R演習

パネルデータ解析

慶應義塾大学 片山翔太

Rコードはこちら

パネルデータ(panel data)とは

- **パネルデータ**
  - ユニットの特徴を一定期間に渡り繰り返し観測
  - 以下のようなロング型で表される

nr	year	lwage	union	exper
13	1980	1.1975402	0	1
13	1981	1.8530600	1	2
13	1982	1.3444617	0	3
13	1983	1.4332134	0	4
13	1984	1.5681251	0	5
13	1985	1.6998910	0	6
13	1986	-0.7202626	0	7
13	1987	1.6691879	0	8
17	1980	1.6759624	0	4
17	1981	1.5183982	0	5
17	1982	1.5591905	0	6

nr : ユニット識別番号

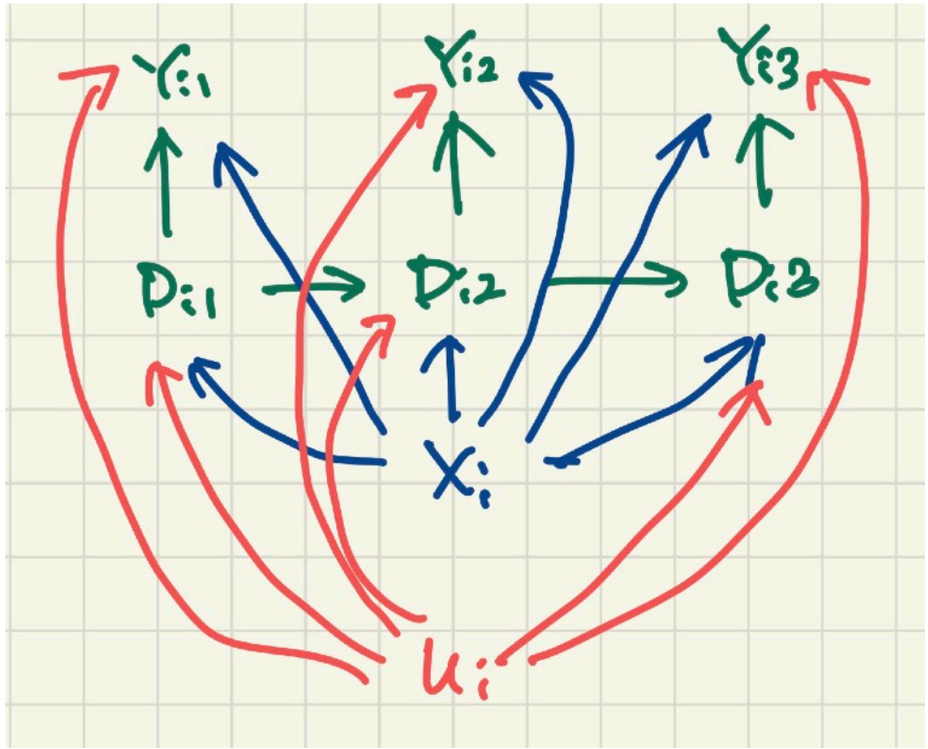
lwage : 賃金 (対数)

union : 労働組合参加

exper : 労働市場参加年数

- **DAGの例 (Imai and Kim (2017))**
  - 3期間のパネルデータ
  - Y_{i1}, Y_{i2}, Y_{i3} : Outcome
  - D_{i1}, D_{i2}, D_{i3} : 興味のある変数
  - u_i : 未観測の交絡変数 and unit-specific

    #時点に依存しない
  - X_i : 観測される交絡変数 and unit-specific



Example

Y : 年収, D : 教育, u : 知能

- **Notes**
  - Dは内生変数(誤差がYに依存する)
  - 未観測の交絡uが存在するケース → unobserved heterogeneity
    - Uは**固定効果(fixed-effects)**とも呼ばれる
  - 因果効果はFixed-effects estimatorで推定可能
    - Within estimatorとも呼ばれる

パネルデータにおける推定

- **DからYへの因果効果を推定したい**
  - 簡単のためXは省略
  - 未観測の交絡変数uの存在は認める
  - Y は1次元, $D = (D_1, \ldots, D_K)^T$ はK次元とする
- **観測データ**
  - 各ユニット$i \in \{1,\ldots,N\}$に対して, T時点までのデータ

$$\{(Y_{it}, D_{it}) : t = 1,\ldots,T\},\ D_{it} = (D_{it1}, \ldots, D_{itK})$$

  - 各ユニット(u含む)はi.i.d. : $\{Y_i, D_i, u_i\} \overset{i.i.d.}{\sim} \mathbb{P}_0$, ただし

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{iT} \end{pmatrix},\ D_i = \begin{pmatrix} D_{i11} & \cdots & D_{i1K} \\ \vdots & \ddots & \vdots \\ D_{iT1} & \cdots & D_{iTK} \end{pmatrix}$$

