Causal Inference with Observational Data

Lê Việt Phú
Fulbright School of Public Policy and Management

Ngày 1 tháng 7 năm 2019

Các thiết kế nghiên cứu để thiết lập quan hệ nhân quả

- Sử dụng thử nghiệm ngẫu nhiên hóa (RCT) để tạo nhóm hưởng lợi và nhóm đối chứng hoàn toàn tương đồng về các điều kiện quan sát được và không quan sát được $\rightarrow$ tiêu chuẩn vàng để thiết lập quan hệ nhân quả.
- Khi không thể thực hiện RCT thì chúng ta có thể sử dụng dữ liệu quan sát được (observational data) nhằm *xây dựng một tình huống nghiên cứu tương tự như thử nghiệm ngẫu nhiên:*
  - Sử dụng các thuật toán thống kê để xây dựng nhóm hưởng lợi và đối chứng tương đồng như thử nghiệm ngẫu nhiên (DiD, matching).
  - Sử dụng tình huống can thiệp tự nhiên (natural experiment) nhằm mô phỏng lại thiết kế thử nghiệm ngẫu nhiên (Regression Discontinuity, IV, regression adjustment)

Thiết kế nghiên cứu

Giải thích quá trình phân bổ nhóm đối tượng hưởng lợi và đối chứng.

- Với thử nghiệm ngẫu nhiên: ngẫu nhiên hóa quá trình lựa chọn đối tượng tham gia.
- Với dữ liệu quan sát được: không đảm bảo việc tham gia là ngẫu nhiên. Có thể kiểm chứng bằng kiểm định điều kiện cân bằng giữa nhóm hưởng lợi và đối chứng.
  - Nếu quá trình tham gia là ngoại sinh, không phụ thuộc ý muốn của đối tượng nghiên cứu $\rightarrow$ Tình huống thử nghiệm tự nhiên $\rightarrow$ Có thể mô phỏng gần giống với thử nghiệm RCT nhất!
  - Nếu quá trình tham gia có hiện tượng lựa chọn mẫu (self selection into treatment) $\rightarrow$ Phải có thiết kế nghiên cứu phù hợp với nguyên nhân gây ra hiện tượng tự lựa chọn mẫu.

Threat to validity (confounding)

- Không thể nhận định được các đặc tính không quan sát được (unobservables) có cân đối giữa nhóm hưởng lợi và đối chứng.
- Thuộc tính không quan sát được tương quan với tình trạng tham gia chính sách và kết quả.

$\rightarrow$ Nhóm đối chứng không hợp lệ, và kết quả có thể bị sai lệch.

Khi nào sử dụng dữ liệu quan sát được cho kết quả tin cậy?

Cần có chiến lược nhận diện mô hình (identification strategy) hợp lý! Ví dụ:

- Tìm cách thiết kế nhóm đối chứng sao cho các đặc tính không quan sát được có thể cân bằng (ví dụ sử dụng thử nghiệm tự nhiên - “treatment is as-if random").
- Chấp nhận có sự khác biệt về đặc tính không quan sát được, nhưng nếu chúng không thay đổi theo thời gian (time invariant unobservables) thì có thể dùng sai phân dữ liệu để loại bỏ.
- Sử dụng biến công cụ với điều kiện loại trừ.
- Ghép cặp hoặc dùng synthetic controls để xây dựng nhóm đối chứng.
- Sử dụng hồi quy gián đoạn để loại trừ tác động của nhân tố không quan sát được.

Tất cả những vấn đề trên phải được thảo luận khi đề xuất một nghiên cứu sử dụng dữ liệu quan sát được.

