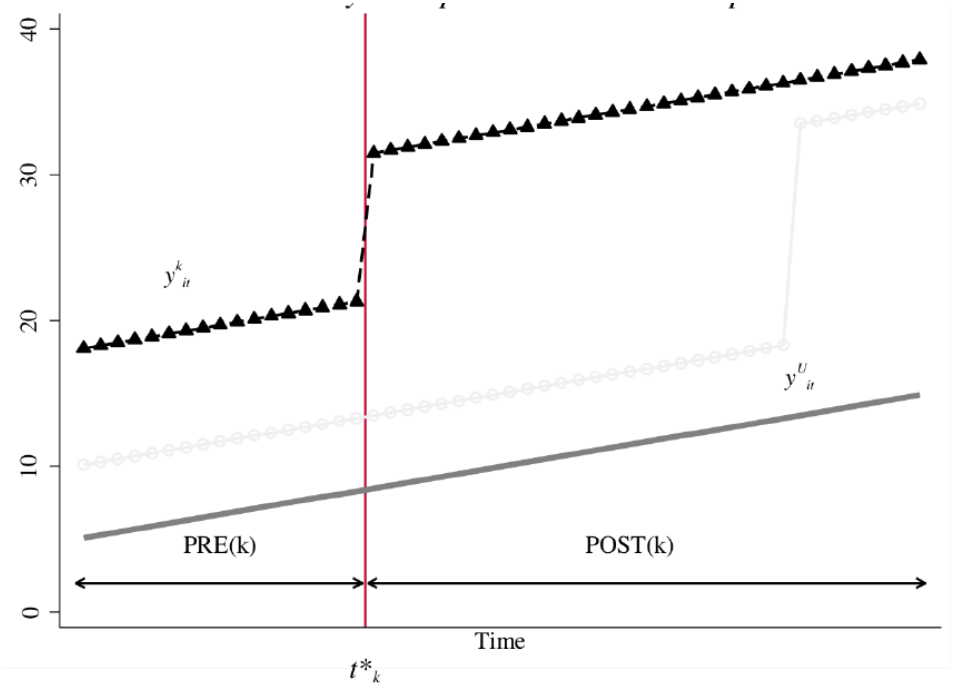
› **Consequência** Se o efeito cresce com o tempo, esse 2×2 produz contaminação — pode até trocar de sinal.

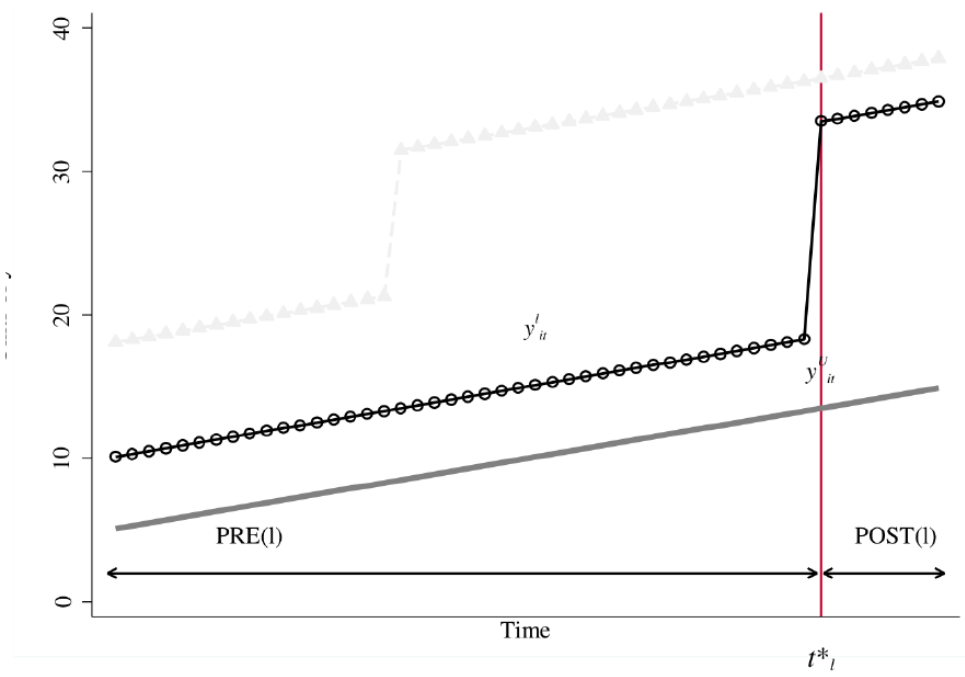
› **Leitura** TWFE = média ponderada desses 2×2. Bom diagnóstico, mas não corrige.

Goodman-Bacon — o diagnóstico

TWFE é uma média ponderada de comparações 2×2

› **2×2 OK** Tratado vs. nunca tratado; tratado cedo vs. tarde ANTES dos tarde entrarem.





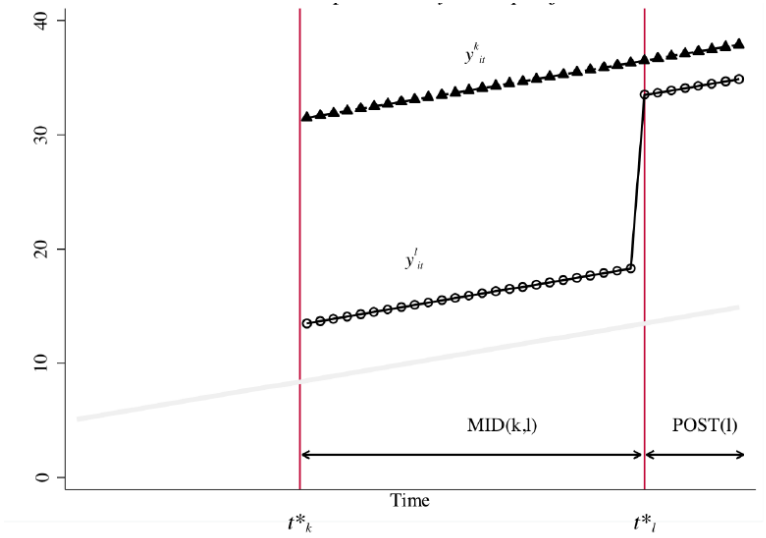
Goodman-Bacon — o diagnóstico

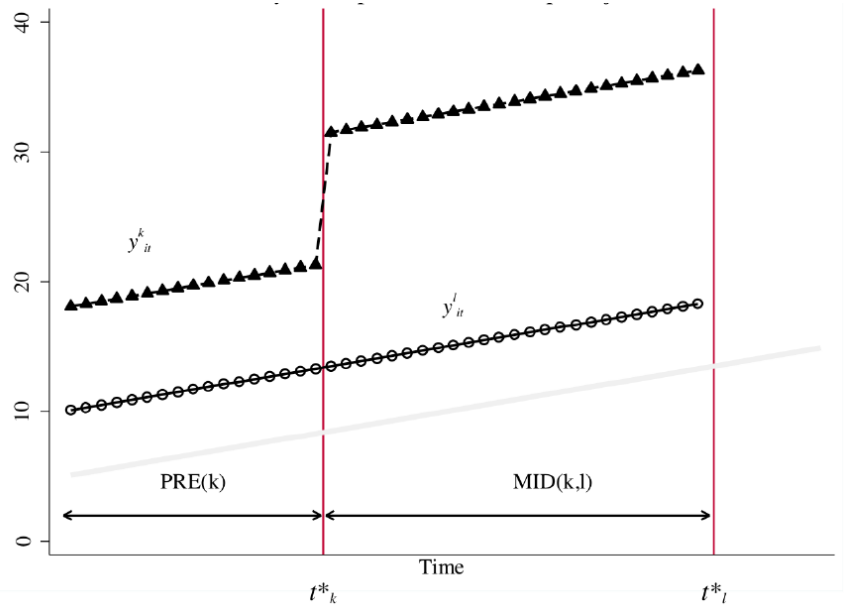
TWFE é uma média ponderada de comparações 2×2

› **2×2 OK** Tratado vs. nunca tratado; tratado cedo vs. tarde ANTES dos tarde entrarem.

› **2×2 problemático** Tratado tarde vs. cedo DEPOIS dos cedo entrarem — usa tratados como 'controle'.