・考えるモデル

$$Y_{it} = \delta^T D_{it} + u_i + \varepsilon_{it}, \quad i = 1,\ldots,N; t = 1,\ldots,T$$

$$\text{where} \quad E(\varepsilon_i \mid D_i, u_i) = 0_{T\times 1}$$

・Pooled OLS　#未観測の交絡変数を無視してOLS

$$\hat{\delta} = \arg\min_b \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - b^T D_{it})^2$$

- 交絡変数u_iを誤差項にpoolしている
- 興味のある変数D_iがu_iと相関 ➡ バイアスが生じる (why)

Fixed-effects (within) estimator

$$(\hat{\delta}, \hat{u}_1, \ldots, \hat{u}_N) = \underset{b, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - b^T D_{it} - m_i)^2$$

#交絡変数も最適化変数に入れて推定

- **First-order conditions**

$$\begin{cases} \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}(Y_{it} - \hat{\delta}^T D_{it} - \hat{u}_i) = 0_{K \times 1} \\ \sum_{t=1}^{T} (Y_{it} - \hat{\delta}^T D_{it} - \hat{u}_i) = 0, \quad i = 1, \ldots, N \end{cases}$$

$$\hat{u}_i = \frac{1}{T} \sum_{t=1}^{T} (Y_{it} - \hat{\delta}^T D_{it}) = \bar{Y}_i - \hat{\delta}^T \bar{D}_i$$

where $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{it}, \quad \bar{D}_i = \frac{1}{T} \sum_{t=1}^{T} D_{it}$

first-order conditionsの第1式

$$\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}(Y_{it} - \hat{\delta}^T D_{it} - \hat{u}_i) = 0 \;\Leftrightarrow\; \sum_{i=1}^{N}\sum_{t=1}^{T} (D_{it} - \bar{D}_i)(Y_{it} - \hat{\delta}^T D_{it} - \hat{u}_i) = 0$$

なぜなら第2式より $\sum_{i=1}^{N} \bar{D}_i \sum_{t=1}^{T} (Y_{it} - \hat{\delta}^T D_{it} - \hat{u}_i) = 0$

前スライドの$\hat{u}_i$を変形した青字の式に代入

$$\sum_{i=1}^{N}\sum_{t=1}^{T} (D_{it} - \bar{D}_i)\{Y_{it} - \bar{Y}_i - \hat{\delta}^T (D_{it} - \bar{D}_i)\} = 0$$

$$\hat{\delta} = \left\{ \sum_{i=1}^{N}\sum_{t=1}^{T} (D_{it} - \bar{D}_i)(D_{it} - \bar{D}_i)^T \right\}^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} (D_{it} - \bar{D}_i)(Y_{it} - \bar{Y}_i)$$

$$= \left(\sum\sum \ddot{D}_{it}\ddot{D}_{it}^T \right)^{-1} \sum\sum \ddot{D}_{it}\ddot{Y}_{it}$$

where $\ddot{D}_{it} = D_{it} - \bar{D}_i, \quad \ddot{Y}_{it} = Y_{it} - \bar{Y}_i$

Fixed-effects (within) estimator

$$\hat{\delta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{D}_{it} \ddot{D}_{it}^{T} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \ddot{D}_{it} \ddot{Y}_{it}$$

- **Notes**
  - $\ddot{Y}$ を $\ddot{D}$ へと回帰した形になっている
  - Pooled OLSとの大きな違い → **時点平均を最初に除いている**
    - これによって**未観測の交絡変数の影響を除去!**

$$Y_{it} = \delta^T D_{it} + u_i + \varepsilon_{it} \quad \Rightarrow \quad \bar{Y}_i = \delta^T \bar{D}_i + u_i + \bar{\varepsilon}_i$$

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_i = \delta^T \ddot{D}_{it} + \ddot{\varepsilon}_{it}$$

Rでやってみよう

用いるデータとモデル

・スライド2のデータを用いる

・労働組合参加や労働市場参加年数が賃金に影響を与えるか?

$$\texttt{lwage}_{it} = \delta_1 \texttt{union}_{it} + \delta_2 \texttt{exper}_{it} + u_i + \varepsilon_{it}$$

・個人の能力(未観測)をu_iで処理しておく

・データ解析の説明 → URL

Difference-in-Differences

Motivating data

- **John Snow医師によるコレラの感染経路分析**
  - コレラの感染経路は空気 or 井戸水?
  - 19世紀半ばで因果推論はまだ存在しない時代
    - Difference-in-Differences法 (DiD法) によく似た方法で特定
- **データの収集方法**
  - 水の供給会社：Southwark and Vauxhall (SV) or Lambeth
    - SVはテムズ川から水を取得
      - 結果から言うとテムズ川の水が汚染されていた
    - Lambethはテムズ川の**上流**から取得(1849年~)
      - 1849以前はテムズ川から取得
  - 2つの供給会社 × 2時点のパネルデータ
    - 対象地域の変更はなし → **“空気”は一定**