Potential outcome framework

Đối với thử nghiệm ngẫu nhiên đảm bảo việc phân bổ vào nhóm tham gia hay đối chứng hoàn toàn độc lập với kết quả chương trình:

$$Y_i^1, Y_i^0 \perp D_i$$

thì chúng ta ước lượng được tác động can thiệp trung bình bằng sự khác biệt về kết quả của hai nhóm:

$$ATE = E[Y_i^1 - Y_i^0] = \frac{1}{N}\sum_{i=1}^{N}(Y_i^1 - Y_i^0)$$

Đối với dữ liệu quan sát được

$$ATE = \underbrace{\mathbf{E}(Y_i^1|D=1) - \mathbf{E}(Y_i^0|D=1)}_{ATT} + \underbrace{\mathbf{E}(Y_i^0|D=1) - \mathbf{E}(Y_i^0|D=0)}_{Bias}$$

$$ATE = ATT + Selection\ Bias$$

Khi nào thì $ATE \neq ATT$?

- Khi xảy ra hiện tượng lựa chọn mẫu (selection into treatment)
- Khi xác xuất phân bổ vào nhóm tham gia hay đối chứng tương quan với kết quả chương trình, $Y_i^1, Y_i^0 \sim D_i$

Thiết kế DiD

- Để xử lý trường hợp lựa chọn mẫu theo đặc tính không quan sát được (selection on unobservables) nhưng không thay đổi theo thời gian (time invariant unobserved heterogeneity).
- Khác biệt với vấn đề lựa chọn mẫu dựa trên đặc tính quan sát được (selection on observables) là gì?
  - Giả định quan sát được các nhân tố ảnh hưởng đến việc phân bổ vào nhóm tham gia hay đối chứng.
  - Có thể xác lập được nhóm đối chứng hợp lệ dựa trên các đặc tính quan sát được.

Potential Outcome Framework in DiD Design

- Hai nhóm đối tượng:
  - $D = 1$ nhóm hưởng lợi
  - $D = 0$ nhóm kiểm soát
- Hai thời điểm
  - $T = 0$ trước khi thực hiện chương trình
  - $T = 1$ sau khi thực hiện chương trình
- Kết quả có thể xảy ra
  - $Y_{1i}(t)$ Kết quả tiềm năng của đối tượng i thuộc nhóm hưởng lợi tại thời điểm t
  - $Y_{0i}(t)$ Kết quả tiềm năng của đối tượng i thuộc nhóm kiểm soát tại thời điểm t