› **Consequência** Se o efeito cresce com o tempo, esse 2×2 produz contaminação — pode até trocar de sinal.





SUTVA ↔ Bacon — a mesma álgebra com nomes diferentes

O "bad 2×2" do Goodman-Bacon é uma violação de SUTVA aparecendo no desenho escalonado

Lembrete — NA é outro caso: a contaminação está no próprio tratado (pré já em Y(1)) → **δ = ATT(pós) + viés − ATT(pré)**

Violação de SUTVA

Quem se contamina: o **controle** — recebe spillover do tratado ou já era tratado.

Forma identificada:

$$\delta = ATT(tratado) + viés - \Delta ATT(controle)$$

Termo extra: $-\Delta ATT(controle)$ — quanto o efeito sobre o controle mudou entre pré e pós

"Bad 2×2" de Goodman-Bacon

Quem se contamina: coorte tratada antes (**k**) usada como **controle** da coorte tratada depois (**l**) — k já tem Y(1) na janela.

Forma identificada:

$$\delta\hat{}_{l\ vs\ k} = ATT(l) + viés - \Delta ATT(k)$$

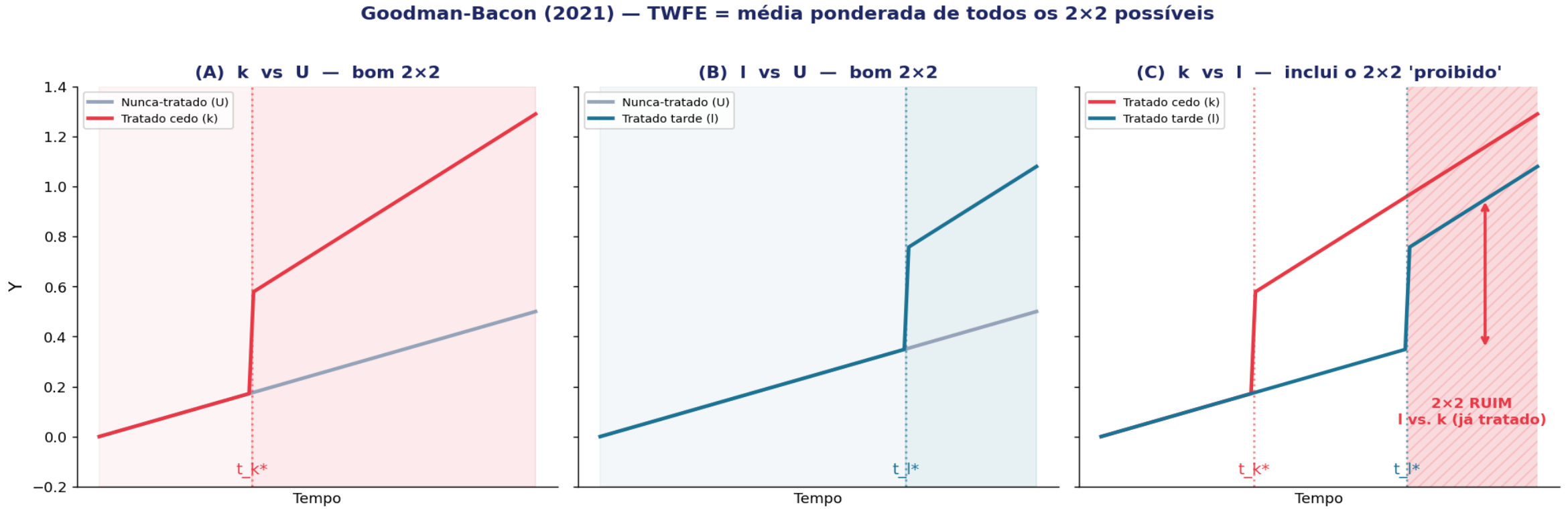
Termo extra: $-\Delta ATT(k)$ — quanto o efeito sobre k mudou entre o pré e o pós da janela

Mesma álgebra, contextos diferentes

Bacon é a violação de **SUTVA** aparecendo no desenho escalonado — por isso CS proíbe coortes já tratadas como controle.

Goodman-Bacon — os três tipos de 2×2 do TWFE

TWFE com adoção escalonada é uma média ponderada de TODAS as comparações 2×2 possíveis



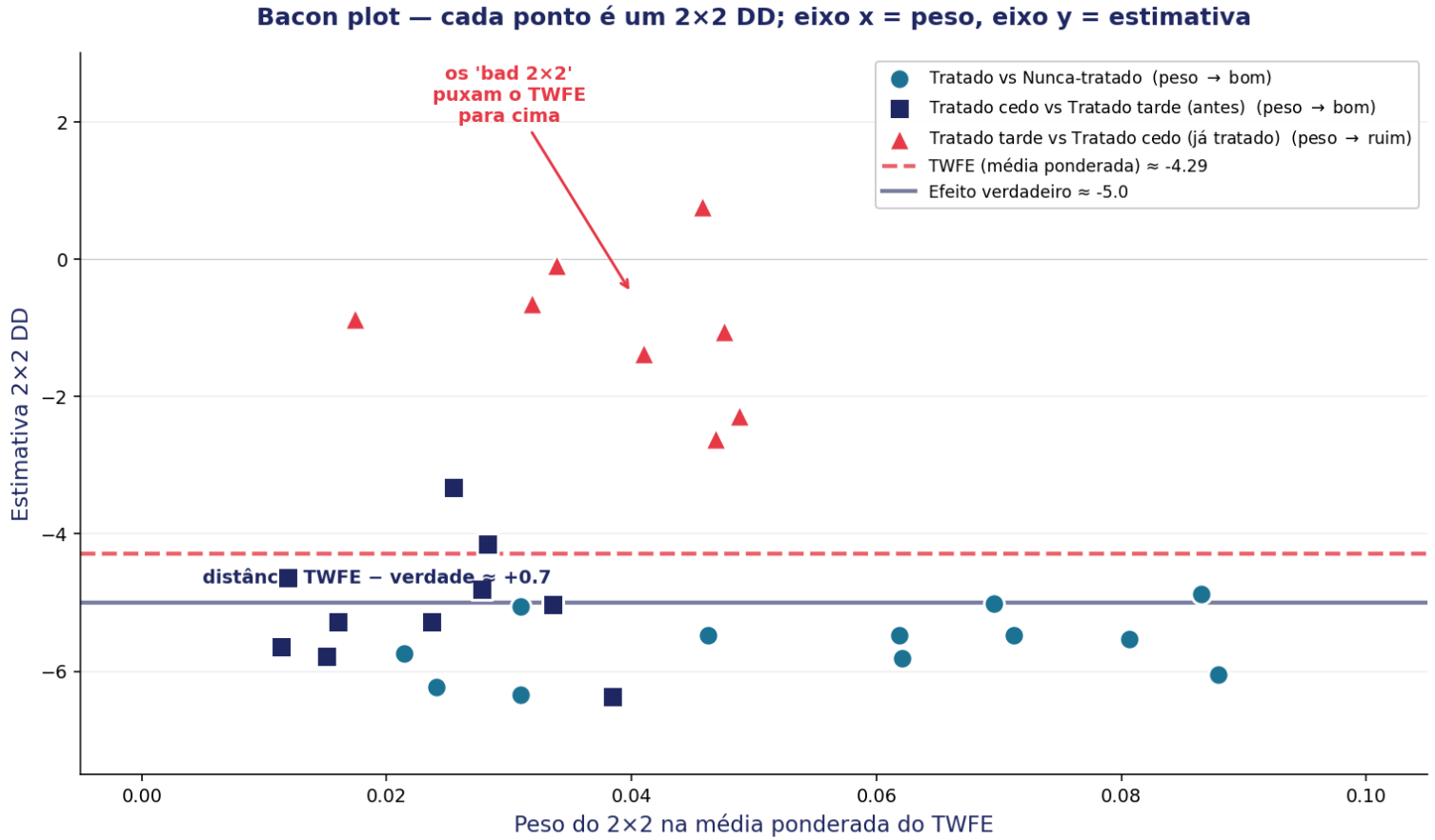
(A) k vs U
Coorte tratada cedo vs. nunca-tratado. Comparação válida — controle nunca foi exposto.

(B) l vs U
Coorte tratada tarde vs. nunca-tratado. Mesma lógica de (A) — controle limpo.