パネルデータとDiD推定値

・表内の数値は10000人あたりの感染者数

	1849年	1854年
Southwark and Vauxhall	135	147
Lambeth	85	**19**

$$\texttt{DiD est.} = \underbrace{(19 - 85)}_{\textbf{before and after}} - \underbrace{(147 - 135)}_{\textbf{before and after}} = -78$$

difference

なぜ単純な差の分析ではダメなのか?

- **1954年(処置後)における会社間比較**
  - D : 処置変数 (clean water)
  - L, SV : 未観測の交絡変数
    - 前述のuに該当するもの
  - Y : 結果変数 (感染者数)

	1849年	1854年
Southwark and Vauxhall	135	147
Lambeth	85	19

	1854年
Southwark and Vauxhall	$Y_{SV} = SV$
Lambeth	$Y_L = L + D$

個体間の差

$$D_{indi} = Y_L - Y_{SV} = D + \underbrace{(L - SV)}_{\text{観測不可}}$$

なぜ単純な差の分析ではダメなのか?

- **Lambeth社における時点間比較**
  - T : 時間効果
    - コレラの繁殖力など

	1849年	1854年
Southwark and Vauxhall	135	147
Lambeth	85	19

	Lambeth
Before	$Y_{L,B} = L$
After	$Y_{L,A} = L + T + D$

時点間の差

$$D_{time} = Y_{L,A} - Y_{L,B} = D + T$$

観測不可

Difference-in-differences analysis (差の差分析)

	Time	Outcome	After - Bofore
Lambeth	Before	L	
	After	$L+T+D$	$T+D$
Southwark and Vauxhall	Before	SV	
	After	$SV+T$	T

差を取れば
Dを得る！

- **重要な仮定**
  - 未観測の交絡変数L, SVは時間依存しない
  - 時間効果Tは個体(供給会社)に依存しない
    - **平行トレンド仮定(parallel trends assumption)**と呼ばれる

もう少し数理的に定式化

- **Settings for unit i**
  - $T = 1$: after treatment, $T = 0$: before treatment
  - $D_i = 1$: treated group, $D_i = 0$: control, $D_i(t) = D_i I(T = t)$
  - $Y_i^d(t)$: potential outcome for $T = t$ and $D_i(t) = d$
  - $Y_i(t)$: observation when $T = t$

- **SUTVA**
  - $Y_i(t) = D_i(t)Y_i^1(t) + (1 - D_i(t))Y_i^0(t)$
    - 処置前(T=0)においては$Y_i(0) = Y_i^0(0)$ $\because D_i(0) = 0$

- **Goal : ATTの推定** $E\{Y_i^1(1) - Y_i^0(1) \mid D_i = 1\}$
  - 独立なユニットを仮定 ➡ 以後添字iは省略

スライド12の表との関係

	1849年 (T=0)	1854年 (T=1)
Southwark and Vauxhall (D=0)	$E\{Y(0) \mid D=0\}$	$E\{Y(1) \mid D=0\}$
Lambeth (D=1)	$E\{Y(0) \mid D=1\}$	$E\{Y(1) \mid D=1\}$

SUTVA

	1849年 (T=0)	1854年 (T=1)
Southwark and Vauxhall (D=0)	$E\{Y^0(0) \mid D=0\}$	$E\{Y^0(1) \mid D=0\}$
Lambeth (D=1)	$E\{Y^0(0) \mid D=1\}$	$E\{Y^1(1) \mid D=1\}$

$$\mathrm{DiD} = E\{Y(1) \mid D = 1\} - E\{Y(0) \mid D = 1\} - \left[E\{Y(1) \mid D = 0\} - E\{Y(0) \mid D = 0\}\right]$$

