Synthetic Control Method & Synthetic Difference-in-Differences

Prof. Daniel de Abreu Pereira Uhr

Estrutura

- Introdução ao Método de Controle Sintético e ao Método de Diferenças em Diferenças Sintético
- Synthetic Control Method - SCM
  - Formalizando o SCM
  - Hipóteses de identificação do SCM
  - Procedimento Geral do SCM
  - Artigo de Abadie et al (2010)
- Synthetic Difference-in-Differences - SDD
  - Requerimentos do SDD
  - Formalizando o SDD
  - Condicionando em Covariáveis
  - Desenho de adoção escalonada (*The Staggered Adoption Design*)
  - Aplicação no Python

Referências

Principais

- Abadie, Alberto, and Javier Gardeazabal. “The economic costs of conflict: a case study of the basque country”. American Economic Review, 93 (1): 113-132. 2003
- Alberto Abadie, Alexis Diamond & Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: estimating the effect of california’s tobacco control program, Journal of the american statistical association, 105:490, 493-505, 2010.
- Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. American Journal of Political Science, 59(2), 495–510. https://doi.org/10.1111/ajps.12116
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. "Synthetic Difference-in-Differences." American Economic Review, 111 (12): 4088-4118. DOI: 10.1257/aer.20190159
- Clément deChaisemartin, Xavier d’Haultfoeuille. (2022) Difference-in-Differences Estimators of Intertemporal Treatment Effects. hal-03873903
- Roth et al. (2023) What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. https://www.jonathandroth.com/assets/files/DiD_Review_Paper.pdf

Complementares

- Masa Asami - notebook/ReproductionExperiment_CaliforniaSmoking.ipynb
- Masa Asami (gitHub) https://github.com/MasaAsami/pysynthdid/blob/main/notebook/ReproductionExperime
- David Hirshberg Notes: https://davidahirshberg.bitbucket.io/static/synth-did-slides.pdf
- Abadie, A., and J. L'Hour, 2021. A Penalized Synthetic Control Estimator for Disaggregated Data. Journal of the American Statistical Association, 116(536): 1817-1834.
- Ben-Michael, E., Feller, A. and J. Rothstein, 2021. The Augmented Synthetic Control Method. Journal of the American Statistical Association, 116(536): 1789-1803.
- Victor Chernozhukov, Kaspar Wüthrich & Yinchu Zhu (2021) An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls, Journal of the American Statistical Association, 116:536, 1849-1864, DOI: 10.1080/01621459.2021.1920957
- Wiltshire, J.C., 2022. allsynth: (Stacked) Synthetic Control Bias-Correction Utilities for Stata. Working paper.
- Clarke et al (2023) Synthetic Difference In Differences Estimation. https://doi.org/10.48550/arXiv.2301.11859
- Charles F. Manski, John V. Pepper; How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. The Review of Economics and Statistics 2018; 100 (2): 232–244. doi: https://doi.org/10.1162/REST_a_00689
- Rambachan, A. and Roth. J. (2022) A More Credible Approach to Parallel Trends. https://www.jonathandroth.com/assets/files/HonestParallelTrends_Main.pdf
- Alyssa Bilinski, Laura A. Hatfield (2018). Nothing to see here? Non-inferiority approaches to parallel trends and other model assumptions. https://doi.org/10.48550/arXiv.1805.03273
- Andrew Goodman-Bacon (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2021.03.014
- Nikolay Doudchenko, Guido W. Imbens (2016) Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. https://doi.org/10.48550/arXiv.1610.07748
- Bruno Ferman, Cristine Pinto (2021). Synthetic controls with imperfect pretreatment fit. Quantitative Economics. https://doi.org/10.3982/QE1596
- Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." Journal of Economic Literature, 59 (2): 391-425. DOI: 10.1257/jel.20191450

Introdução ao Método de Controle Sintético e ao Método de Diferença em Diferenças Sintético

Um desafio de modelagem para estimar o impacto da exposição à algum evento ou política particular, quando as observações estão disponíveis em um painel ou corte transversal repetido de grupos e tempo é determinar o que teria acontecido com as

unidades expostas se elas não tivessem sido expostas. Caso tal contrafactual seja estimável a partir de dados, a inferência causal pode ser conduzida comparando os resultados em unidades tratadas com aqueles em estados contrafactuais teóricos não tratados, sob a estrutura de resultados potenciais.

Um número substancial de estudos empíricos em economia e nas ciências sociais em geral procura estimar os efeitos nesse cenário usando designs no estilo de diferenças em diferenças (DD). No DD os impactos são inferidos comparando unidades tratadas com unidades de controle, onde são permitidas efeitos fixos individuais invariantes no tempo entre as unidades, bem como tendências gerais comuns. No entanto, para inferência causal requer-se a suposição de **tendências paralelas**, a qual afirma que, na ausência de tratamento, as unidades tratadas teriam seguido caminhos paralelos às unidades não tratadas.

Em muitos casos, **tendências paralelas podem ser uma suposição de modelagem questionável**. Uma solução específica para o desafio tem sido a aplicação de **métodos de controle sintéticos**. Os primeiros trabalhos em controle sintético exploram o cenário de estudos de caso comparativos, onde uma **única unidade tratada é observada** e **se deseja construir um controle sintético** combinado a partir de um número maior de unidades doadoras potenciais (Abadie e Gardeazabal2003; Abadie et al. 2010, 2015). **Esse método busca gerar um único controle sintético** a partir de uma ponderação convexa única de unidades de controle subjacentes, de modo que esse controle sintético seja o mais próximo possível da unidade tratada em resultados de pré-tratamento e determinado potencialmente por outras covariáveis. São gerados pesos ótimos e fixados ao longo do tempo, potencialmente atribuindo peso zero a certas unidades de controle e pesos positivos às outras.

A primeira aparição do Método de Control Sintético foi em um artigo de 2003, onde foi utilizado para estimar o impacto do terrorismo na actividade económica (Abadie e Gardeazabal 2003) . Desde essa publicação, tornou-se muito popular – especialmente após o lançamento de um pacote R e Stata coincidindo com Abadie, Diamond e Hainmueller (2010). O estimador tem sido tão influente que Athey e Imbens (2017) afirmaram que foi **"indiscutivelmente a inovação mais importante na literatura de avaliação de políticas nos últimos 15 anos"**.

Recentemente, uma série de metodologias têm procurado afrouxar a suposição de tendências paralelas. Isso inclui procedimentos nos quais as tendências contrafactuais podem se desviar de forma não linear, levando à identificação parcial (Manski e Pepper, 2018; Rambachan e Roth, 2019), procedimentos flexíveis para controlar adequadamente quaisquer diferenças existentes entre unidades tratadas e de controle (Bilinski e Hatfield 2018), frequentemente baseados apenas em períodos de pré-tratamento (Goodman-Bacon 2021).

A abordagem de Controle Sintético atraiu atenção considerável tanto em aplicações empíricas quanto em extensões teóricas, com avanços recentes, incluindo procedimentos de eliminação de viés (Ben-Michael et al.2021) que podem adicionalmente abrigar várias unidades de tratamento (Abadie e L'Hour 2021), esquemas de ponderação mais flexíveis

ou diferenças fixas constantes entre unidades de controle tratadas e sintéticas (Doudchenko e Imbens, 2016; Ferman e Pinto, 2021).

Arkhangelsky et al.(2021) propõem o estimador **Synthetic Difference-in-Differences** (SDD), que traz pontos fortes do DD e SCM. Como os modelos DD, **o SDID permite que as unidades tratadas e de controle tenham tendências em níveis totalmente diferentes** antes de uma intervenção de interesse. E como o SCM, o SDD procura gerar de forma otimizada uma unidade de controle correspondente que **reduz consideravelmente a necessidade de suposições de tendências paralelas**.