(C) k vs l — inclui o "bad 2×2"
Após t_l*, k já está tratada mas serve como "controle" de l — contamina a comparação.

Bacon plot — visualizando o TWFE como média ponderada

Replica o exemplo Stevenson & Wolfers (2006): TWFE ≈ −4.3 vs efeito verdadeiro ≈ −5



Inspirado em Goodman-Bacon (2021, Fig. 6) — replicação ilustrativa do exemplo Stevenson & Wolfers (2006).

Como ler

Cada **ponto** é um 2×2 que entra no TWFE.

Eixo **x = peso**; eixo **y = estimativa** daquele 2×2.

O que ver

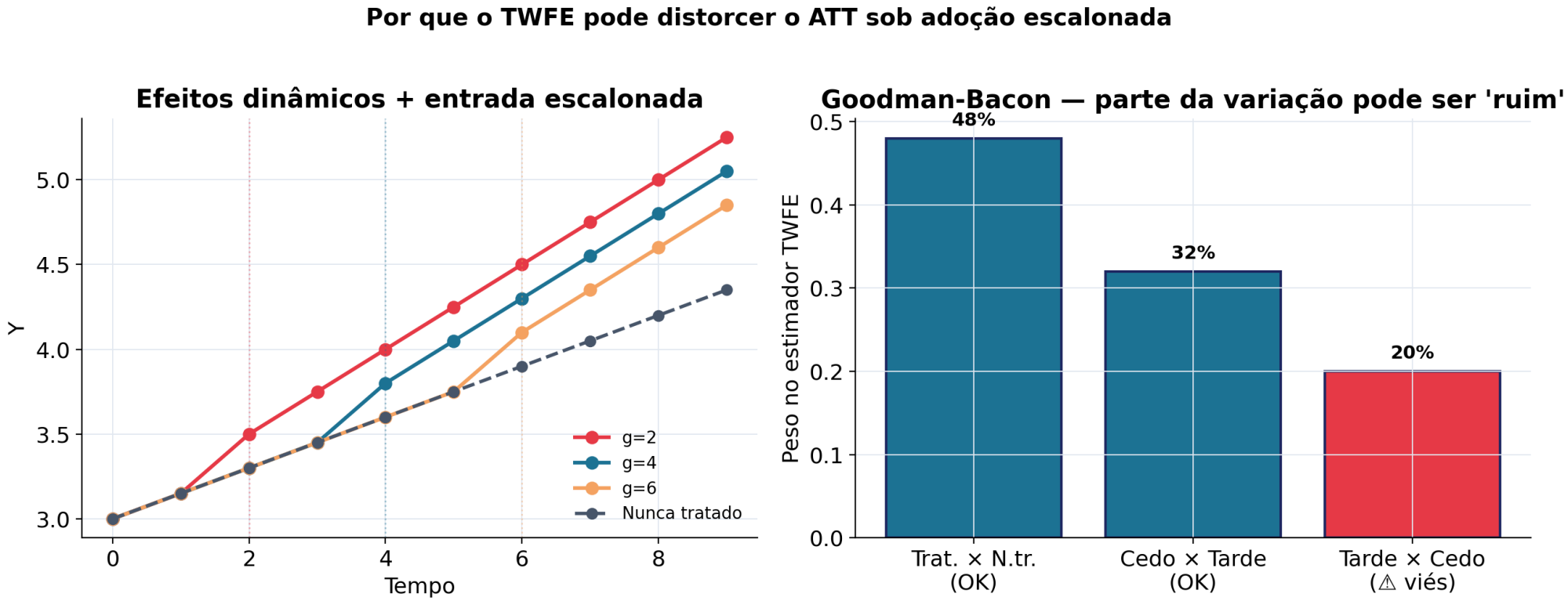
- ● **círculos**: tratado vs nunca-tratado.
- ■ **quadrados**: cedo vs tarde (tarde ainda intacto).
- ▲ **triângulos**: tarde vs cedo (cedo já tratado) — **os ruins**.

Bandeira vermelha: triângulos com peso alto ou puxando o TWFE para longe da nuvem.

Para rodar no Stata: `bacondecomp` · No R: `bacondecomp::bacon()`

Pesos negativos e o viés

Quando 'bons' 2×2 são minoria, o TWFE se afasta do ATT



› **Leitura** O TWFE mistura comparações boas e ruins com pesos dados pela variação em D.

› **Sinal de alerta** Efeitos dinâmicos + adoção escalonada = TWFE pode estar enviesado mesmo sob PTA.

Por que pesos negativos podem aparecer no TWFE

A intuição por trás do diagnóstico de Goodman-Bacon

2×2 bom

Tratado vs. nunca-tratado, ou tratado cedo vs. tratado tarde ANTES de o tarde virar tratado.

Peso na média ponderada: positivo

Contribuição: soma corretamente a comparação que o pesquisador queria fazer.

2×2 ruim (forbidden contrast)

Tratado tarde (l) usando como "controle" a coorte que JÁ foi tratada (k) — k está com Y(1).

Peso na média ponderada: pode ser negativo quando o efeito de k cresce no tempo.

Contribuição: subtrai o ΔATT(k), contaminando o estimador de l.

Por que o sinal pode inverter — Imagine que TODAS as coortes ganham com a política (efeito positivo, crescente). O 2×2 ruim subtrai o ΔATT da coorte k já tratada com peso negativo. Se esses **pesos negativos** dominarem (poucos controles nunca-tratados, espaçamento grande entre coortes), o estimador TWFE pode dar **negativo** — sinal oposto ao efeito real.

Quando isso machuca mais

Efeitos **dinâmicos** (crescem/decrescem) · **poucos nunca-tratados** · **coortes muito espaçadas no tempo**. Soma: mesmo com PTA válida, o TWFE pode estar distante do ATT.

Bloco 4

O estimador de Callaway & Sant'Anna

ATT(g, t) — a abordagem moderna para adoção escalonada

A ideia central — ATT(g, t)

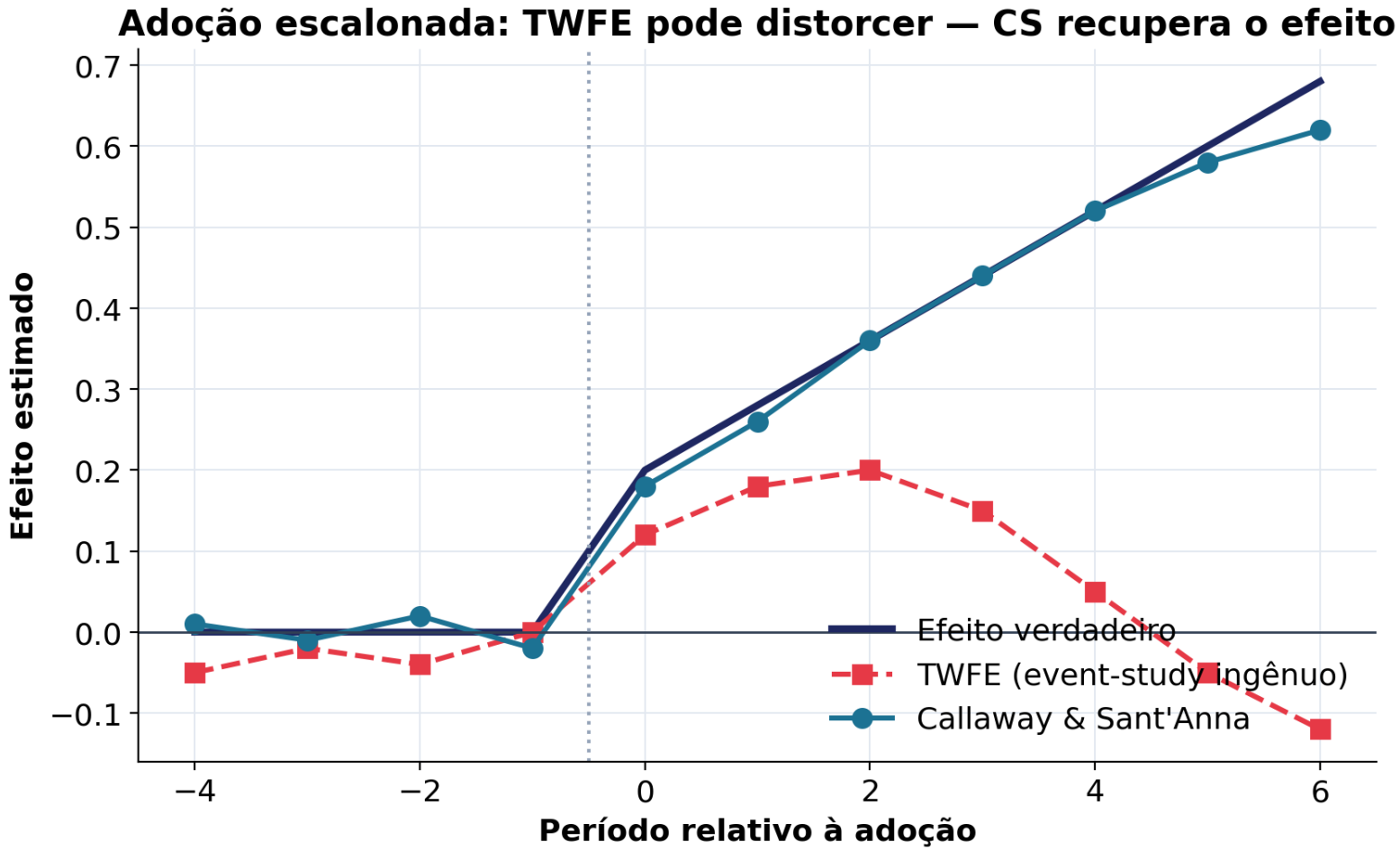
Um 'bloco' de efeito para cada coorte e cada período