#処置群におけるafter-beforeから対照群における同一のものを引く

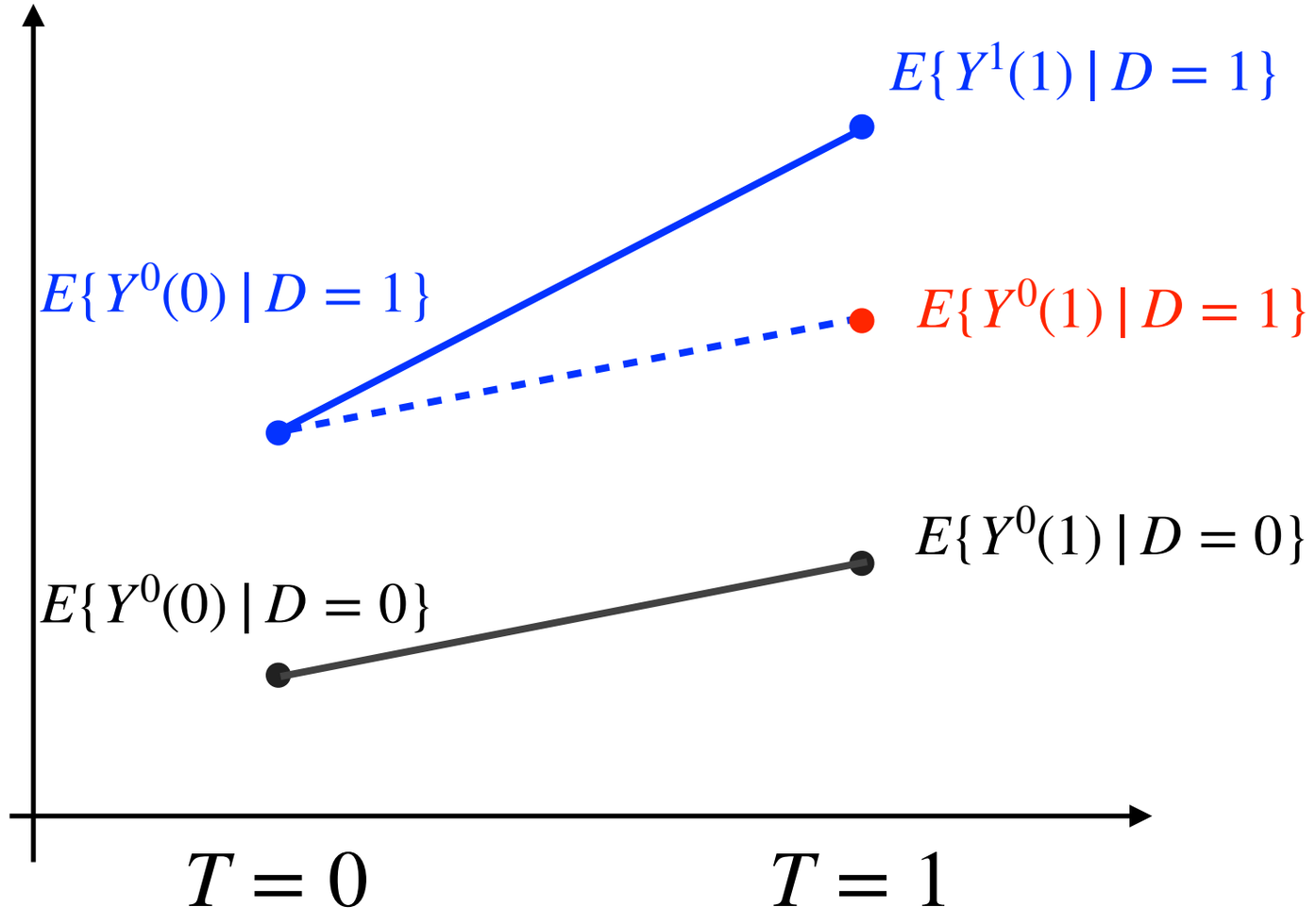
$$\begin{aligned}
\mathrm{DiD} &= E\{Y^1(1) \mid D = 1\} - E\{Y^0(0) \mid D = 1\} \\
&\quad - \left[E\{Y^0(1) \mid D = 0\} - E\{Y^0(0) \mid D = 0\}\right] \\
&= E\{Y^1(1) \mid D = 1\} - E\{Y^0(1) \mid D = 1\} \quad \text{➡ ATT} \\
&\quad + \left[E\{Y^0(1) \mid D = 1\} - E\{Y^0(0) \mid D = 1\}\right] \\
&\quad - \left[E\{Y^0(1) \mid D = 0\} - E\{Y^0(0) \mid D = 0\}\right]
\end{aligned}$$

#ATTを推定するために消えて欲しい

- **平行トレンド仮定**
  - DiD = ATTであるためには次の仮定が必要

$$E\{Y^0(1) \mid D=1\} - E\{Y^0(0) \mid D=1\}$$
$$= E\{Y^0(1) \mid D=0\} - E\{Y^0(0) \mid D=0\}$$

青線：D=1
黒線：D=0



DiDの推定

$$\mathrm{DiD} = E\{Y(1) \mid D = 1\} - E\{Y(0) \mid D = 1\} - \big[E\{Y(1) \mid D = 0\} - E\{Y(0) \mid D = 0\}\big]$$

$$\widehat{\mathrm{DiD}} = \frac{1}{N_1}\sum_{D_i=1} Y_i(1) - \frac{1}{N_1}\sum_{D_i=1} Y_i(0) - \left[\frac{1}{N_0}\sum_{D_i=0} Y_i(1) - \frac{1}{N_0}\sum_{D_i=0} Y_i(0)\right]$$