O SDD evita armadilhas comuns do DD e do SCM. Ou seja, uma incapacidade de estimar relações causais se tendências paralelas não forem atendidas em dados agregados no caso DD, e um requisito de que a unidade tratada seja alocada dentro de uma "combinação convexa" das unidades de controle no caso de SC. Arkhangelsky et al. (2021) propõem procedimentos de estimação e inferência, provando formalmente a consistência e a normalidade assintótica do estimador. Além disso, os autores discutem vários pontos importantes aplicados, como seu estimador pode incorporar covariáveis e como seu estimador pode ser aplicado a várias unidades de tratamento e até a várias unidades de tratamento que adotam tratamento em diferentes períodos de tempo.

Synthetic Control Method - SCM

A abordagem de controle sintético foi inicialmente formulada para quando havia apenas um indivíduo afetado pela intervenção numa estrutura de dados em painel (método para “estudo de caso”). A ideia fundamental dessa abordagem é que **uma combinação de unidades geralmente fornece uma comparação melhor para a unidade exposta à intervenção do que uma única outra unidade sozinha (ou, ainda, a média das não tratadas)**.

Então, o **Abadie and Gardeazabal (2003)** e **Abadie et al. (2010)** propuseram um método que pondera o grupo de indivíduos do grupo de controle (chamado de “donor pool” ou “conjunto de doadores”) de modo a **construir uma unidade sintética** mais próxima possível da unidade tratada no período pré-intervenção. Essa estratégia **não requer a hipótese de tendências paralelas**.

Formalizando o SCM

Suponha que observamos $J+1$ indivíduos e apenas o primeiro esteja exposto à intervenção. Os demais indivíduos estão no “donor pool”. Seja Y_{it}^N a variável de resultado para i no tempo t na ausência de intervenção, para unidades $i = 1, \cdots, J+1$ e períodos de tempo $t = 1, \cdots, T$. Considere T_0 o número de períodos pré-intervenção, com $1 \leq T_0 < T$. Seja Y_{it}^I o resultado para i no momento t se a unidade i é exposta à intervenção nos períodos T_{0+1} até T.

Seja $\alpha_{it} = Y_{it}^I - Y_{it}^N$ o **efeito da intervenção**. Então gostaríamos de estimar:

$$\alpha_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

Como Y_{1t}^{I} é observado, precisamos estimar Y_{1t}^{N} (**contrafactual**), o qual é baseado num vetor de covariáveis que não são afetadas pela intervenção. Considere um vetor $(J+1)$ de pesos $W = (w_2, \ldots, w_{J+1})$ tal que $w_j \geq 0$ para $j = 2, \cdots, J+1$ e a soma dos pesos é tal que: $\sum_{j=2}^{J+1} w_j = \mathbf{1}$. O efeito estimado da intervenção em $t \in T_{0+1}, \ldots, T$, pode ser entendido como:

$$\widehat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

O controle sintético fornece **pesos para um estimador não-viesado** de Y_{1t}^{N}. Seja $X_{\mathbf{1}}$ um vetor de características pré-intervenção para i, e X_0 uma matriz (kxJ) que contém as mesmas variáveis para os indivíduos não afetados. O vetor de pesos ótimos, W^*, é escolhido para minimizar a distância entre $X_{\mathbf{1}}$ e $X_{\mathbf{0}}W$, sujeito a $w_2 \geq 0, \ldots, w_{J+1} \geq 0$ e $w_2 + \ldots + w_{J+1} = 1$.

Hipóteses de identificação do SCM

Hipóteses e requerimentos do método

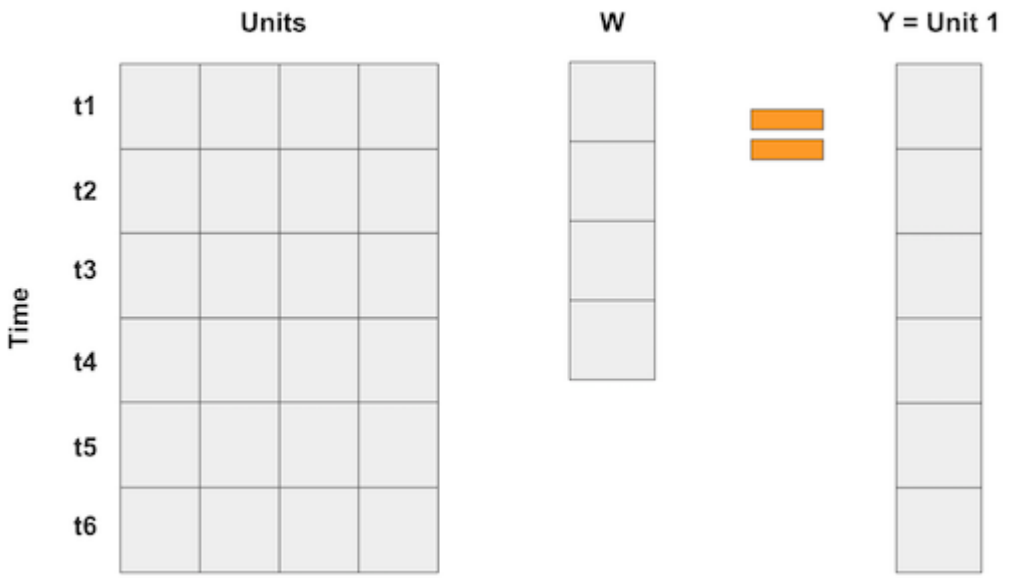
- "Donor Pool" comparável;
- Não interferência: Y_{1t} e Y_{js} não podem afetar-se para qualquer $i \neq j$ e para qualquer t e s;
- Não antecipação: Y_{1t} não é afetado no período pré-tratamento;
- Informação suficiente no período pré-tratamento;
- Informação suficiente no período pós-tratamento.

Condições de convexidade

- Abadie et al (2010): pesos não-negativos e somando um.
- Literatura apresenta questionamentos.
- Chernozhukovet.al.(2021): Os pesos podem ser negativos desde que a soma dos valores absolutos dos pesos seja menor que um. No entanto, a dispersão dos pesos é altamente recomendada para evitar o overfitting.

Controle sintético como uma regressão linear invertida

Para encontrar a combinação de estados que melhor se aproxima da tendência de pré-tratamento da Califórnia, o método de Controle Sintético executa uma regressão horizontal, onde as linhas são os períodos de tempo e as colunas são os estados. Tenta encontrar os pesos que, quando multiplicados pelos estados de controle, melhor se aproximam do estado tratado



Controle Sintético impõe duas restrições:

- Os pesos devem somar 1;
- Os pesos devem ser não negativos;

Combinadas, estas restrições significam que estamos definindo o controle sintético como uma combinação convexa das unidades de controle.

Procedimento Geral do SCM

A análise dos resultados do SCM é predominantemente gráfica.

1. Verificamos se o controle sintético acompanha o indivíduo tratado antes do período de intervenção.
2. Verificamos se há afastamento entre o indivíduo tratado e seu controle sintético no momento da intervenção.
3. Analisamos a robustez do modelo através de testes de placebo.
4. Calcula-se um conjunto de valores do erro quadrático médio previsto (RMSPE) para o período pré e pós-tratamento como a estatística de teste usada para inferência (donor pool). Calcula-se a proporção do RMSPE pós-tratamento e pré-tratamento. Ordena-se essa proporção em ordem decrescente do maior para o maior. Queremos saber se o efeito do tratamento é extremo.

Artigo de Abadie et al (2010)

Contexto do problema de pesquisa: estimar o efeito da tributação dos cigarros sobre o seu consumo. Um lado do argumento diz que os impostos aumentarão o custo dos charutos, o que diminuirá a sua procura. O outro lado argumenta que, uma vez que os cigarros causam dependência, a mudança no seu preço não alterará muito a sua procura. Em termos económicos, diríamos que a procura de cigarros é inelástica em relação ao preço e que um aumento dos impostos é apenas uma forma de aumentar as receitas do governo à custa dos fumadores.