$$ATT(g, t) = E[\, Y_t(g) - Y_t(0) \mid G = g \,]$$

- **g (grupo)** Período em que o tratamento começou para a coorte — geralmente um ano.
- **t (tempo)** Período em que medimos o efeito (t ≥ g).
- **Ideia** Em vez de um único δ, estimamos muitos parâmetros — um para cada (g, t).
- **Agregamos depois** Combinamos os ATT(g, t) em estatísticas interpretáveis (dinâmicas, por coorte, gerais).

TWFE × Callaway & Sant'Anna — simulação ilustrativa

Event study ingênuo pode distorcer a dinâmica do efeito



› **Linha preta** Efeito verdadeiro simulado — cresce com o tempo.

› **TWFE (vermelho)** Inicialmente acompanha, depois desvia — contaminação dos cedo servindo de controle aos tarde.

› **CS (azul)** Recupera a trajetória verdadeira usando comparações válidas.

› **Implicação** Para PP com efeitos que maturam no tempo, usar o estimador correto é crítico.

Quando usar CS — quatro motivações práticas

A heterogeneidade do efeito é o motivo principal para abandonar o TWFE

1

Efeito heterogêneo por coorte de adoção

Quem aderiu antes (early-adopters) tem efeito diferente de quem aderiu depois (late-adopters) — comum em políticas de transferência de renda, vacina, ampliação de cobertura.

2

Efeito heterogêneo no tempo (dinâmica)

O efeito cresce ou se dissipa ao longo do tempo — o ATT médio "esconde" essa trajetória, e o event study agregado pode até inverter de sinal.

3

Curto prazo ≠ longo prazo

Política tem efeito imediato (alívio) e efeito de longo prazo (estrutural) com sinais ou magnitudes diferentes — CS permite separar os horizontes.

4

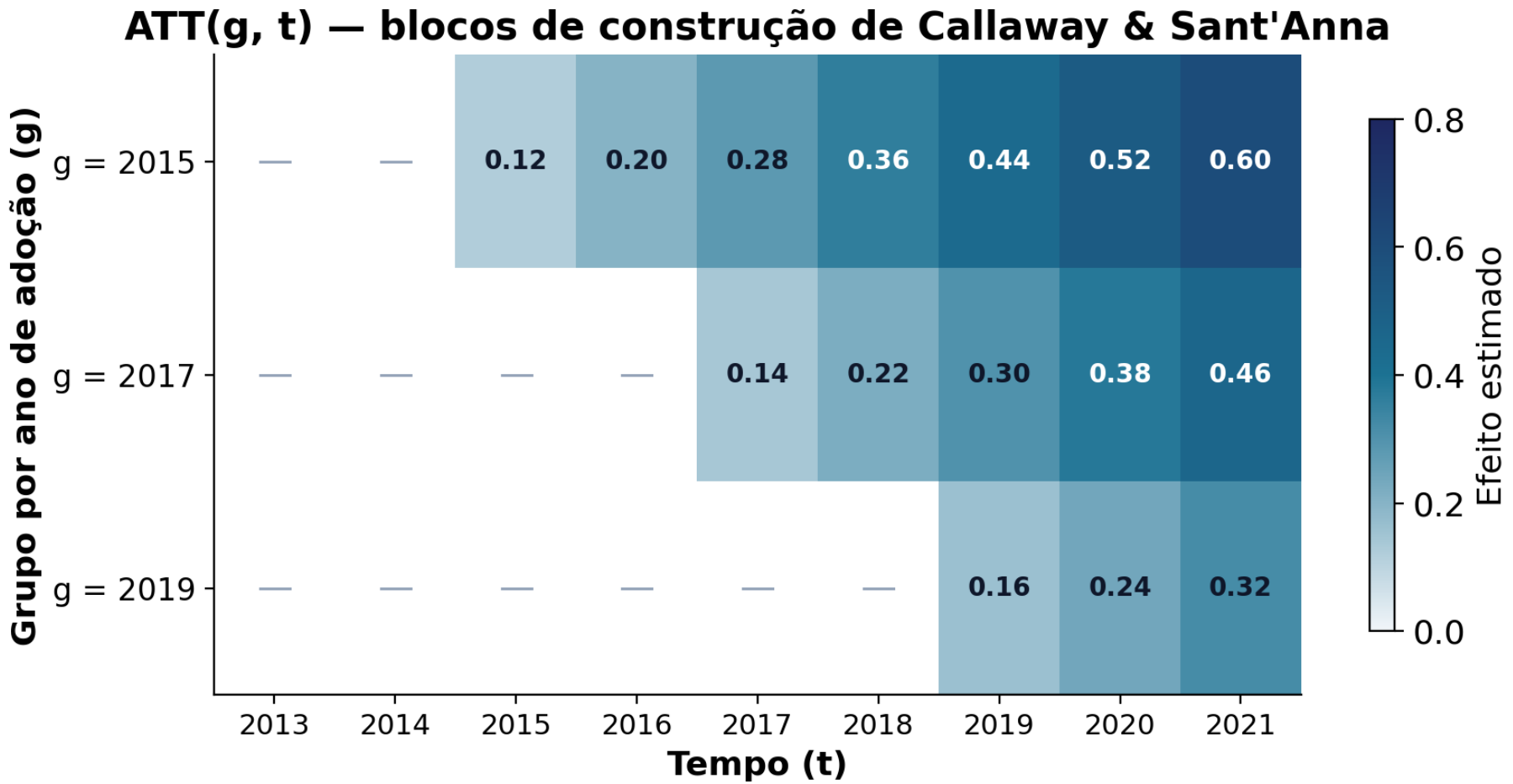
Contexto econômico do momento da adoção

Tratados em recessão respondem diferente de tratados em expansão — CS aceita condicionalismo em X e separa por coorte.

Em todos os quatro cenários o TWFE pode mascarar ou inverter o sinal — CS evita isso identificando ATT(g, t) sem mistura.

ATT(g, t) — blocos de construção

Cada célula é um efeito; os '—' são relações não identificadas



› **Linhas** Uma por coorte g (ano de adoção).

› **Colunas** Períodos calendário t.

› **Pré-tratamento** Usados como placebo.

› **Pós-tratamento** Efeitos dinâmicos por coorte.

Escolha do grupo de comparação

Never-treated vs. not-yet-treated

CS: dois grupos de comparação válidos

Nunca tratados (never-treated)

- Nunca recebem o tratamento, em nenhum período
- Exige que esse grupo exista na amostra
- Dá a identificação mais limpa
- Pode ser pequeno ou pouco representativo

Ainda não tratados (not-yet-treated)

- Ainda não tratados no período t — serão tratados depois
- Aproveita quase toda a amostra disponível
- Permite usar uma janela de antecipação
- Exige a hipótese de não antecipação

- **Na prática** Use not-yet-treated quando o grupo nunca-tratado for pequeno; use never-treated quando for representativo.
- **Reporte ambos** Comparar os dois grupos de comparação é uma checagem de robustez padrão.