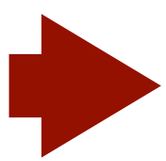
where $N_d = \sum_{D_i=d} 1$

単純なモデル化

- **スライド15の例との関係**
  - $\texttt{Comp}_i$: ユニットiが契約している会社 (ダミー変数)
  - $\texttt{Time}_t$: 時点のダミー変数, $\texttt{Time}_1 = 1, \texttt{Time}_0 = 0$

$$\begin{cases} Y_i^0(t) = \beta_0 + \beta_1 \texttt{Comp}_i + \beta_2 \texttt{Time}_t + \varepsilon_{it} \\ Y_i^1(t) = Y_i^0(t) + \delta D_i(t) \end{cases}$$

- 仮定 : $E(\varepsilon_{it} \mid D_i(1)) = 0$



$$Y_i(t) = \beta_0 + \beta_1 \texttt{Comp}_i + \beta_2 \texttt{Time}_t + \delta D_i(t) + \varepsilon_{it}$$

SUTVA

このとき $\mathrm{DiD} = \delta$ ➡ **回帰分析でも推定できる！**

・次の表を埋めて示してみよう (βとδで書き直す)

	1849年 (T=0)	1854年 (T=1)
Southwark and Vauxhall (D=0)		
Lambeth (D=1)		

#参考

	1849年 (T=0)	1854年 (T=1)
Southwark and Vauxhall (D=0)	$E\{Y(0) \mid D = 0\}$	$E\{Y(1) \mid D = 0\}$
Lambeth (D=1)	$E\{Y(0) \mid D = 1\}$	$E\{Y(1) \mid D = 1\}$

参考

- **回帰モデルに共変量を追加することも可能**

$$Y_i(t) = \beta_0 + \beta_1 \texttt{Comp}_i + \beta_2 \texttt{Time}_t + \delta D_i(t) + \theta^T X_{it} + \varepsilon_{it}$$

- Outcomeの説明力を上昇させる
- 仮定：$E(\varepsilon_{it} \mid D_i(1), X_{i0}, X_{i1}) = 0$

RでDiD

- **利用するデータ**
  - 最低賃金と雇用者数 (Card and Krueger (1994))
    - “最低賃金を増やすと雇用は減る”か検証
  - 地域 : New Jersey (NJ)とPennsylvania (PA)
    - NJは1992年に最低賃金を$4.25から$5.05へ引き上げ
    - PAは$4.25のままであり，NJに隣接している
- 時点
  - T=0 : 引き上げ前, T=1 : 引き上げ後
- 実際にデータ解析してみる → URL

平行トレンド仮定

- **どんなケースで崩れるか**
  - スライド21において$\beta_1 = \beta_{1t}$
    - 時間による効果が会社に依存している
    - Afterにおいて会社独自の感染対策が取られていた場合など

- **どうやって確認するのか**
  - 基本的にデータからは確認不可能
  - プラシーボテストはひとつの選択肢
    - Before期間においてDiDは0に近く推定されるはず
      - Before期間を分割する時点を選ぶ必要がある
    - 過去に変化がなければ未来も変化しないはずという仮説
      - 未来は誰にも分からない…

Synthetic control method

DiDの欠点

- **対照群をどう選択するか？**
  - コレラ分析では処置群に地理的に近い地区を選択
    - 客観的な判断であるとは言い難い…

- **Synthetic control methodでは…**
  - 処置を受けていない地区を複数用意できる場合
  - 複数の対照地区を結合して仮想的なひとつの地区を作る

具体的なデータ

- **California's Proposition 99**
  - CAは1988年にたばこ税を引き上げた
  - たばこの販売もいくらか制限された
    - 値段が上がるから需要が下がる or
    - 中毒性があるから需要は一定

- **データ**
  - 1970年-2000年までのたばこの売り上げ
  - カリフォルニアの他に38州からも入手
    - CA以外ではProposition 99は実施されていない
    - CA → 処置群; その他の38州 → 対照群
  - DiDを実行する場合は対照群からひとつ選ぶ必要がある

定式化

- $Y_i(t)$: 時点 t におけるユニット i の観測値 (たばこの売上)
  - $t = 1,\ldots, T_0, T_0+1,\ldots, T$
  - 時点 T_0+1 において介入が起きる (たばこ税の引き上げ)
  - $i = 1,2,\ldots, N;$ $i=1$を処置群とする (カリフォルニア)