Kết quả thực hiện của hai nhóm tại hai thời điểm

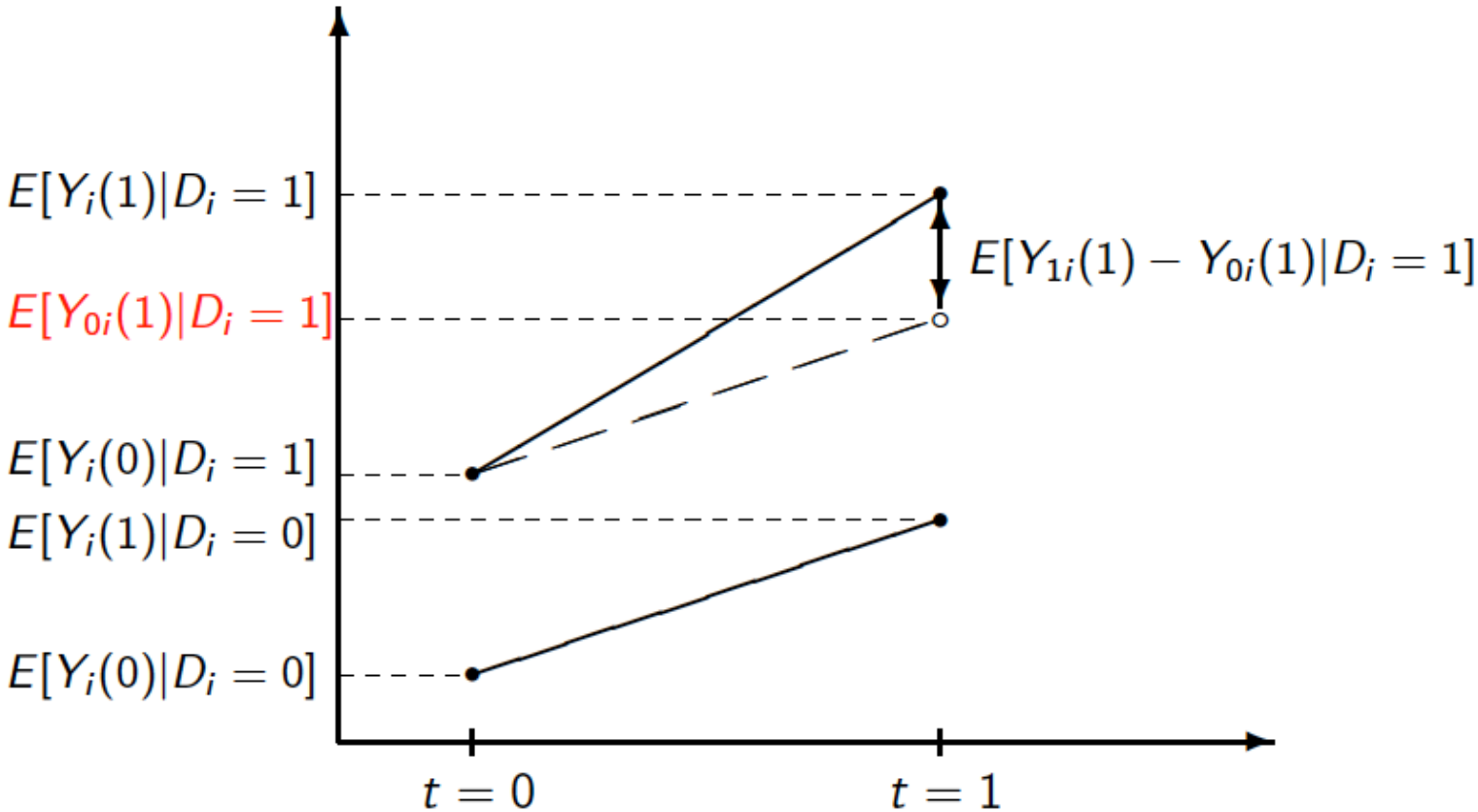
	$T=0$	$T=1$	**Differences**
$D=0$	$E[Y_{0i}(0)\|D_i=0]$	$E[Y_{0i}(1)\|D_i=0]$	$E[Y_{0i}(1)-Y_{0i}(0)\|D_i=0]$
$D=1$	$E[Y_{1i}(0)\|D_i=1]$	$E[Y_{1i}(1)\|D_i=1]$	$E[Y_{1i}(1)-Y_{1i}(0)\|D_i=1]$

Với giả định song song,

$$\underbrace{E[Y_{0i}(1)|D=1]-E[Y_{0i}(0)|D=1]}_{\textit{what if treated unit didn't receive treatment}} = \underbrace{E[Y_{0i}(1)|D=0]-E[Y_{0i}(0)|D=0]}_{\textit{change in control unit}}$$

Chứng minh:

$$\begin{aligned} ATT &= E[Y_{1i}(1)|D=1]-E[Y_{0i}(1)|D=1] \\ &= E[Y_{1i}(1)-Y_{1i}(0)|D_i=1]-E[Y_{0i}(1)-Y_{0i}(0)|D_i=0] \end{aligned}$$


$E[Y_i(1)|D_i = 1]$
$E[Y_{0i}(1)|D_i = 1]$
$E[Y_i(0)|D_i = 1]$
$E[Y_i(1)|D_i = 0]$
$E[Y_i(0)|D_i = 0]$
$E[Y_{1i}(1) - Y_{0i}(1)|D_i = 1]$
$t = 0$
$t = 1$

Identification với DiD

- Giả định song song thỏa nếu nhân tố quan sát được ảnh hưởng đến lựa chọn mẫu không thay đổi theo thời gian (time invariant) và mang tính cộng dồn (additive).
- Nếu có nhân tố không quan sát được thay đổi theo thời gian ảnh hưởng đến kết quả chương trình (time-varying unobservables) thì giả định song song bị vi phạm.

