# PS207 Quantitative Causal Inference

## Synthetic Controls + Panel Analysis

Matto Mildenberger

UC Santa Barbara

Special thanks to Chad Hazlett (UCLA) and Jens Hainmueller (Stanford) for slides

# Comparative Case Studies

Goal:

- Estimate effects of events or policy interventions that take place at an aggregate level (e.g., cities, states, countries).

Comparative Case Studies:

- Compare the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same aggregate for some control group (e.g. Card, 1990, Card and Krueger, 1994, Abadie and Gardeazabal, 2003).

# Comparative Case Studies

Advantages:

- Policy interventions often take place at an aggregate level
- Aggregate/macro data are often available

Problems:

- Selection of control group is often ambiguous
- Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

# Synthetic Control Method: A Motivating Model

- Suppose that we observe $J+1$ units in periods $1, 2, \ldots, T$.
- Region "one" is exposed to the intervention during periods $T_0+1, \ldots, T$.
- Let $Y_{it}^N$ be the outcome that would be observed for unit $i$ at time $t$ in the absence of the intervention.
- Let $Y_{it}^I$ be the outcome that would be observed for unit $i$ at time $t$ if unit $i$ is exposed to the intervention in periods $T_0+1$ to $T$.
- We aim to estimate the effect of the intervention on the treated unit $(\alpha_{1T_0+1}, \ldots, \alpha_{1T})$, where $\alpha_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$ for $t > T_0$.

# Synthetic Control Method: A Motivating Model

- Suppose that $Y_{it}^N$ is given by a factor model:

$$Y_{it}^N = \delta_t + Z_i\theta_t + \lambda_t\mu_i + \varepsilon_{it},$$

  - $\delta_t$ is an unobserved (common) time-dependent factor,
  - $Z_i$ is a $(1 \times r)$ vector of observed covariates,
  - $\theta_t$ is a $(r \times 1)$ vector of unknown parameters,
  - $\lambda_t$ is a $(1 \times F)$ vector of unknown common factors,
  - $\mu_i$ is a $(F \times 1)$ vector of unknown factor loadings,
  - $\varepsilon_{it}$ are unobserved transitory shocks.
- Specification allows heterogeneous responses to multiple unobserved factors.
- In contrast, the Difference-in-Differences (or Fixed-Effects) model restricts $\lambda_t$ to be constant.

# Synthetic Control Method: A Motivating Model

- The vector $Z_i$ may contain pre- and post-intervention values of time-varying variables, as long as they are not affected by the intervention.
- For example, if $T = 2$, $T_0 = 1$, $Z_i = (Z_{i1}, Z_{i2})$,

$$\theta_1 = \begin{pmatrix} \beta \\ 0 \end{pmatrix} \quad \text{and} \quad \theta_2 = \begin{pmatrix} 0 \\ \beta \end{pmatrix},$$

then $Z_i\theta_t = Z_{it}\beta$.

# Synthetic Control Method: A Motivating Model

- Let $W = (w_2, \ldots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \ldots, J+1$ and $w_2 + \cdots + w_{J+1} = 1$. Each value of $W$ represents a potential synthetic control
- Let $\bar{Y}_i^{K_1}, \ldots, \bar{Y}_i^{K_M}$ be $M$ linear functions of pre-intervention outcomes ($M \geq F$)
- Suppose that we can choose $W^*$ such that:

$$\sum_{j=2}^{J+1} w_j^* Z_j = Z_1, \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_1} = \bar{Y}_1^{K_1}, \cdots, \sum_{j=2}^{J+1} w_j^* \bar{Y}_j^{K_M} = \bar{Y}_1^{K_M}.$$

- Then (if $T_0$ is large relative to the scale of $\varepsilon_{it}$), an approximately unbiased estimator of $\alpha_{1t}$ is:

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

for $t \in \{T_0 + 1, \ldots, T\}$

# Synthetic Control Method: Implementation

- Let $X_1 = (Z_1, \bar{Y}_1^{K_1}, \ldots, \bar{Y}_1^{K_M})'$ be a $(k \times 1)$ vector of pre-intervention characteristics.
- Similarly, $X_0$ is a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^*$ is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints.
- We consider $\|X_1 - X_0 W\|_V = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$, where $V$ is some $(k \times k)$ symmetric and positive semidefinite matrix.
- Various ways to choose $V$ (subjective assessment of predictive power of $X$, regression, minimize MSPE, cross-validation, etc.).