Identificação — suposições

Quatro suposições — irreversibilidade, tendências paralelas condicionais, não antecipação e overlap

1. Tratamento irreversível (absorvente)

$D_{it} = 1 \implies D_{is} = 1$ para todo $s > t$ — uma vez tratada, sempre tratada

2. Conditional Parallel Trends (pós)

$E[\Delta Y_t(0) \mid X, G=g] = E[\Delta Y_t(0) \mid X, C]$ para cada g, $t \geq g$

3. Não antecipação

$Y_t(g) = Y_t(0)$ para $t < g$ (opcionalmente admite antecipação limitada)

4. Overlap (suporte comum)

Existe suporte comum de X entre a coorte g e o grupo de comparação

Estimação — Doubly Robust

A recomendação prática dos autores

- **Combina OR + IPW** Modele Y e modele a adoção condicionada a X.
- **Consistente se um acertar** Robustez a má especificação de apenas um componente.
- **Implementação** Versões em R (`did`) e Stata (`csdid`) calculam automaticamente.
- **ATT(g, t) DR** Mesma estrutura para cada (g, t); variância analítica disponível.
- **Escolha de X** Pré-tratamento, impactando Y e/ou adoção; evite colliders e mediadores.

Por que DR?

Raramente sabemos se o modelo do resultado ou o da seleção está correto.

O DR dá duas chances de acertar: se qualquer um dos dois estiver bem especificado, o estimador é consistente.

Callaway & Sant'Anna — o estimador IPW em fórmula

A receita explícita para identificar ATT(g, t) a partir de propensity score

Identificação — versão IPW (Abadie, adaptada por CS)

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{p(X)\cdot C/(1-p(X))}{E[p(X)\cdot C/(1-p(X))]} \right) \cdot \left(Y_t - Y_{g-1} \right) \right]$$

diferença entre a média do grupo tratado g e a média ponderada dos controles, repesados pelo propensity score

Lendo cada peça

G_g indicador da coorte tratada no período g (vale 1 se i pertence à coorte, 0 se não)

C indicador de controle (never-treated ou not-yet-treated) — nunca usa a coorte já tratada como controle

$p(X)$ propensity score generalizado: P(Gg = 1 | X, Gg + C = 1) — chance de pertencer à coorte g dado o perfil X

$Y_t - Y_{g-1}$ variação do resultado entre o último período pré-tratamento (g−1) e t — a "longa diferença"

Três sabores de estimador IPW (acima) · Outcome Regression (modela E[Y(0)|X]) · **Doubly Robust** (combina os dois). Sant'Anna recomenda DR como default.

Como o IPW de CS pondera — leitura visual

A fórmula é apenas uma diferença ponderada — cada lado tem um peso natural

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\hat{p}(X)\cdot C}{(1-\hat{p}(X))\cdot E[\ldots]} \right) \cdot (Y_t - Y_{g-1}) \right]$$

Para uma unidade i da coorte g (G=1)

Peso = $1 / Pr(G=g)$ → cada tratado contribui com sua trajetória $\Delta Y = Y_t - Y_{g-1}$.

Interpretação: a média entre os tratados é simples — todos têm o mesmo peso (uniforme dentro da coorte).

Para uma unidade i do controle (C=1)

Peso = $\hat{p}(X) / (1 - \hat{p}(X))$ → controles parecidos com tratado (alto $\hat{p}$) pesam mais, controles "muito diferentes" pesam pouco.

Interpretação: repesagem reconstrói uma "população sintética" de controles com a mesma distribuição de X dos tratados.

A receita em três passos

1. Estimar $\hat{p}(X)$ com probit/logit nos controles (C=1) e na coorte g.
2. Calcular o peso de cada controle: $\hat{p}(X) / (1-\hat{p}(X))$. Aparar $\hat{p}$ extremos.
3. Tomar a diferença ponderada entre média do ΔY da coorte g e média ponderada do ΔY dos controles.

Inferência — bootstrap multiplier

Bandas uniformes para não induzir ao erro

- **Problema de múltiplos testes** Event study tem muitos coeficientes; cada IC a 95% ignora a dependência.
- **Bandas pontuais × uniformes** Pontuais: cobrem cada coeficiente com 95%. Uniformes: cobrem a curva toda com 95%.
- **Bootstrap multiplier** Método rápido proposto por CS (2021): preserva estrutura de painel, clusterizado em i.
- **Leitura prudente** Em gráficos de event study, reporte bandas uniformes para evitar 'pescar' significâncias.
- **Sensibilidade** Rode com clusters diferentes e com n de bootstrap ≥ 999 para estabilizar IC.

Bootstrap multiplier — mecânica e bandas uniformes

Como CS produz inferência válida sem assumir formas paramétricas

Por que bootstrap em CS

Os ATT(g, t) são correlacionados — compartilham covariáveis, controles e o mesmo painel. A variância analítica seria complicada. O bootstrap captura a estrutura de dependência: **reamostra unidades inteiras (cluster por i)**, preservando a trajetória temporal de cada unidade.

Bootstrap multiplier (Mammen, 1993)

Em vez de reamostrar com reposição, multiplica cada unidade por um peso aleatório (média 0, variância 1). Roda o estimador. Repete $B \geq 999$ vezes. Cada repetição produz uma curva de ATT(g, t) — a dispersão dessas curvas é a variância amostral.

Vantagem — muito mais rápido que o bootstrap empírico tradicional e mantém validade quando há heterocedasticidade.

Bandas pontuais × uniformes — por que importa

Pontuais (95% para cada τ): se você olhar um event study com 10 leads e 10 lags e 95% de cobertura **por ponto**, em média **1 a cada 20 pontos** vai "estourar" o intervalo só por acaso — leitura falsamente significativa.

Uniformes (95% para a curva inteira): bandas calibradas pelo bootstrap multiplier para que, em 95% das amostras, **NENHUM** ponto da curva fique fora. Bandas mais largas, mas leitura honesta.

Para uso prático

No R, `att_gt(..., bstrap = TRUE, biters = 999, cband = TRUE)`. No Stata, `csdid …, wboot rseed(.)` com **cband** nos plots.

Bootstrap multiplier — o algoritmo passo a passo

Por dentro do método de inferência do CS — da função de influência à banda uniforme

1 · Função de influência

Cada ATT(g, t) é uma média das contribuições individuais ψ_i de cada unidade. Calcular essas contribuições uma única vez é o que torna o bootstrap rápido — não se reestima o modelo a cada repetição.

2 · Sorteie os pesos

Para cada unidade i, sorteia-se um peso aleatório V_i de média 0 e variância 1 (Mammen ou Rademacher). Um único peso por unidade, aplicado a toda a sua trajetória — isso preserva a correlação serial (cluster por unidade).

3 · Recompute o estimador

Perturbe as contribuições pelos pesos sorteados e recalcule: cada repetição devolve uma curva inteira de ATT(g, t).

4 · Repita B ≥ 999 vezes

A dispersão das B curvas é a própria distribuição amostral do estimador — sem precisar da fórmula analítica da variância.

5 · Construa a banda uniforme

Calibre a banda para que, em 95% das amostras-bootstrap, nenhum ponto da curva fique fora ao mesmo tempo — mais larga que a pontual, mas honesta com o problema de múltiplos testes.

Os quatro tipos de agregação em CS — fórmulas

Cada ATT(g, t) é um tijolo; a agregação escolhe quais tijolos somar e com que peso

ATT geral (overall)

$$ATT = \Sigma_{g,t \geq g}\ w(g,t) \cdot ATT(g, t)$$

Média ponderada sobre todas as células viáveis. Resumo único — útil para reports executivos. Esconde heterogeneidade.

Dinâmico ATT(e) — estudo de evento

$$ATT(e) = \Sigma_g\ w(g) \cdot ATT(g, g+e)$$

Efeito **e períodos após a adoção**, somado sobre coortes. Mostra trajetória — adoção→efeito imediato→maturação.