Threat to validity

- Đánh giá tác động ngắn hạn và dài hạn.
- Dữ liệu repeated cross-sectional và panel data.
- Attrition.
- Functional forms.
- DiD không xử lý được vấn đề simultaneity/reverse causality.

Những dạng mô hình sử dụng DiD để thiết lập quan hệ nhân quả

- DiD cơ sở (plugged-in DiD)
- Subgroup effects
- DiD mở rộng
- Placebo test
- Matching with DiD
- DDD
- Multilevel/hierarchical model
- Synthetic controls

Plugged-in Estimator

Dùng khác biệt nhóm trước và sau (diff-in-diff in means) để ước tính ATT:

$$\begin{aligned} ATT &= E[Y_{1i}(1)|D=1] - E[Y_{0i}(1)|D=1] \\ &= E[Y_{1i}(1) - Y_{1i}(0)|D_i=1] - E[Y_{0i}(1) - Y_{0i}(0)|D_i=0] \end{aligned}$$

Có thể áp dụng với repeated cross-sectional data, tuy nhiên cảnh giác thay đổi cấu trúc nhóm.

Ước lượng ATT bằng hồi quy

$$Y = \beta_0 + \beta_1 * T + \beta_2 * Year + \beta_{\mathbf{3}} * (T \times Year) + u$$

	Trước ($Year = 0$)	**Sau** ($Year = 1$)	$\triangle$**Y**
Đối chứng ($T = 0$)	$Y = \beta_0$	$Y = \beta_0 + \beta_2$	β_2
Hưởng lợi ($T = 1$)	$Y = \beta_0 + \beta_1$	$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
			$ATT_{DiD} = \beta_{\mathbf{3}}$

Có thể đưa thêm các biến kiểm soát khác vào mô hình hồi quy $\sum_j X_j \beta j$.

DiD để ước lượng subgroup effects

- Nếu can thiệp được thực hiện ở cấp độ cao hơn cấp độ cá nhân (ví dụ với dữ liệu doanh nghiệp, chúng ta muốn ước lượng tác động của môi trường kinh doanh cấp tỉnh lên hiệu quả hoạt động của doanh nghiệp) thì vẫn có thể áp dụng DiD, tuy nhiên khi đó giả định là tất cả các doanh nghiệp trong cùng một tỉnh đều bị ảnh hưởng giống nhau (do đó phải dùng cluster standard errors để điều chỉnh tương quan nội nhóm.)
- Có thể có sự khác biệt về tác động giữa các nhóm doanh nghiệp có đặc tính khác nhau trong cùng một tỉnh (heterogeneous effects), ví dụ doanh nghiệp lớn bị tác động khác với doanh nghiệp nhỏ.

Diff-in-diff-in-diff

DDD ước lượng sự khác biệt giữa DiD của chính sách cần nghiên cứu và DiD của nhóm đối chứng giả (placebo).