Objeto de pesquisa: Em 1988, a Califórnia aprovou uma famosa Lei de Imposto sobre o Tabaco e Proteção à Saúde, que ficou conhecida como Proposição 99 . "Seu principal efeito é impor um imposto estadual de 25 centavos por maço sobre a venda de cigarros de tabaco na Califórnia, com impostos especiais de consumo aproximadamente equivalentes cobrados de forma semelhante sobre a venda no varejo de outros produtos comerciais de tabaco, como charutos e tabaco de mascar. Restrições adicionais impostas à venda de tabaco incluem a proibição de máquinas de venda automática de cigarros em áreas públicas acessíveis a jovens e a proibição da venda individual de cigarros individuais. A receita gerada pela lei foi destinada a vários programas ambientais e de saúde, e a anúncios antitabaco."

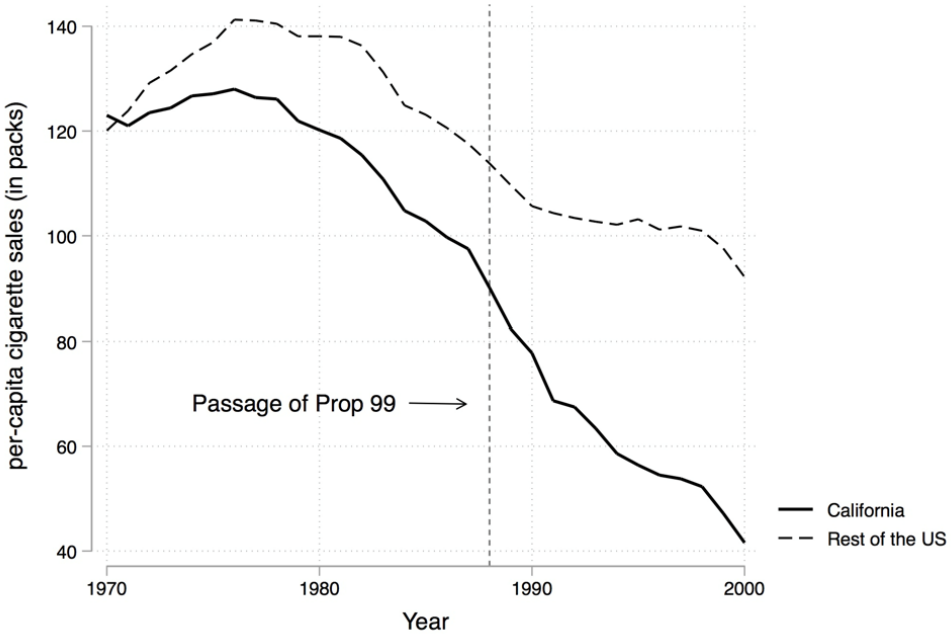
Método e Dados: Para avaliar o efeito do programa, utiliza-se dados sobre as vendas de cigarros (per capita) em vários estados e ao longo de vários anos. No nosso caso, dados do ano de 1970 a 2000 de 39 estados. Outros estados tinham programas semelhantes de controle do tabaco e foram retirados da análise. Aqui está a aparência de nossos dados.

Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program

Alberto ABADIE, Alexis DIAMOND, and Jens HAINMUELLER

Building on an idea in Abadie and Gardeazabal (2003), this article investigates the application of synthetic control methods to comparative case studies. We discuss the advantages of these methods and apply them to study the effects of Proposition 99, a large-scale tobacco control program that California implemented in 1988. We demonstrate that, following Proposition 99, tobacco consumption fell markedly in California relative to a comparable synthetic control region. We estimate that by the year 2000 annual per-capita cigarette sales in California were about 26 packs lower than what they would have been in the absence of Proposition 99. Using new inferential methods proposed in this article, we demonstrate the significance of our estimates. Given that many policy interventions and events of interest in social sciences take place at an aggregate level (countries, regions, cities, etc.) and affect a small number of aggregate units, the potential applicability of synthetic control methods to comparative case studies is very large, especially in situations where traditional regression methods are not appropriate.

KEY WORDS: Observational studies; Proposition 99; Tobacco control legislation; Treatment effects.



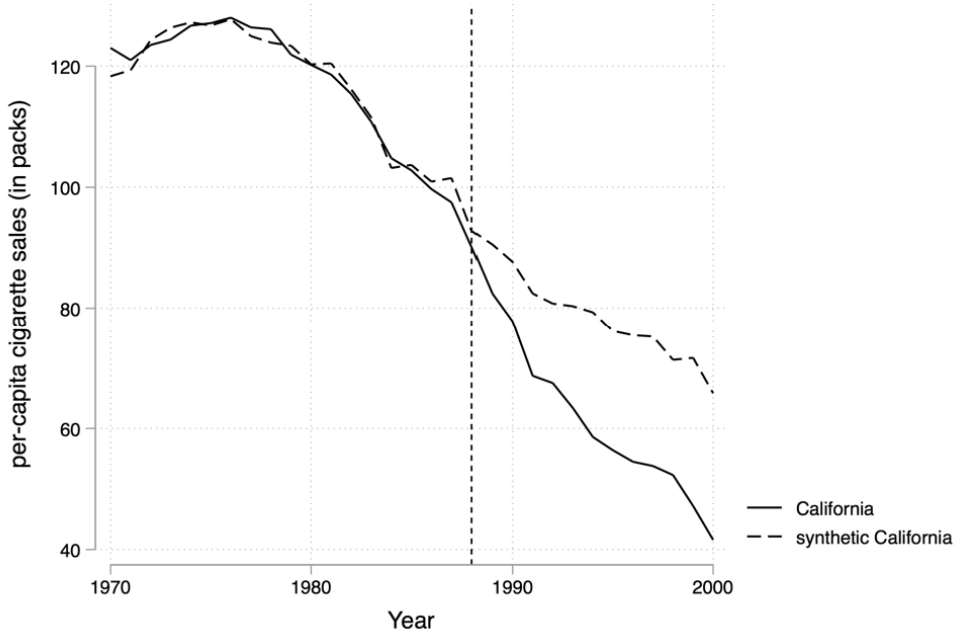


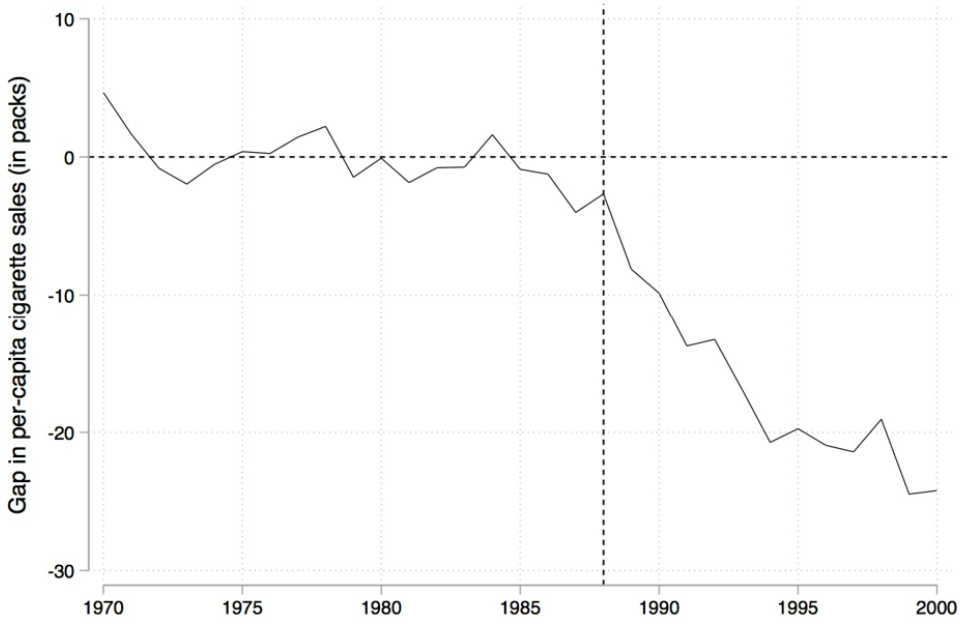
Table 1. Cigarette sales predictor means

Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

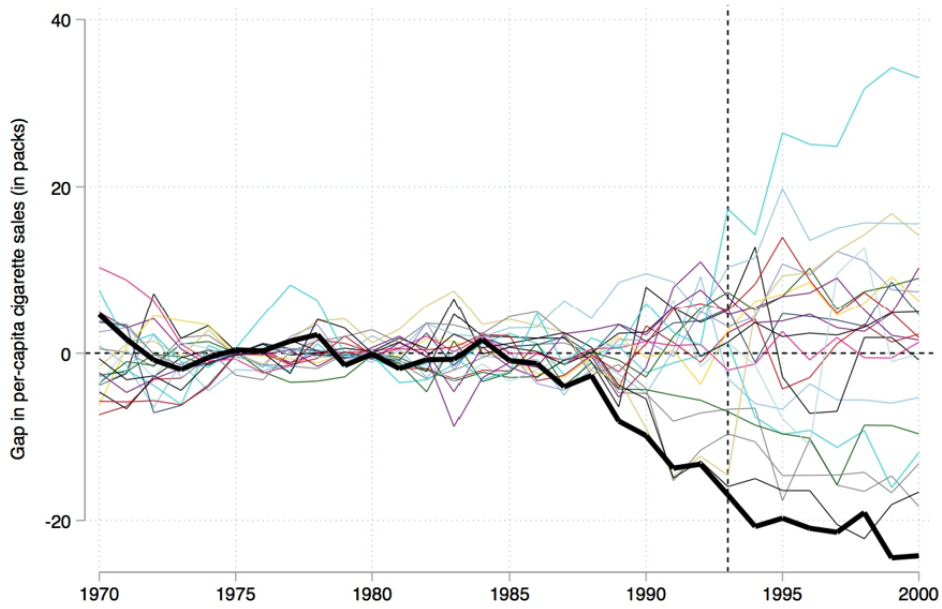


Para inferência estatística, os autores propõe o teste de placebo, e a razão Post-treatment MSPE/Pre-treatment MSPE.

1. Permutation distribution (Placebo test)
2. Razão: Post-treatment MSPE/Pre-treatment MSPE

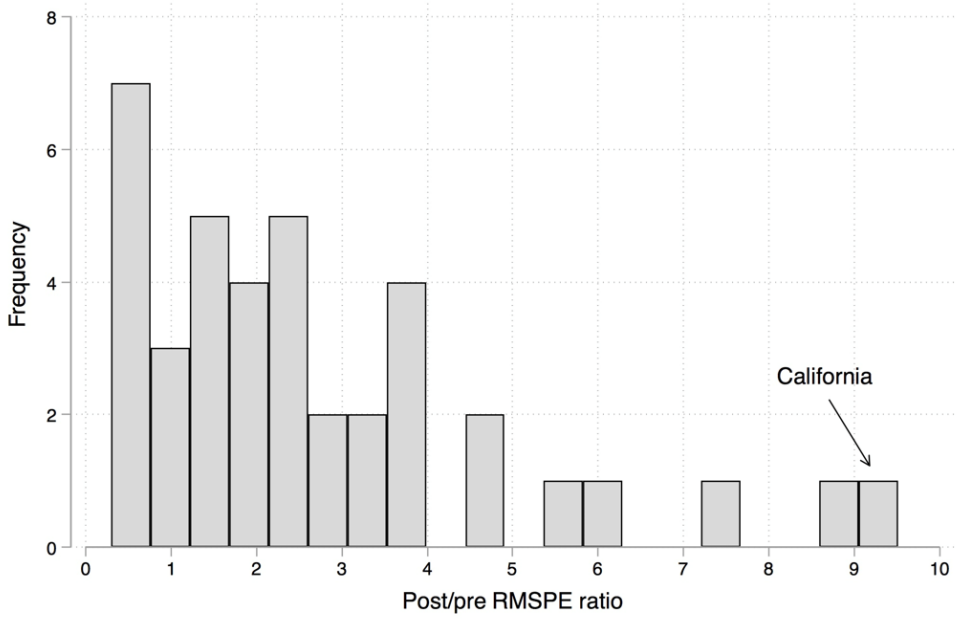
No **primeiro caso**, podemos obter a distribuição de permutação pela redistribuição do tratamento nas unidades de controle da "piscina de doadores" ("donor pool") e estimar efeitos placebo em cada interação.

O **efeito do tratamento** estimado é **significativo se ele se diferencia da distribuição de permutação**.



No segundo caso,

- **MSPE**: Mean Squared Prediction Error
- Prediction Error = Valor observado - Valor contrafactual
- Em termos de inferência para o ATT, o teste estatístico é a razão do MSPE pós-tratamento sobre o MSPE pré-tratamento. **Essa razão deve ser maior para a unidade tratada em comparação com as unidades de controle**.



Posteriormente, **Abadie, Diamond e Hainmueller (2015)** propuseram estimativas de **p-valores** identificados pela raiz da RMSPE (raiz do erro quadrático médio de previsão) com base nos testes de placebo em que o status do tratamento é permutado aleatoriamente em unidades não tratadas. A ideia é que se houver efeito, o RMSPE será grande. Assim como o MSPE comentado anteriormente. Por exemplo, se MSPE é 100, o RMSPE é 10. Se o MSPE é 25, o RSMPE é 5. O p-valor será a raiz do pós sobre o pré.

Considerações sobre o SCM

Aprendemos que se tivermos apenas dados de nível agregado sobre entidades como cidades ou estados, a comparação entre diferenças não nos permitirá fazer inferências. Além disso, tem algumas outras limitações, uma vez que tem de definir uma unidade de controlo e uma única unidade de controlo pode não ser uma representação muito boa do cenário contrafactual para a unidade tratada.

Para corrigir isso, aprendemos que podemos construir um controle sintético que combina múltiplas unidades de controle para torná-las semelhantes à unidade tratada. Com este controlo sintético, pudemos ver o que teria acontecido à nossa unidade tratada na ausência de tratamento.

Finalmente, vimos como poderíamos usar os Testes Exatos de Fisher para fazer inferências com controle sintético. Ou seja, fingimos que as unidades não tratadas eram na verdade as tratadas e calculamos o seu efeito. Estes foram os efeitos placebo: os efeitos que observaríamos mesmo sem tratamento. Nós os usamos para ver se o efeito do tratamento que estimamos foi estatisticamente significativo.

Synthetic Difference-in-Differences - SDD

O DD é um dos muitos métodos de dados em painel possíveis. Uma das alternativas é o SCM. As literaturas de DD e SCM evoluíram separadamente, usando diferentes processos de geração de dados como linha de base (Abadie, 2021). Trabalhos recentes começaram a combinar *insights* das duas literaturas (por exemplo, Arkhangelsky et al 2021; Ben-Michael, Feller e Rothstein, 2021; Doudchenko e Imbens, 2016).

Em geral, **os métodos DD são aplicados** nos casos em que temos um **número substancial de unidades expostas à política**, e os pesquisadores estão dispostos a fazer uma **suposição de "tendências paralelas"** que implica que podemos controlar adequadamente os efeitos de seleção contabilizando efeitos aditivos específicos de unidade e específicos de tempo.

Em contraste, **os métodos SCM**, introduzidos em uma configuração com apenas **um único (ou pequeno número) de unidades expostas**, procura compensar a falta de tendências paralelas **reponderando as unidades para corresponder às suas tendências pré-exposição**.

Synthetic Diference in Diferences (SDD) combina características atrativas de ambos os métodos.

- Assim como o SCM, repondera e combina as tendências de pré-exposição para enfraquecer a dependência de suposições de tendências paralelas.
- Assim como o DD, o método é invariante para deslocamentos aditivos em nível de unidade e permite inferências válidas em dados em painel.

Requerimentos do SDD

- Um **painel balanceado** de N unidades observados em T períodos de tempo
- Uma variável de resultado, denominada Y_{it} observada para cada unidade i em cada período t.
- Algumas dessas unidades são tratadas (D_{it}). Esta variável de tratamento $D_{it} = 1$ se a observação i for tratada no tempo t, caso contrário, $D_{it} = 0$ indica que a unidade i não é tratada no tempo t.
- Importante: Consideramos que uma unidade uma vez tratada, devem permanecer expostas ao tratamento para sempre.
- Não consideramos as unidades sempre tratadas (Always treated) no período.