Por coorte ATT(g)

$$ATT(g) = (1/(T-g+1))\ \Sigma_{t \geq g}\ ATT(g, t)$$

Efeito médio para quem adotou em g. Útil para comparar **early- vs late-adopters** (a coorte é fonte de heterogeneidade).

Por calendário ATT(t)

$$ATT(t) = \Sigma_{g \leq t}\ w(g) \cdot ATT(g, t)$$

Efeito em cada ano t — útil para narrativas históricas (ex.: "em 2018 o efeito médio foi X").

Pesos w(·, ·) são sempre **não-negativos** (frações de tamanho de coorte) — diferente do TWFE, nada de pesos negativos.

Quatro agregações — como cada uma funciona

Cada agregação é uma média dos mesmos ATT(g, t) — muda apenas quais células somar e como organizá-las

1 · ATT geral (overall)

Soma todas as células pós-tratamento num número único, ponderado pelo tamanho de cada coorte. É a manchete executiva — mas achata toda a heterogeneidade do efeito.

2 · Dinâmico — ATT(e)

Fixa o tempo desde a adoção (e = 0 efeito imediato, e = 1 um ano depois…) e soma sobre as coortes. É a trajetória do efeito — vira o gráfico de event study.

3 · Por coorte — ATT(g)

Fixa a coorte g e tira a média dos seus efeitos ao longo dos períodos pós-adoção. Responde “qual foi o efeito de quem adotou em g?” — compara early- × late-adopters.

4 · Por calendário — ATT(t)

Fixa o ano-calendário t e soma sobre as coortes já tratadas nele. Útil para narrativas históricas e para cruzar o efeito com o contexto macroeconômico do ano.

Em comum. Toda agregação é uma média ponderada dos mesmos ATT(g, t), sempre com pesos não-negativos (frações de tamanho de coorte) — nada de pesos negativos como no TWFE.

Transparência. Mostre os ATT(g, t) brutos antes de agregar — toda agregação esconde alguma heterogeneidade.

Event study no CS — short gap × long difference

O default do pacote pode atrapalhar a interpretação do gráfico pré-tratamento

Short gap (rolling)

Baseline muda em cada par

$$\hat{\delta}_{t-3} = (E[Y|D=1,t-3]-E[Y|D=1,t-2]) - (E[Y|D=0,t-3]-E[Y|D=0,t-2])$$

Default no R **did** e no Stata **csdid** — pode dar coeficientes pré que não se comparam ao gráfico clássico do event study.

Long difference (universal)

Baseline fixo em t−1

$$\hat{\delta}_{t-3} = (E[Y|D=1,t-3]-E[Y|D=1,t-1]) - (E[Y|D=0,t-3]-E[Y|D=0,t-1])$$

Use `base_period="universal"` no R ou `long2` no Stata — coeficientes pré comparáveis ao event study TWFE clássico.

Recomendação prática (Roth, 2024)

Use **long difference** para tornar o event study de CS comparável ao TWFE: as estimativas pré ficam alinhadas ao "zero antes, efeito depois" que a plateia espera ver.

Pipeline analítico

Sequência prática para aplicar CS a uma avaliação de PP

- **Iteração** Na prática, itere entre 2 e 3 ao refinar escolha de X e aparamento.
- **Robustez** Reporte always/never-treated comparisons e varie suposições de antecipação.
- **Transparência** Mostre ATT(g, t) bruto ANTES da agregação — a agregação esconde heterogeneidade.

CS × outros estimadores modernos

Panorama das alternativas a TWFE sob adoção escalonada

Callaway & Sant'Anna (2021)

ATT(g, t) com DR, never- ou not-yet-treated, agregações

Recomendado como default para adoção absorvente

Sun & Abraham (2021)

Event-study com interações cohort × relative time

Muito similar a CS; robusto em especificações de event study

de Chaisemartin & D'Haultfœuille (2020+)

DiD multiperíodo e tratamentos não-absorventes

Essencial quando o tratamento liga/desliga

Borusyak, Jaravel & Spiess (2024)

Imputação de contrafactual via modelo de efeitos fixos

Eficiente sob homogeneidade; complementar ao CS

Sun & Abraham e de Chaisemartin & D'Haultfœuille

Dois estimadores robustos para event study sob adoção escalonada — e o caso do tratamento reversível

Sun & Abraham (2021) — estimador IW

› **Receita em 3 passos.** uma regressão DD com indicadores de coorte × tempo relativo estima cada CATT(e, l); os pesos são a participação amostral de cada coorte; por fim, a média ponderada.

› **Grupo de comparação.** usa a última coorte tratada ou os never-treated como controle; descarta os always-treated.

› **Inferência.** consistente sob tendências paralelas e não antecipação; assintoticamente normal — variância analítica, sem depender de bootstrap como o CS.

› **Software.** eventstudyinteract (Stata) · fixest::sunab() (R).

de Chaisemartin & D'Haultfœuille — DIDM

› **O diferencial.** é o único do grupo que admite tratamento que liga e desliga (reversível).

› **Como estima.** mede o efeito em torno de cada transição: “joiners” (DID+, entram no tratamento) e “leavers” (DID−, saem).

› **Estimador DIDM.** média ponderada de DID+ e DID−; evita os pesos negativos do TWFE.

› **Diagnóstico.** os autores recomendam reportar a fração de ATT(g, t) com peso negativo. Software: did_multiplegt.

Em comum. Os dois “pulam” os forbidden contrasts — nunca usam coortes já tratadas como controle. Sob timing diferencial, SA e dCDH são equivalentes ao CS a menos da escolha de pesos.

Estimador de imputação — Borusyak, Jaravel & Spiess

Modela diretamente o contrafactual Y(0) com as observações não tratadas e o extrapola para as tratadas

1 · A ideia

Toda inferência causal é imputação de contrafactuais (Imbens & Rubin) — o BJS apenas torna isso explícito: modela-se o resultado potencial Y(0) e o estende às unidades tratadas.

2 · Receita em 3 passos

(1) estima o modelo de Y(0) por OLS só com as observações não tratadas — efeitos fixos de unidade e de tempo; (2) imputa $\hat{Y}(0)$ para cada célula tratada; (3) calcula $\delta_{st} = Y_{st} - \hat{Y}_{st}(0)$ e agrega em somas ponderadas.

3 · Por que funciona

A PTA é uma afirmação sobre Y(0). Estimando os efeitos fixos só com quem não foi tratado, recupera-se Y(0) sem contaminação do efeito do tratamento — e daí se extrapola para as células tratadas.

4 · Vantagens e custo

Usa todas as não tratadas (pré dos tratados + nunca/ainda-não tratados) → mais potência com mais períodos pré; eficiente sob efeitos homogêneos e rápido (continua OLS). Custo: sem suporte comum, impõe forma funcional.

“Feasible ATT” — o CS não estima todos os ATT(g, t)

Algumas células (g, t) não têm grupo de comparação válido — o CS estima a média só sobre as viáveis

Coorte g \ Ano t	1986	1992	1998	2004	2009
1986	✓	✓	✓	✓	✓
1992	pré	✓	✓	✓	✓
1998	pré	pré	✓	✓	✓
2004 (última)	pré	pré	pré	—	—

✓ ATT(g, t) estimável (pós-tratamento) **pré** placebo pré-tratamento — não identificado (sem grupo de comparação)

Por que algumas células somem

› Cada ATT(g, t) pós só é estimável se existir grupo de comparação naquele t — never-treated ou not-yet-treated.

› A última coorte (2004) não tem ninguém “ainda não tratado” depois dela. Sem never-treated, seus efeitos pós ficam **não identificados**.

› **Feasible ATT** = média ponderada só sobre os ATT(g, t) que o CS consegue estimar (as células ✓).

› Pode diferir do ATT geral verdadeiro: se a coorte excluída tem efeito atípico, o feasible ATT cobre só parte da população tratada.

No exemplo de Cunningham (adoção escalonada simulada). O ATT geral verdadeiro ≈ 82, mas o feasible ATT verdadeiro ≈ 68 — e o CS estima ≈ 64. A queda de 82 para 68 é o preço de não haver grupo never-treated para a última coorte.

Em PP: sempre que possível, preserve um grupo never-treated (municípios elegíveis não atendidos) — isso amplia o conjunto de ATT(g, t) estimáveis e aproxima o feasible ATT do efeito completo.