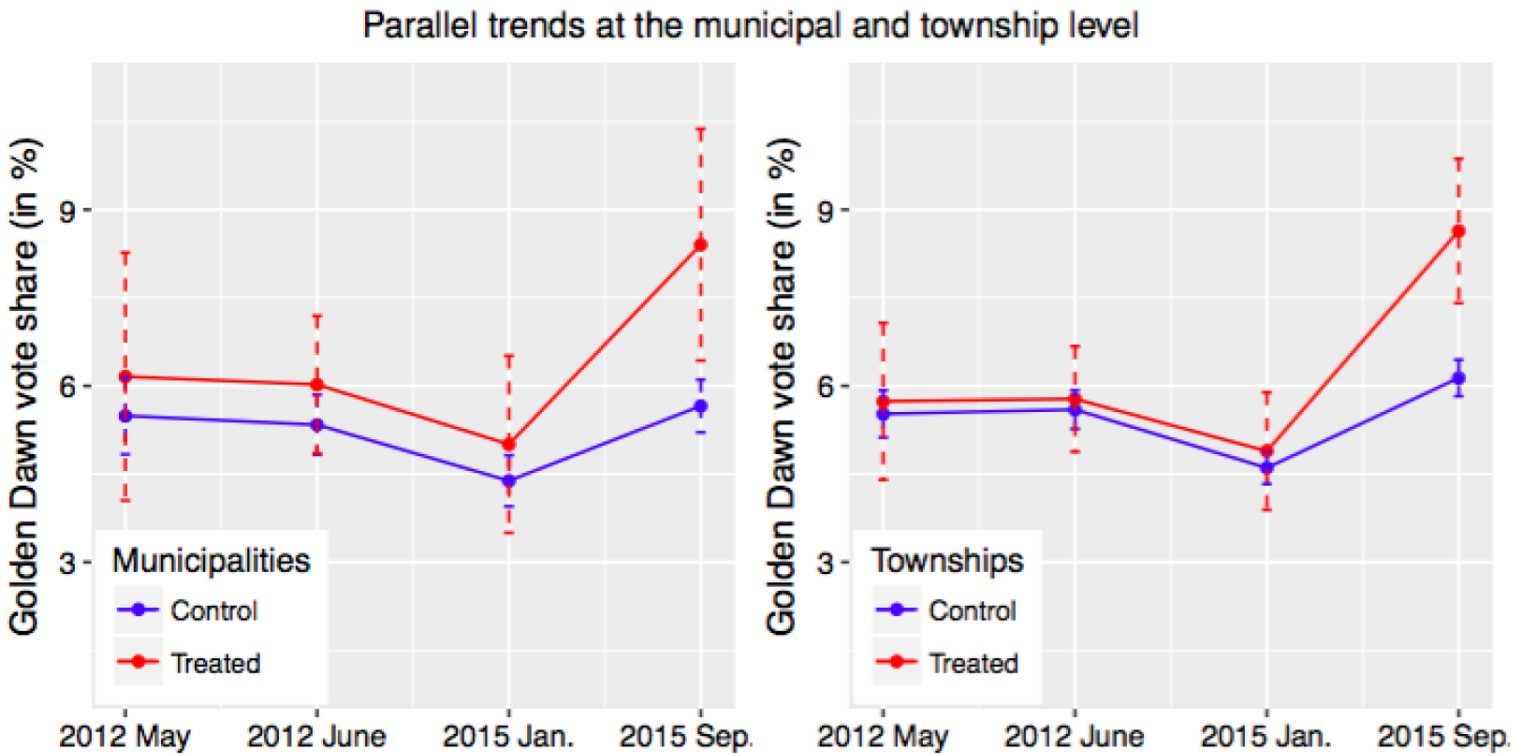
- Về nguyên tắc, $DiD_{placebo} = 0$, do đó DiD và DDD cho tác động giống nhau. Tuy nhiên DDD sẽ loại bỏ vấn đề lựa chọn mẫu có thể mắc phải trong DiD nếu giả định song song bị vi phạm. Có thể chọn DiD, và dùng DDD để kiểm định độ nhạy.
- Nếu $DiD_{placebo} \neq 0$ thì cần xem xét lại mô hình.

Áp dụng khi nào?