Formalizando o SDD

Nosso objetivo é estimar consistentemente o **efeito causal do tratamento** sobre D_{it}, *Efeito Médio do Tratamento sobre os Tratados - ATT*, mesmo que **não acreditemos na suposição de tendências paralelas** entre todas as unidades de tratamento e controle em média. As estimativas do ATT procedem da seguinte forma para os métodos DD, SCM e SDD.

DD

O **método de DD** canônico pode ser representado pelo TWFE. O DD é uma regressão não ponderada, com efeitos fixos individuais e de tempo. Repare que o DD atribui **pesos iguais a todos os períodos de tempo** (não estamos ponderando nada).

$$\hat{\tau}^{did} = \underset{\mu,\alpha,\beta,\tau}{argmin} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - (\mu + \alpha_i + \beta_t + \tau D_{it}) \right)^2 \right\}$$

SCM

O **método de controle sintético (SCM)** mantém os pesos específicos da unidade escolhidos de maneira ideal ω, e **omite os efeitos fixos da unidade** αi. Assim, no SCM a unidade tratada e seu controle sintético devem manter níveis de pré-tratamento aproximadamente equivalentes, bem como tendências pré-intervenção. O SCM é uma regressão ponderada, sem efeitos fixos individuais.

$$\hat{\tau}^{sc} = \underset{\beta,\tau}{argmin} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - \beta_t - \tau D_{it} \right)^2 \hat{w}_i^{sc} \right\}$$

Repare que o SCM **não considera os períodos de tempo de maneira ideal** por meio de alguma ponderação aos termos de tempo.

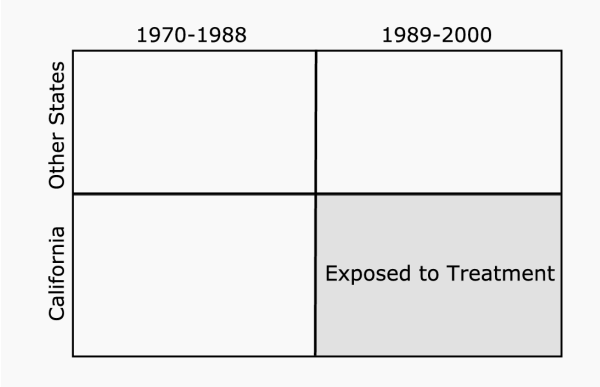
SDD

O **método de SDD** canônico pode ser representado pelo TWFE. Os efeitos individuais (α_i) capturam a diferença nos interceptos para cada unidade, enquanto os efeitos temporais (β_t) capturam a tendência geral nas unidades tratadas e de controle. A presença de efeitos fixos individuais implica que o SDD simplesmente procurará combinar as unidades tratadas e de controle nas tendências de pré-tratamento, e não necessariamente nas tendências e níveis de pré-tratamento, permitindo uma diferença constante entre as unidades de tratamento e controle.

$$\hat{\tau}^{sdid} = \underset{\mu,\alpha,\beta,\tau}{argmin} \left\{ \sum_{i=1}^{N} \sum_{t=1}^{T} \left(Y_{it} - (\mu + \alpha_i + \beta_t + \tau D_{it})\right)^2 \hat{w}_i^{sdid} \hat{\lambda}_t^{sdid} \right\}$$

Repare que no SDD, agora possuímos duas variáveis de ponderação: $\hat{w}_i^{sdid}$ e $\hat{\lambda}_t^{sdid}$. A primeira pondera as unidades de tratamento e controle, enquanto a segunda pondera os períodos de tempo.

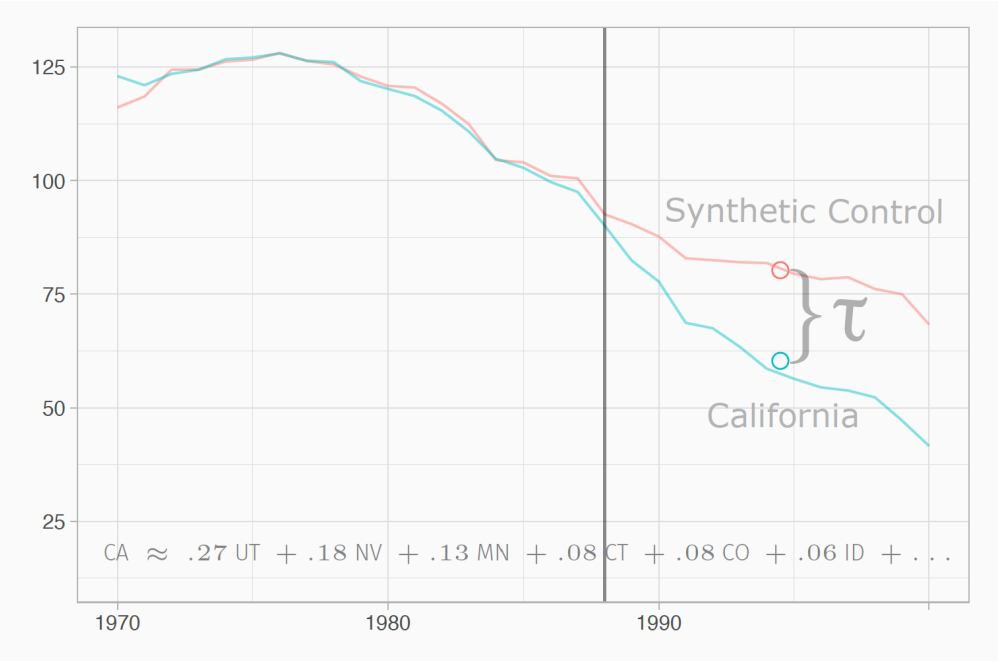
Para entender de modo intuitivo essas ponderações, vamos combinar teoria com exemplo do caso da Califórnia. Isto é, podemos ver a estrutuda dos dados da seguinte forma:



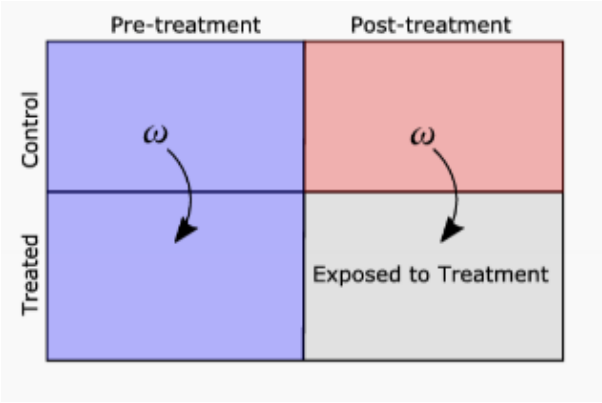
Para calcular o efeito ATT através do método DD, apenas comparamos as médias PRÉ e PÓS tratamento entre as unidas tratadas e de controle. Considerando o SCM, partimos para um tipo de estimador automatizado (sem influência direta do pesquisador sobre a escolha da amostra), que relaxa a hipótese de tendências paralelas, e que "simplifica" a comparação entre tratado e controle. A ideia é uma regressão (comparação) entre a unidade tratada e seu "controle sintético" no período pós-intervenção. Voltando ao exemplo da Califórnia, pressupomos que no período Pré-Intervenção, são semelhantes. Ou seja:

$$Califórnia_i \approx \sum_i w_i \times Control_{it}$$

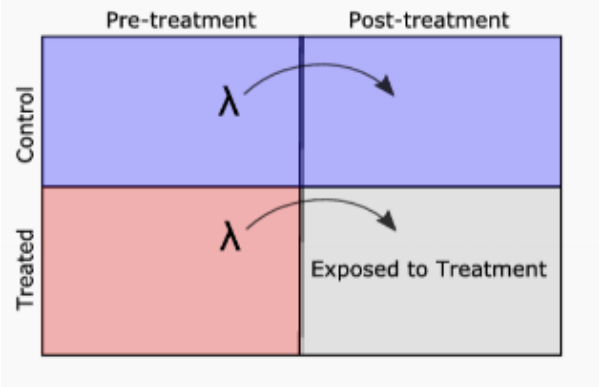
dessa forma, atribuimos à diferença entre a unidade tratada e seu contrafactual no período pós-tratamento à intervenção. Vejamos o resultado do SCM para a análise da Califórnia:



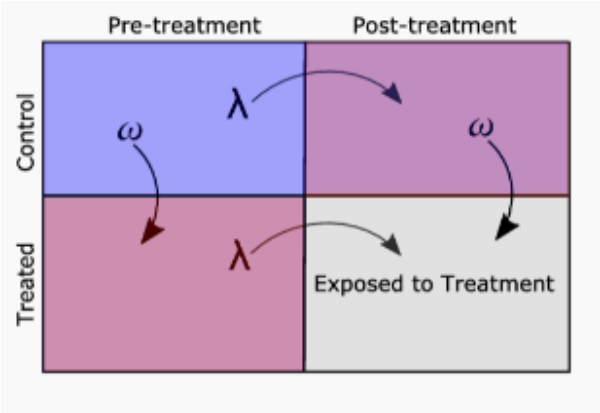
Voltando ao quadro onde dividimos os dados em quadrantes. O SCM usa os dados pré-tratamento entre tratados e conotroles (área azul), tomamos a média ponderada dos controles que melhor se ajustam à unidade tratada (também conhecida como **regressão vertical**). Assumindo que essa relação se mantem válida no período pós-tratamento, nós utilizamos a mesma média dos controles para imputar o contrafactual da unidade tratada.



Agora vamos tentar conectar essa ideia às demais possibilidades que essa visão em quadrantes da estrutura dos dados nos permite. Por exemplo, podemos utilizar os dados para projetar (prever/forecasting) o ponto de referência no termo pré-tratamento (também conhecida como **regressão horizontal**). Então, usando os controles (área azul), podemos obter uma ponderação dos períodos que melhor prevêem o período médio pós-tratamento. Supondo que essa relação permaneça válida para os tratados, usamos a mesma ponderação dos períodos para imputar observações "livres da intervenção" para o tratado.



Então, o SDD será uma combinação dessas duas ideias. Ou seja, o SDD é uma combinação da regressão vertical e horizontal.



Cabe destacar que o SDD permite um intercepto no peso ω e adiciona um termo de penalidade $L2$. O intercepto representa o fato de que o objetivo de ω não é mais corresponder perfeitamente ao grupo tratado no período pré-tratamento, mas apenas imitar a tendência dele. Além disso, a penalidade tenta evitar uma concentração excessiva de peso em um pequeno número de unidades de controle, estabilizando a estimativa. Com relação ao λ, ele ajusta o ponto de referência no termo pré-tratamento.

Em suma, o SDD é uma combinação do DD com o SCM, em que realizamos as seguintes etapas:

- 1. Estimamos o peso de unidades (ω) definindo uma unidade de controle sintético usando os dados pré-tratamento.

$$\hat{\omega_0} + \hat{\omega^T} Y_{co,pre} \approx \bar{Y_{tr,pre}}$$

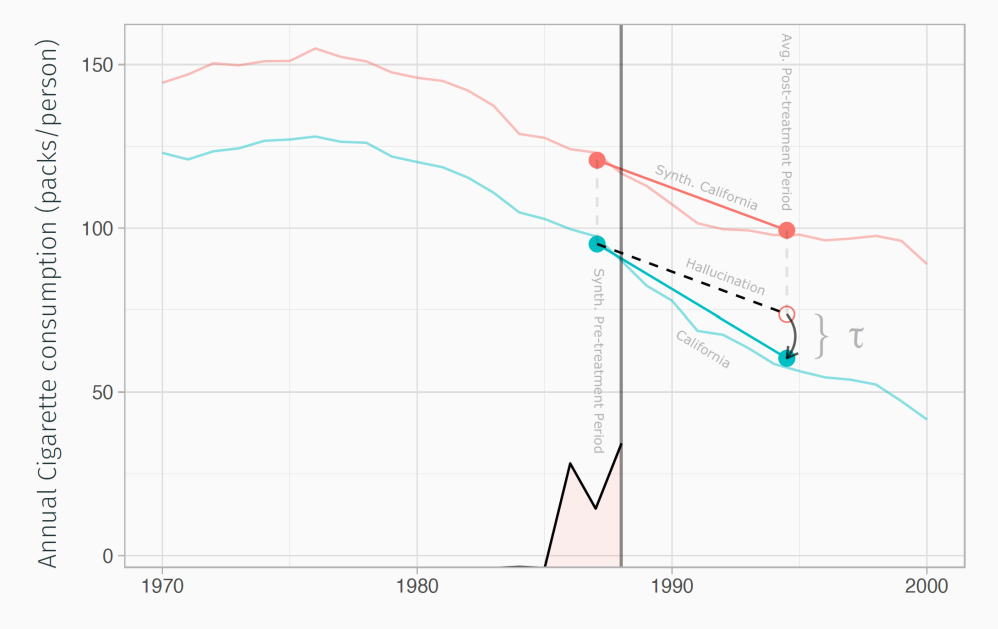
- 2. Estimamos os pesos de tempo (λ) definindo um período pré-tratamento sintético usando os dados do controle.

$$\hat{\lambda_0} + Y_{co,pre}\hat{\lambda} \approx \overline{Y_{co,post}}$$

- 3. Aplica-se o DD 2x2 conforme a figura abaixo.

	Synthetic Pre-Treatment	Average Post-treatment
Synthetic Control	$\hat{\omega}^T Y_{co,pre}\hat{\lambda}$	$\hat{\omega}^T Y_{co,\overline{post}}$
Average Treated	$Y_{\overline{tr},pre}\hat{\lambda}$	$Y_{\overline{tr},\overline{post}}$

Então, voltando ao caso da California, temos:



Caso você queira entender o processo para identificar os pesos ótimos de modo formal, ver em Arkhangelskyetal.(2021,pp.4091-4092).

Condicionando em Covariáveis

Em certas configurações, pode ser relevante condicionar covariáveis que variam no tempo X_{it}. Observe que, neste caso, podemos prosseguir aplicando o algoritmo SDD aos resíduos calculados como:

$$Y_{it}^{res} = Y_{it} - X_{it}\theta$$

Onde θ vem da regressão de Y_{it} sobre X_{it}. Nesse sentido, o SDD difere do SCM (Abadie et al, 2010). Na concepção de Abadie et al (2010), quando as covariáveis são incluídas no modelo de controle sintético são escolhidas para garantir que essas covariáveis sejam aproximadas o melhor possível entre a unidade tratada e a unidade de controle sintético.

Entretanto, no SDD o ajuste pelas covariáveis é visto como uma tarefa de pré-processamento, que remove o impacto das mudanças das covariáveis no resultado Y_{it} antes de calcular o controle sintético.

Arkhangelsky et al.(2021) condicionam essas variáveis X_{it} encontrando θ num procedimento de otimização que adicionalmente permite o cálculo eficiente de pesos

ótimos.

Desenho de adoção escalonada (*The Staggered Adoption Design*)

Arkhangelsky et al. (2021, Apêndice A), eles observam que esse procedimento pode ser estendido para uma adoção escalonada, em que as unidades tratadas adotam o tratamento em momentos variados.

No design de adoção escalonada, várias datas de adoção são observadas. Considere por exemplo a matriz de tratamento abaixo, composta por 8 unidades, sendo 2 (1 e 2) não tratadas, enquanto as outras 6 são tratadas, porém em pontos variados.

$$D = \begin{bmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 5 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 7 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Essa matriz de tratamento escalonada D pode ser quebrada em matrizes específicas por data. D^1, D^2 e D^3, ou genericamente, $D^1, \ldots, D^A$, onde A indica o número distinto de datas de inicio do tratamento. E o vetor A, contem os diferentes períodos de adoção. Nesse exemplo:

$$A = (3, 4, 6)$$

Finalmente, as matrizes específicas D^1,D^2 e D^3 consistem:

$$D^1 = \begin{bmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

$$D^2 = \begin{bmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$D^3 = \begin{bmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 8 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Conforme estabelecido em Arkhangelsky et al. (2021, Apêndice A), o efeito médio do tratamento no tratado pode então ser calculado aplicando o SDD a cada uma dessas 3 amostras específicas e calculando uma média ponderada dos três, onde os pesos são atribuídos com base no número relativo de unidades tratadas e períodos de tempo em cada grupo de adoção.