HonestDiD — análise de sensibilidade da PTA

A suposição-rainha não é testável — então medimos o quanto o resultado depende dela

1 · O problema

Tendências paralelas não é uma suposição testável. Pré-tendências ≈ 0 são evidência a favor, nunca prova — e nada garante que a PTA continue valendo no pós-tratamento.

2 · A virada de chave

Rambachan & Roth (2023) trocam o teste "passou / não passou" por uma pergunta de robustez — não rejeitam a PTA, quantificam o quanto o resultado depende dela.

3 · Como funciona

Toma-se a maior violação observada nas pré-tendências e supõe-se que a PTA quebra nessa magnitude no pós. Procura-se o breakdown: o quanto a violação precisa crescer até o IC do ATT cobrir o zero.

4 · Duas famílias de restrição

Magnitude relativa ($\bar{M}$) — o viés pós é no máximo $\bar{M}$ vezes o pior gap pré. Suavidade — a violação só muda de forma gradual entre períodos. Reporta-se o ATT para uma faixa de hipóteses.

HonestDiD — as duas famílias de restrição

Como o HonestDiD define o conjunto de violações plausíveis da PTA — e o que reportar

Magnitude relativa (M̄)

› **Definição.** o desvio de tendências paralelas no pós é, no máximo, M̄ vezes o maior desvio observado nas pré-tendências.

› **A escala.** M̄ = 1 — a violação pós não é pior que a pior violação pré; M̄ = 2 admite o dobro, e assim por diante.

› **Intuição.** usa os próprios pré-trends como régua — não exige escolher uma escala externa, arbitrária.

› **Quando usar.** default quando há pré-períodos suficientes para medir a "pior" violação observada.

Suavidade (M)

› **Definição.** restringe o quanto a violação pode mudar de forma entre períodos consecutivos (segundas diferenças).

› **O parâmetro.** M = 0 — a violação extrapola o pré-trend em linha reta; M > 0 admite curvatura, até M por período.

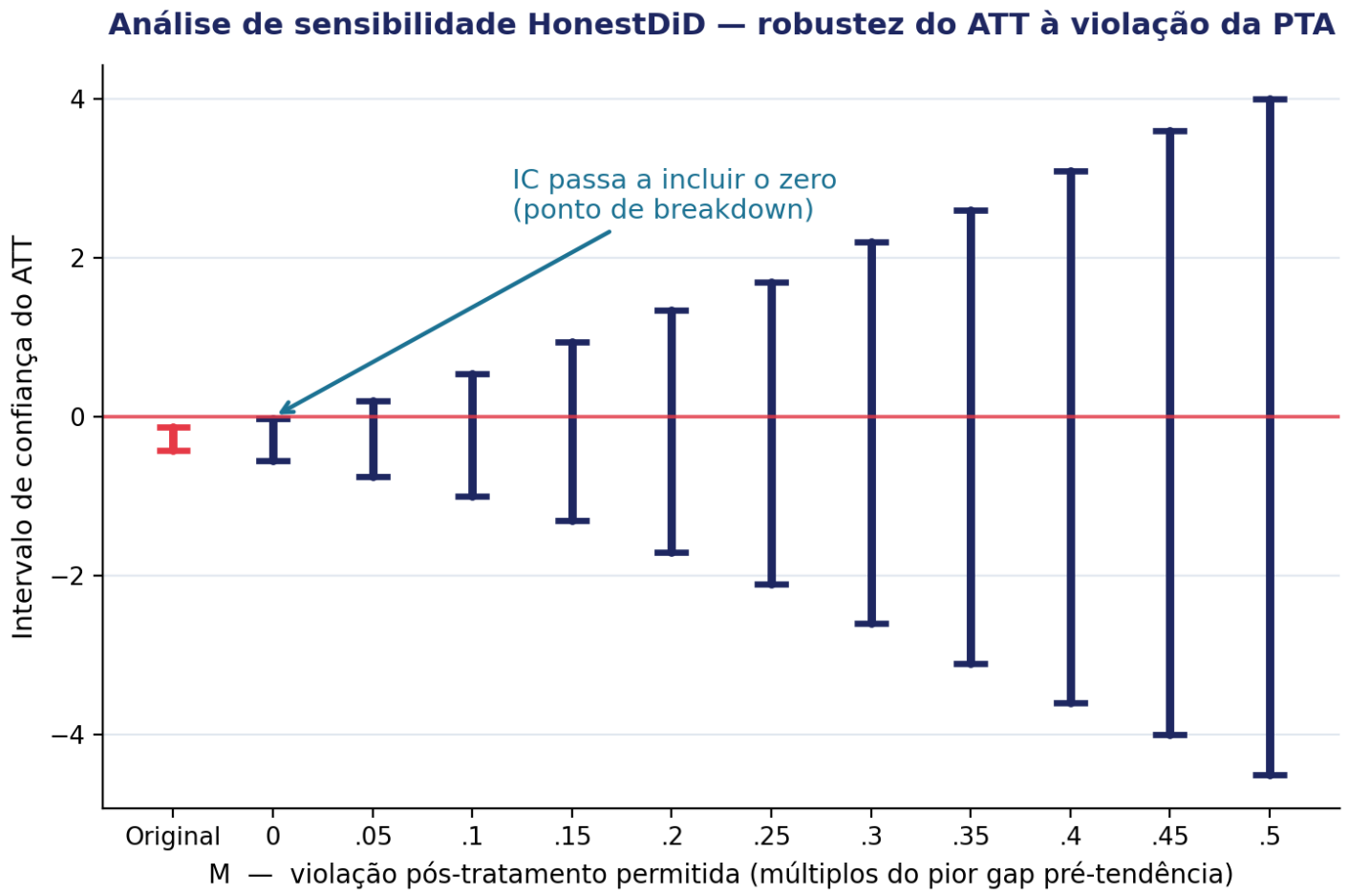
› **Intuição.** tendências diferenciais costumam evoluir de modo gradual, não em saltos abruptos.

› **Quando usar.** útil quando se espera uma violação suave e não há um pré-período longo para medir o pior gap.

O que reportar. O resultado é um intervalo, não um ponto — ao admitir violação da PTA, o ATT passa a ser parcialmente identificado. Reporte o breakdown: o menor M̄ (ou M) em que o IC robusto passa a incluir o zero.

HonestDiD + Callaway & Sant'Anna na prática

O CS produz exatamente a entrada de que o HonestDiD precisa — os dois se encaixam



Encaixe natural

O HonestDiD recebe os coeficientes de um event study mais a matriz de variância. O CS, agregado na forma dinâmica ATT(e), entrega exatamente isso.

Implementação

O pacote honestdid (R) tem um wrapper que lê a saída do pacote did (CS) diretamente. Em Stata, honestdid roda após o csdid.

Lendo o gráfico

"Original" é o IC do CS. Conforme M cresce (violação permitida), o IC alarga; o breakdown é o M em que ele passa a incluir o zero.

Em política pública

Reportar o M de breakdown é honesto — vira um número que o gestor entende: "o efeito se sustenta, a não ser que a violação passe de X".

Encerramento

Conclusão e próximos passos

Síntese, checklist e referências

Mensagens centrais do curso

Cinco ideias para levar para a sua próxima avaliação

1 **DiD é uma estratégia de desenho, não uma fórmula**

Pense primeiro em contrafactual e grupo de comparação; depois na regressão.

2 **A suposição-rainha é tendências paralelas**

Não é testável diretamente, mas é diagnosticada por pré-tendências e falsificações.

3 **Em adoção escalonada, TWFE pode mentir**

Efeitos heterogêneos no tempo contaminam o estimador. Sempre diagnostique.

4 **Callaway & Sant'Anna é o default moderno**

ATT(g, t) + DR + bootstrap é o pacote recomendado; agregue com intenção.

5 **Comunique com event studies bem interpretados**

Pré-trends, efeito imediato, dinâmica — e bandas uniformes, não pontuais.