処置時点

$$
\begin{array}{ll}
 & \\
\textbf{処置群} & \\
 & \\
\textbf{対照群} & \\
 &
\end{array}
\left(\begin{array}{cccc|ccc}
Y_1(1) & Y_1(2) & \cdots & Y_1(T_0) & Y_1(T_0+1) & \cdots & Y_1(T) \\
\hline
Y_2(1) & Y_2(2) & \cdots & Y_2(T_0) & Y_2(T_0+1) & \cdots & Y_2(T) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
Y_N(1) & Y_N(2) & \cdots & Y_N(T_0) & Y_N(T_0+1) & \cdots & Y_N(T)
\end{array}\right)
$$

- **SUTVA**
  - $Y_i(t) = D_i(t)Y_i^1(t) + (1 - D_i(t))Y_i^0(t)$
  - $D_i(t) = D_i I(t \geq T_0)$ **#時点tで処置されるかどうかを表す指示変数**

- **興味の対象**

$$E\{Y_i^1(t) - Y_i^0(t) \mid D_i = 1\}, \quad t \geq T_0 + 1$$

推定

$$E\{Y_i^1(t) \mid D_i = 1\} \Leftarrow Y_1(t)$$

$$E\{Y_i^0(t) \mid D_i = 1\} \Leftarrow \sum_{i=2}^{N} w_i^* Y_i(t)$$ **#対照群の重み和**

処置されなかった場合の仮想的な結果変数

重みをどうやって定めるか？

直感　処置前において処置群と対照群が似るように定める

$$\left(\begin{array}{cccc|ccc} Y_1(1) & Y_1(2) & \cdots & Y_1(T_0) & Y_1(T_0+1) & \cdots & Y_1(T) \\ Y_2(1) & Y_2(2) & \cdots & Y_2(T_0) & Y_2(T_0+1) & \cdots & Y_2(T) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ Y_N(1) & Y_N(2) & \cdots & Y_N(T_0) & Y_N(T_0+1) & \cdots & Y_N(T) \end{array}\right)$$

定式化

$$\sum_{t=1}^{T_0} \left\{ Y_1(t) - \sum_{i=2}^{N} w_i Y_i(t) \right\}^2 \to \text{minimize}$$

Option

- **共変量がある場合**
  - ただし処置に依存するものはダメ

$$\left(\begin{array}{cccccc|ccc} X_{11} & \cdots & X_{1L} & Y_1(1) & \cdots & Y_1(T_0) & \cdots & Y_1(T) \\ X_{21} & \cdots & X_{2L} & Y_2(1) & \cdots & Y_2(T_0) & \cdots & Y_2(T) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{NL} & \cdots & X_{NL} & Y_N(1) & \cdots & Y_N(T_0) & \cdots & Y_N(T) \end{array}\right)$$

$$\sum_{t=1}^{T_0}\left\{Y_1(t)-\sum_{i=2}^{N} w_i Y_i(t)\right\}^2+\sum_{\ell=1}^{L}\left\{X_{1\ell}-\sum_{i=2}^{N} w_i X_{i\ell}\right\}^2 \to \text{minimize}$$