Aplicação no Python

Pacote 'synthdid', estamos replicando o pacote de SDD, para o caso da Califórnia. Ver mais detalhes em:
https://github.com/MasaAsami/pysynthdid/blob/main/notebook/ReproductionExperiment_C

In [48]:
```
import warnings

warnings.filterwarnings("ignore")

import sys
import os

sys.path.append(os.path.abspath("../"))

import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from scipy.stats import spearmanr
plt.style.use('ggplot')

from tqdm import tqdm

from synthdid.model import SynthDID
```

In [49]:
```
df = pd.read_csv("https://github.com/Daniel-Uhr/data/raw/main/california_prop99.
df.head()
```

Out[49]:

	State	Year	PacksPerCapita	treated
0	Alabama	1970	89.800003	0
1	Arkansas	1970	100.300003	0
2	Colorado	1970	124.800003	0
3	Connecticut	1970	120.000000	0
4	Delaware	1970	155.000000	0

In [51]:
```
data = df.pivot("Year", "State", "PacksPerCapita")
data.head().round()
```

Out[51]:

State	Alabama	Arkansas	California	Colorado	Connecticut	Delaware	Georgia	Idaho
Year								
1970	90.0	100.0	123.0	125.0	120.0	155.0	110.0	102.0
1971	95.0	104.0	121.0	126.0	118.0	161.0	116.0	108.0
1972	101.0	104.0	124.0	134.0	111.0	156.0	117.0	126.0
1973	103.0	108.0	124.0	138.0	109.0	155.0	120.0	122.0
1974	108.0	110.0	127.0	133.0	112.0	151.0	124.0	126.0

5 rows × 39 columns

In [52]:
```
PRE_TEREM = [1970, 1988]
POST_TEREM = [1989, 2000]

TREATMENT = ["California"]

data.head()
```

Out[52]:

State	Alabama	Arkansas	California	Colorado	Connecticut	Delaware	Geor
Year							
1970	89.800003	100.300003	123.000000	124.800003	120.000000	155.000000	109.900
1971	95.400002	104.099998	121.000000	125.500000	117.599998	161.100006	115.699
1972	101.099998	103.900002	123.500000	134.300003	110.800003	156.300003	117.000
1973	102.900002	108.000000	124.400002	137.899994	109.300003	154.699997	119.800
1974	108.199997	109.699997	126.699997	132.800003	112.400002	151.300003	123.699

5 rows × 39 columns

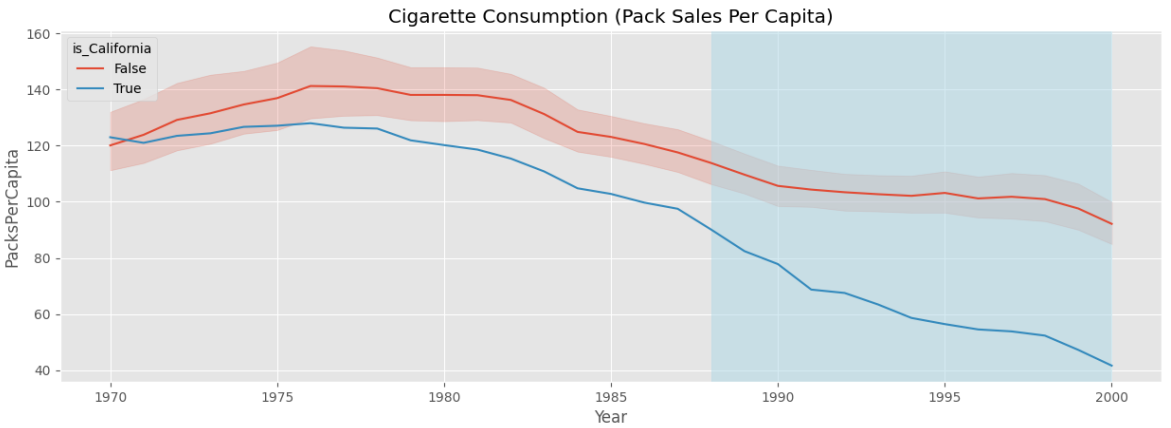
In [53]:
```
melt_df = pd.melt(
    data.reset_index().rename(columns={"index": "Year"}),
    id_vars="Year",
    value_name="PacksPerCapita",
    var_name="State",
)
melt_df["is_California"] = melt_df["State"] == "California"
melt_df
```

Out[53]:

	Year	State	PacksPerCapita	is_California
0	1970	Alabama	89.800003	False
1	1971	Alabama	95.400002	False
2	1972	Alabama	101.099998	False
3	1973	Alabama	102.900002	False
4	1974	Alabama	108.199997	False
...	...	...	...	...
1204	1996	Wyoming	110.300003	False
1205	1997	Wyoming	108.800003	False
1206	1998	Wyoming	102.900002	False
1207	1999	Wyoming	104.800003	False
1208	2000	Wyoming	90.500000	False

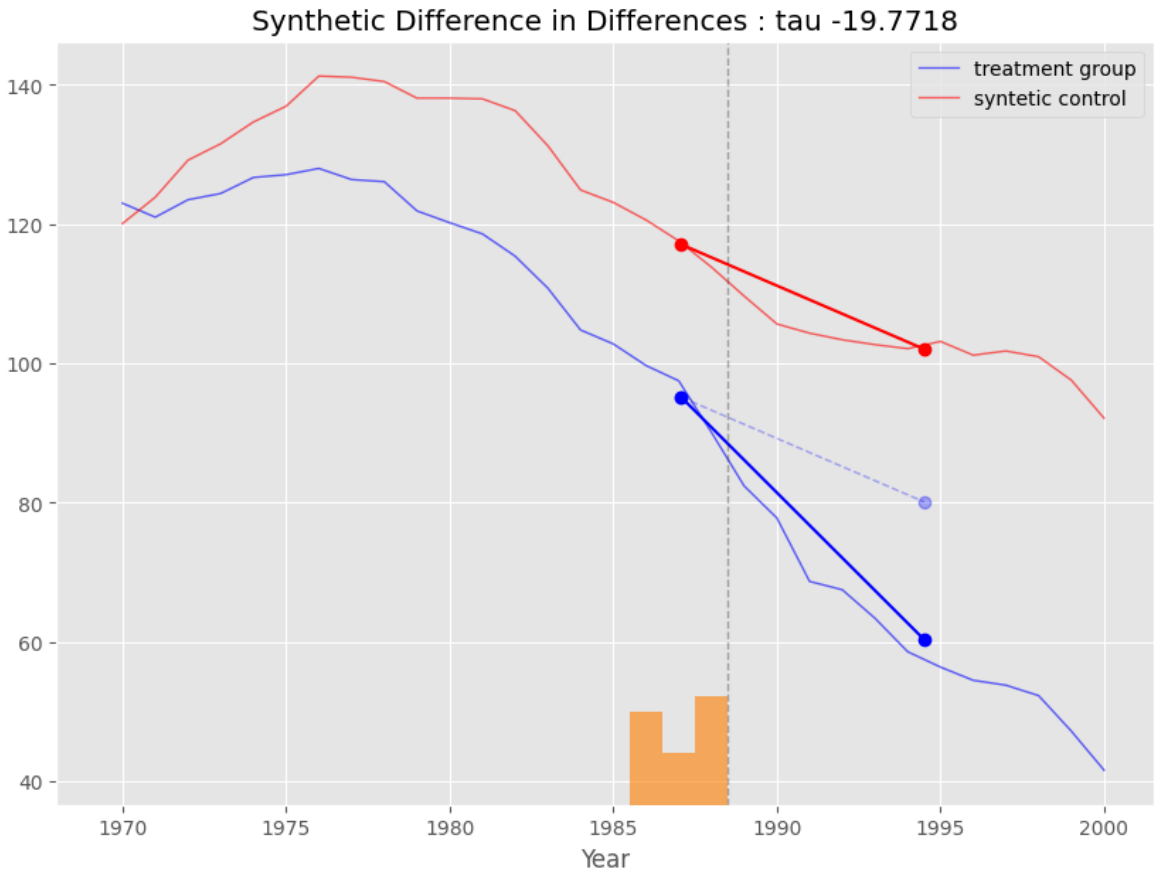
1209 rows × 4 columns

In [54]:
```
fig, ax = plt.subplots()
fig.set_figwidth(15)

sns.lineplot(x="Year", y="PacksPerCapita", hue="is_California", data=melt_df, ax

ax.axvspan(1988, 2000, alpha=0.5, color="lightblue")

plt.title("Cigarette Consumption (Pack Sales Per Capita)")
plt.show()
```