- Có nhiều hơn một nhóm kiểm soát
- Có nhiều hơn là hai thời điểm

Placebo test

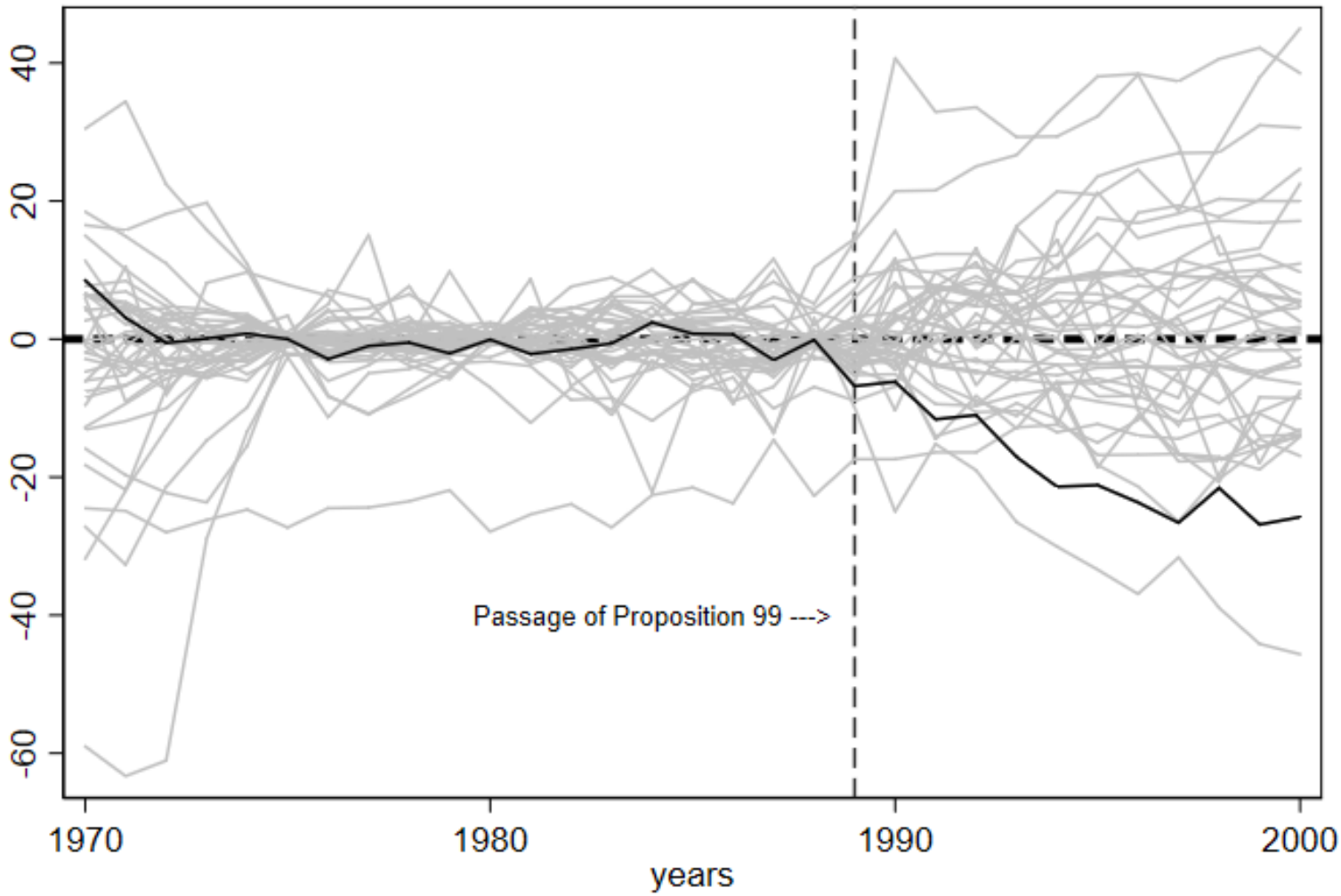
Xu hướng thay đổi của nhóm kiểm soát và hưởng lợi có tương đồng hay không?