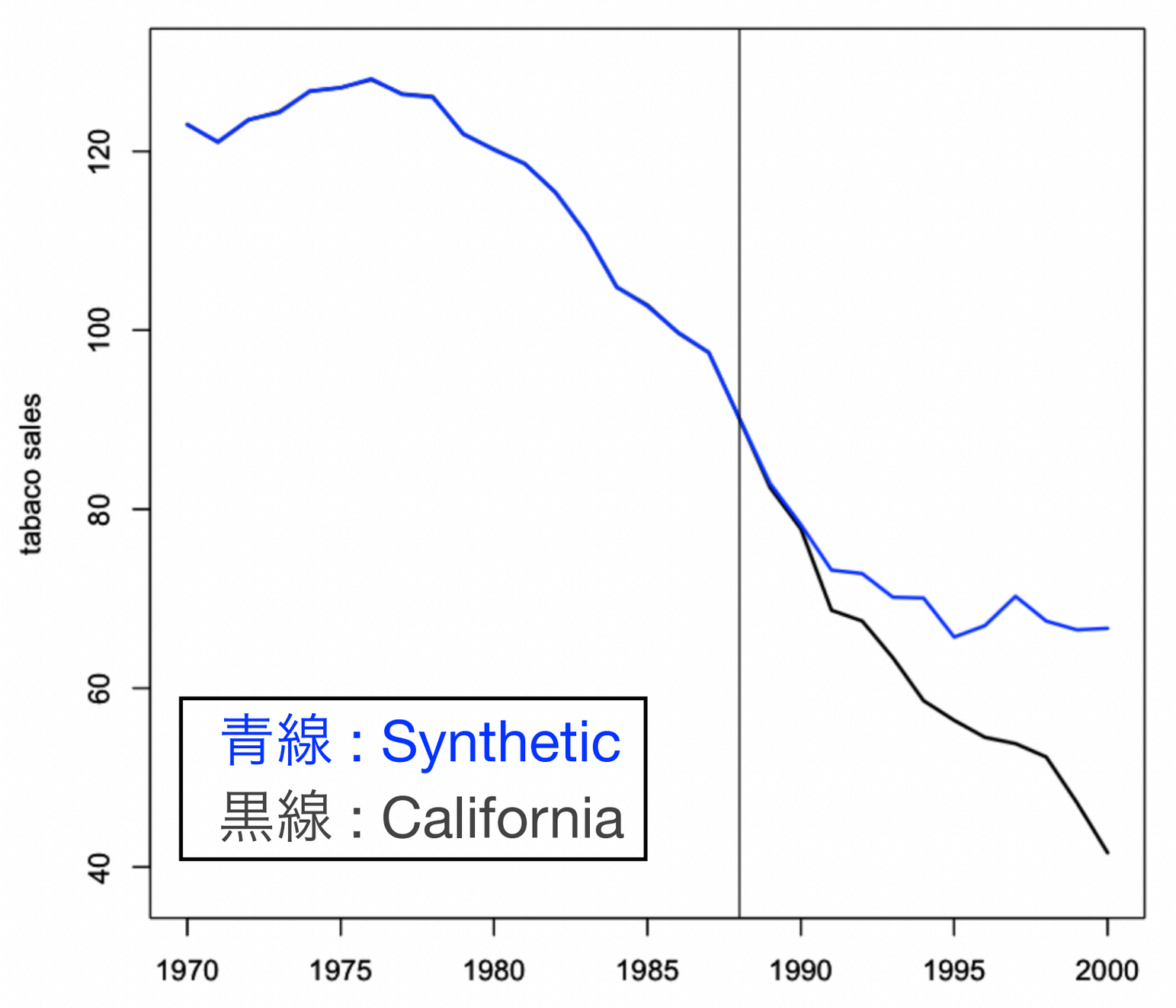
全て共変量とみなせば

$$\sum_{k=1}^{K} v_k\left\{X_{1k}-\sum_{i=2}^{N} w_i X_{ik}\right\}^2 \to \text{minimize}$$

v_k：どの共変量を重視するかの重み

Option

- **Note**
  - $\nu_k = 1$ならば重みの計算は単なる線形回帰
  - 結果変数をX_{1k}, 説明変数を$X_{2k}, \cdots, X_{Nk}$ とみなす
- **California's proposition 99**
  - 処置前期間1970-1988の売上のみ用いる (共変量なし)
  - $K = 18, N = 39$ だから係数パラメータの方が多い
  - とりあえずRidge回帰で実行してみる



処置前期間において過剰適合してしまっている...

凸包制約付きの最適化

$$\min \sum_{k=1}^{K} v_k \left\{ X_{1k} - \sum_{i=2}^{N} w_i X_{ik} \right\}^2$$

$$\text{subject to} \quad w_2, \ldots, w_N \geq 0 \quad \text{and} \quad \sum_{i=2}^{N} w_i = 1$$

ベクトル化 簡単のため$v_k = 1$とする

$$\boldsymbol{X}_1 = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1K} \end{pmatrix}, \quad \boldsymbol{X}_0 = \begin{pmatrix} X_{21} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{2K} & \cdots & X_{NK} \end{pmatrix}, \quad \boldsymbol{w} = \begin{pmatrix} w_2 \\ \vdots \\ w_N \end{pmatrix}$$

$$\sum_{k=1}^{K} \left\{ X_{1k} - \sum_{i=2}^{N} w_i X_{ik} \right\}^2 = \|\boldsymbol{X}_1 - \boldsymbol{X}_0 \boldsymbol{w}\|_2^2$$