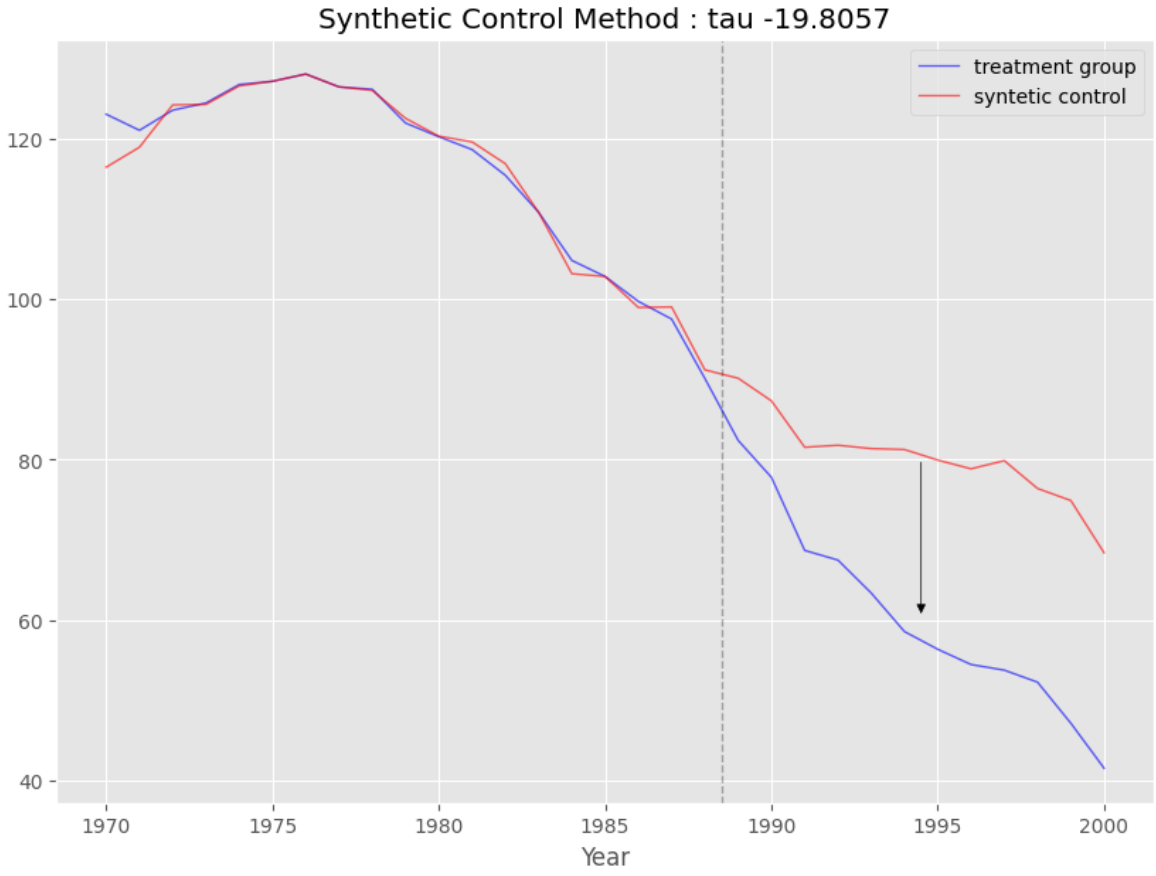


In [55]:
```
sdid = SynthDID(data, PRE_TEREM, POST_TEREM, TREATMENT)
sdid.fit(zeta_type="base")
```

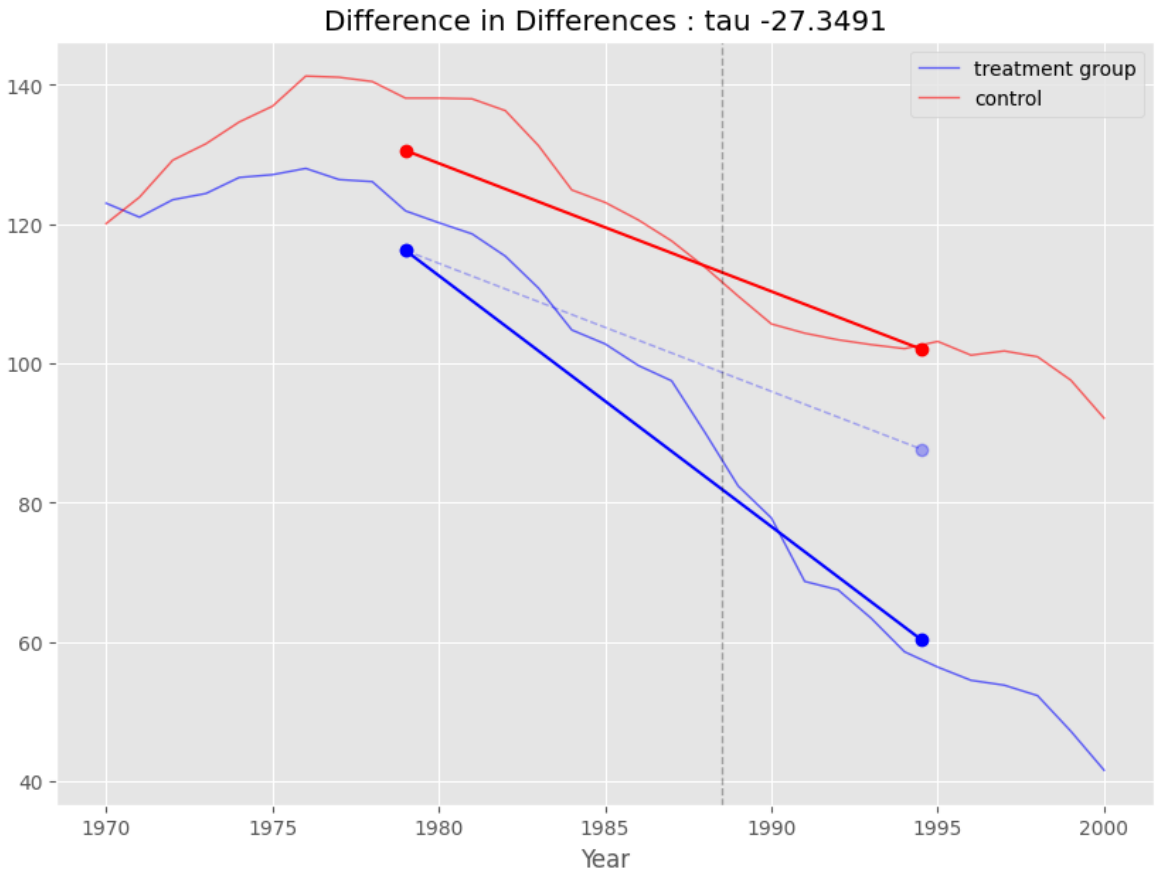
In [56]:
```
sdid.plot(model="sdid")
```



In [57]:
```
sdid.plot(model="sc")
```



In [58]:
```
sdid.plot(model="did")
```


Difference in Differences : tau -27.3491
treatment group
control
140
120
100
80
60
40
1970
1975
1980
1985
1990
1995
2000
Year