Checklist para avaliação de PP com DiD

Antes de publicar ou apresentar a um conselho

Desenho e identificação

- Definição clara do tratamento (binário, absorvente, timing)
- Grupo de comparação justificado (never- vs. not-yet-treated)
- Covariáveis pré-tratamento documentadas (DAG mental)
- Teste de pré-tendências (event study com bandas uniformes)
- Antecipação considerada (janela e diagnóstico)

Robustez e comunicação

- Balanceamento de X reportado (SMD, overlap)
- Resultados por coorte ANTES da agregação global
- Robustez: troca do grupo de comparação, aparamento, clusters
- SUTVA discutido (spillovers, buffers, fronteiras)
- Interpretação do efeito em unidades de política (R$, pp, vidas)

Limitações, extensões e quando NÃO usar DiD

Honestidade sobre os limites do método

› **Tratamentos contínuos** O arcabouço clássico é para binários; há extensões recentes.

› **Tratamento não absorvente** Ligar/desligar exige de Chaisemartin & D'Haultfœuille.

› **Poucas unidades tratadas** DiD pede variação entre grupos; considere synthetic control.

› **Ausência de paralelismo plausível** Considere RDD, matching dinâmico ou desenho instrumental.

› **Efeitos de equilíbrio geral** DiD mede efeito local — alterações estruturais pedem modelos.

› **Dados composicionais** Se o painel é desbalanceado por entrada/saída, cuidado com viés de composição.

Lembrete final

Um DiD impecável em execução mas com suposições frágeis é pior que um desenho simples com escrutínio honesto.

A transparência sobre suposições é o que distingue uma boa avaliação de política pública.

Referências principais

Leituras para aprofundamento

[1] Callaway, B. & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2), 200-230.

[2] Cunningham, S. (2021). Causal Inference: The Mixtape. Yale University Press. Capítulos sobre DiD.

[3] Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2), 254-277.

[4] Sun, L. & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2), 175-199.

[5] de Chaisemartin, C. & D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. American Economic Review, 110(9), 2964-2996.

[6] Borusyak, K., Jaravel, X. & Spiess, J. (2024). Revisiting event study designs: robust and efficient estimation. Review of Economic Studies.

[7] Roth, J., Sant'Anna, P. H. C., Bilinski, A. & Poe, J. (2023). What's trending in difference-in-differences? Journal of Econometrics.

[8] Sant'Anna, P. H. C. & Zhao, J. (2020). Doubly robust difference-in-differences estimators. Journal of Econometrics, 219(1), 101-122.

Obrigada!

Perguntas, críticas, aplicações?

Materiais de referência: Cunningham (Mixtape Sessions) — github.com/Mixtape-Sessions

Anexo

Triple differences (DDD)

Um desenho para quando as tendências paralelas não se sustentam

Triple differences — o que é e quando usar

Não é um teste placebo — é um desenho próprio, para quando o DiD comum é enviesado

1 · Um desenho próprio

O DDD não é um exercício de falsificação (confusão comum) — é uma estratégia de identificação completa, com hipóteses próprias.

2 · Quando usar

Use DDD quando o DiD 2×2 é enviesado — ou seja, quando você suspeita que as tendências paralelas não valem para o seu par tratado × controle.

3 · A ideia

Monte dois DiD: um “verdadeiro” (no grupo afetado pela política) e um “placebo” (num grupo que não deveria responder). Se os dois carregam o mesmo viés, a diferença entre eles o cancela.

4 · A troca de hipótese

O DDD não elimina suposições — substitui a tendência paralela pela hipótese de viés paralelo: o viés do DiD #1 é igual ao viés do DiD #2.

DDD — a lógica do viés paralelo

O exemplo de Gruber (1994/95) — mandatos de licença-maternidade e salários

1 · O contexto (Gruber)

Gruber avalia o efeito de mandatos de licença-maternidade sobre salários. Grupo de risco: mulheres casadas de 20–40 anos. Grupo placebo: homens solteiros e mulheres mais velhas — não afetados pelo mandato.

2 · Dois DiD

DiD #1 (verdadeiro): estados que adotaram × não adotaram, no grupo de risco. DiD #2 (placebo): a mesma comparação de estados, mas no grupo placebo.

3 · A subtração

$\delta(\text{DDD}) = \delta(\text{DiD verdadeiro}) - \delta(\text{DiD placebo})$. Se os dois DiD têm o mesmo viés, a subtração o cancela — e sobra o ATT, mesmo com a PTA violada.

4 · A hipótese e o mito

Substitui a PTA pelo viés paralelo: viés do DiD #1 = viés do DiD #2. Mito a desfazer: não é preciso que o DiD placebo dê zero. (No Gruber 1995, o DDD ≈ −0,054 ficou quase igual ao DiD ≈ −0,062.)

DDD — por que o viés se cancela

Quando o viés é o mesmo nos dois DiD, a terceira diferença o elimina e sobra o ATT

1 · DiD verdadeiro = ATT + viés

No grupo afetado pela política, o DiD 2×2 entrega δ(verdadeiro) = ATT + (tendência dos tratados − tendência dos controles). Com a PTA violada, o segundo termo ≠ 0 — é o viés.

2 · DiD placebo = só o viés

No grupo placebo, a política não tem efeito (não há ATT). O mesmo DiD 2×2 entrega δ(placebo) = (tendência dos tratados − tendência dos controles) — apenas o viés.

3 · A terceira diferença

Subtraindo um do outro: δ(DDD) = δ(verdadeiro) − δ(placebo) = (ATT + viés) − viés = ATT. O viés é eliminado.

4 · A condição

O cancelamento só vale se o viés for idêntico nos dois DiD — a hipótese de viés paralelo. É essa, e não a tendência paralela, a suposição de identificação do DDD.

DDD em regressão

A dupla diferença vira tripla — com um termo de interação tripla

1 · A regressão de tripla diferença

$$Y(i, j, t) = \alpha + \beta_2\tau + \beta_3\delta + \beta_4 D + \beta_5(\delta\times\tau) + \beta_6(\tau\times D) + \beta_7(\delta\times D) + \beta_8(\delta\times\tau\times D) + \varepsilon$$

2 · Três dimensões

Os dados são empilhados em três eixos: tempo (τ — antes/depois), grupo (δ — risco/placebo) e estado (D — tratado/controle).

3 · β_8 é o ATT

O coeficiente da interação tripla (δ×τ×D) estima o efeito causal da política — é o análogo do δ da interação no DiD 2×2.

4 · O papel das interações duplas

As duplas (β_5, β_6, β_7) absorvem os viéses que contaminariam um DiD simples — é exatamente isso que a terceira diferença acrescenta.

Anexo

Sun & Abraham

O estimador interaction-weighted (IW) — em detalhe

Sun & Abraham — o estimador interaction-weighted

De onde vem o IW e como ele estima os efeitos por coorte e tempo relativo

1 · A motivação

No event study TWFE, o coeficiente de cada lead/lag é “contaminado” por informação de outros leads e lags. Sun & Abraham (2021) provam essa decomposição e propõem uma solução — próxima do CS, mas descoberta de forma independente.

2 · Passo 1 — CATT(e, l)

Uma regressão DD com indicadores de coorte (e) × tempo relativo (l) estima o CATT(e, l): o efeito médio da coorte e, l períodos após o tratamento. Descartam-se as unidades always-treated.

3 · Passo 2 — os pesos

Os pesos são a participação amostral de cada coorte nos períodos relevantes — $\Pr(E_i = e)$. Coortes maiores pesam mais na média final.

4 · Passo 3 — média ponderada

O estimador IW é a média dos CATT(e, l) ponderada pelas participações de coorte — daí o nome “interaction-weighted”.

Sun & Abraham — identificação e propriedades

O grupo de comparação, a garantia de não viés e por que dispensa bootstrap

1 · Grupo de comparação

O IW usa a última coorte tratada — ou os never-treated — como controle, nunca uma coorte já tratada. É isso que evita os forbidden contrasts do TWFE.

2 · Estimador DD do CATT

Cada CATT(e, l) é uma dupla diferença: a variação do resultado da coorte e, de um pré-período até e+l, menos a variação das coortes de controle no mesmo intervalo.

3 · Garantia de não viés

Sob tendências paralelas e não antecipação em todos os pré-períodos, o estimador DD é não viesado para o CATT(e, l) — qualquer pré-período e qualquer grupo de controle não vazio servem (Proposição 5).

4 · Inferência e software

Sun & Abraham derivam a variância assintótica analiticamente — não dependem de bootstrap como o CS. Software: eventstudyinteract (Stata) · fixest::sunab() (R).