Matching with DiD

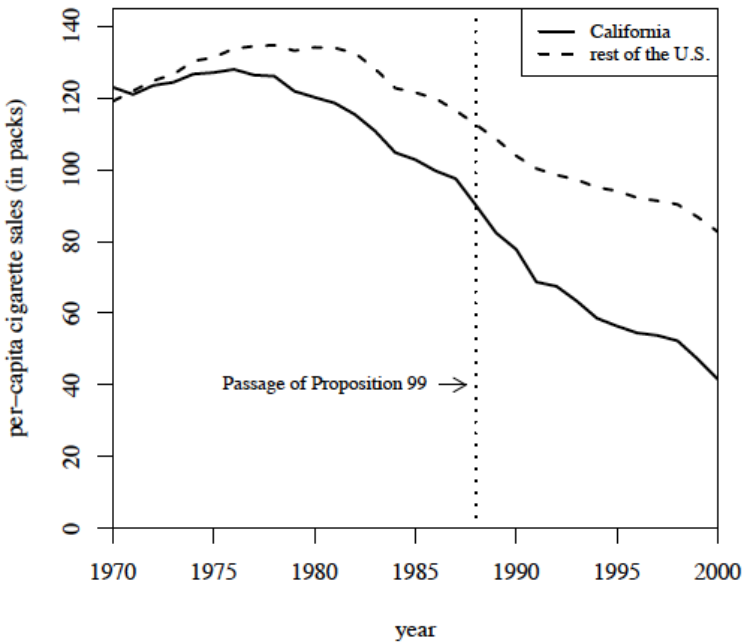
- Kết hợp giữa PSM với DiD để lọc nhóm đối chứng và hưởng lợi dựa trên đặc tính quan sát được.
- Synthetic controls: Khi có nhiều nhóm đối chứng, có thể dùng thuật toán để tìm ra một nhóm đối chứng tối ưu bằng cách kếp hợp giữa các nhóm đối chứng khác nhau. Áp dụng khi có dữ liệu bảng với $T >> n$.

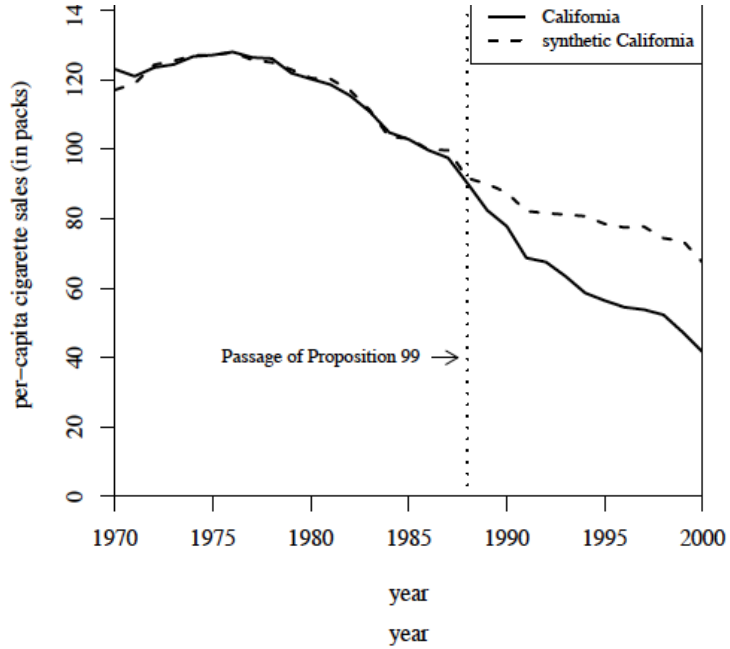
Tác động của thuế thuốc lá lên số điếu hút bình quân theo bang.



Re: Figure 4. Per-cap cig sales gaps in California & placebo gaps in 38 control states

Xây dựng nhóm đối chứng tổng hợp và kiểm định.





Thiết kế DiD như thế nào?

- Dựa vào vị trí địa lý
- Dựa vào thời gian
- Dựa vào các quy định hành chính