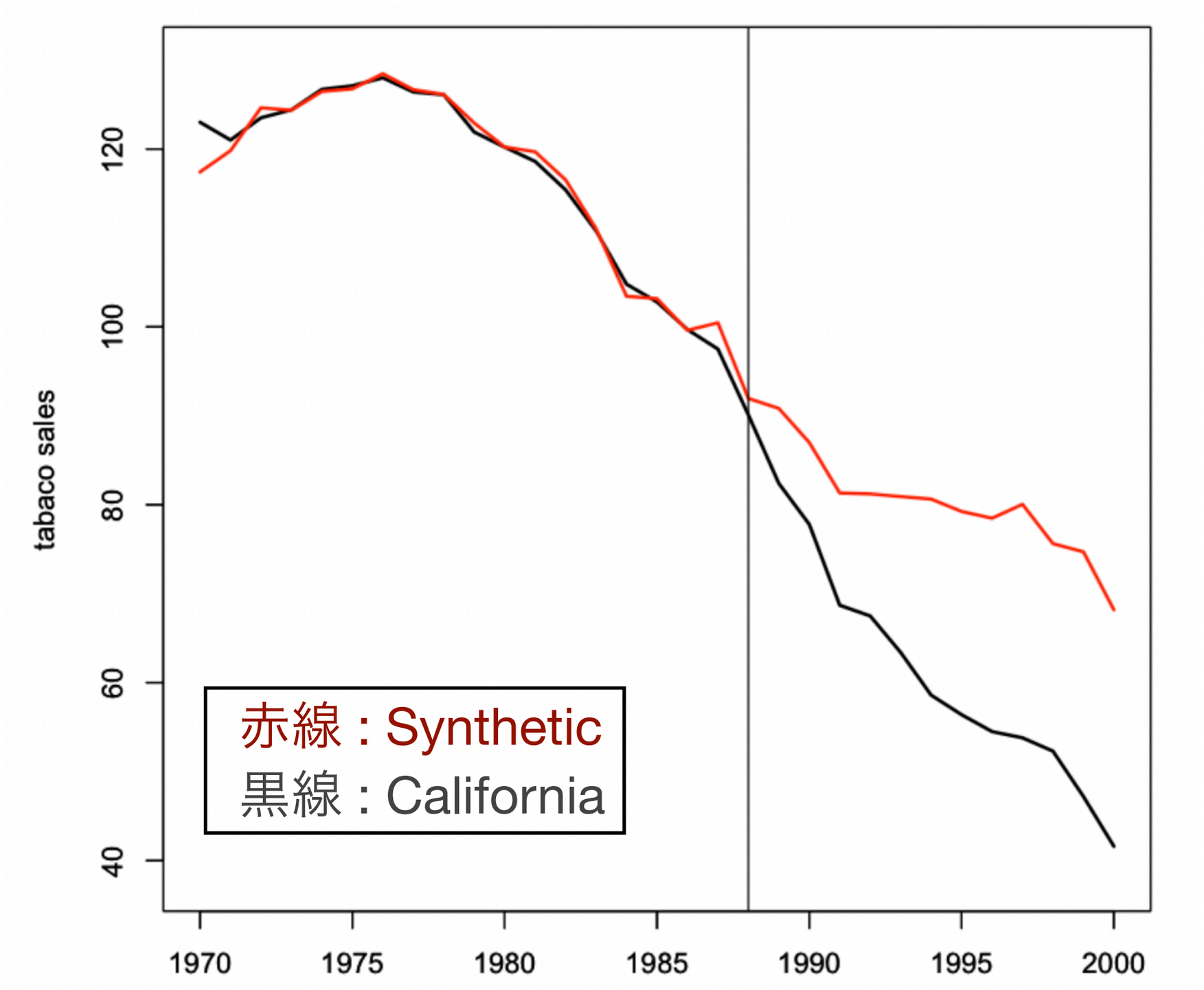
Rで制約付き最適化

- **凸2次計画問題を解くためのパッケージ**

```
> library(quadprog)
> solve.QP(Dmat,dvec,Amat,bvec,meq=1)
```

$$\min_{\boldsymbol{b}} \frac{1}{2}\boldsymbol{b}^T\boldsymbol{D}\boldsymbol{b} - \boldsymbol{d}^T\boldsymbol{b} \quad \texttt{subject to} \quad \boldsymbol{A}^T\boldsymbol{b} \geq \boldsymbol{b}_0$$

注意：“meq”は最初のいくつが等式制約かを指定する



過剰適合を回避し, 処置後にはっきりとした差を検出

有意性の検証

- **Fisherの正確検定**

1. N=39州のそれぞれを仮に処置群だと想定
2. 各データセット(N=39)についてSCを計算
3. 処置前期間におけるMSPEをそれぞれについて(N=39)計算

$$MSPE_{\texttt{before}} = \frac{1}{T_0}\sum_{t=1}^{T_0}\left\{Y_1(t) - \sum_{i=2}^{N}\hat{w}_i Y_i(t)\right\}^2$$

4. 処置後期間についても同様

$$MSPE_{\texttt{after}} = \frac{1}{T - T_0}\sum_{t=T_0+1}^{T}\left\{Y_1(t) - \sum_{i=2}^{N}\hat{w}_i Y_i(t)\right\}^2$$

5. 比 $MSPE_{\texttt{after}}/MSPE_{\texttt{before}}$ を計算
6. 計算した比を降順にソートし，CAの順位rを求める
7. p値：p=r/N を算出

主参考文献

Scott Cunningham. (2021). *Causal inference : The mixtape*, Yale University Press.

Bruce Hansen. (2022). *Econometrics*, Princeton University Press.

Matheus Facure Alves. Causal Inference for The Brave and True.
https://matheusfacure.github.io/python-causality-handbook/landing-page.html