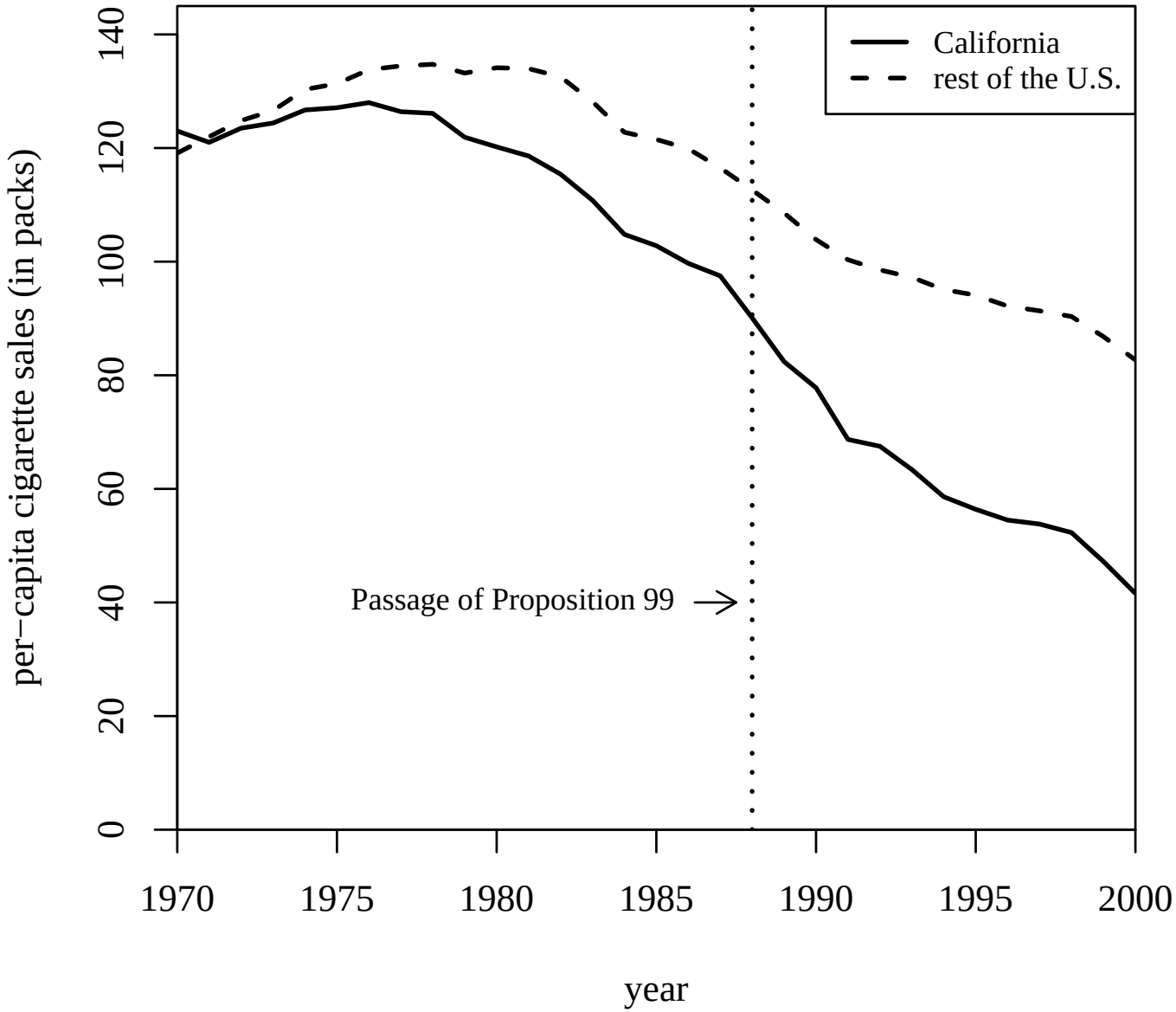
# The Application: California's Proposition 99

In 1988, California first passed comprehensive tobacco control legislation:

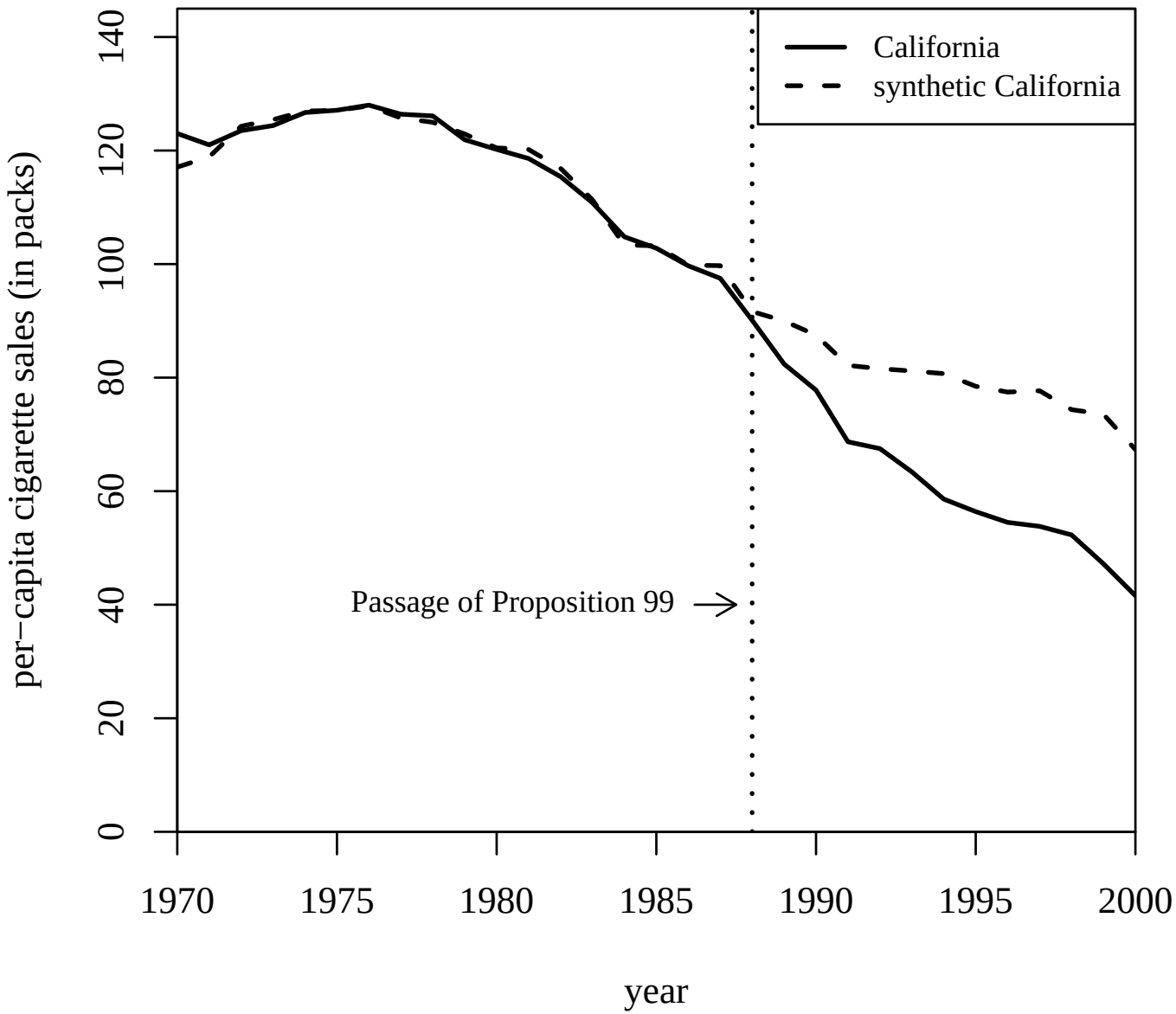
- increased cigarette tax by 25 cents/pack
- earmarked tax revenues to health and anti-smoking budgets
- funded anti-smoking media campaigns
- spurred clean-air ordinances throughout the state
- produced more than $100 million per year in anti-tobacco projects

Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HA, MA, MD, MI, NJ, NY, OR, WA, DC)

# Cigarette Consumption: CA and the Rest of the U.S.



# Cigarette Consumption: CA and synthetic CA

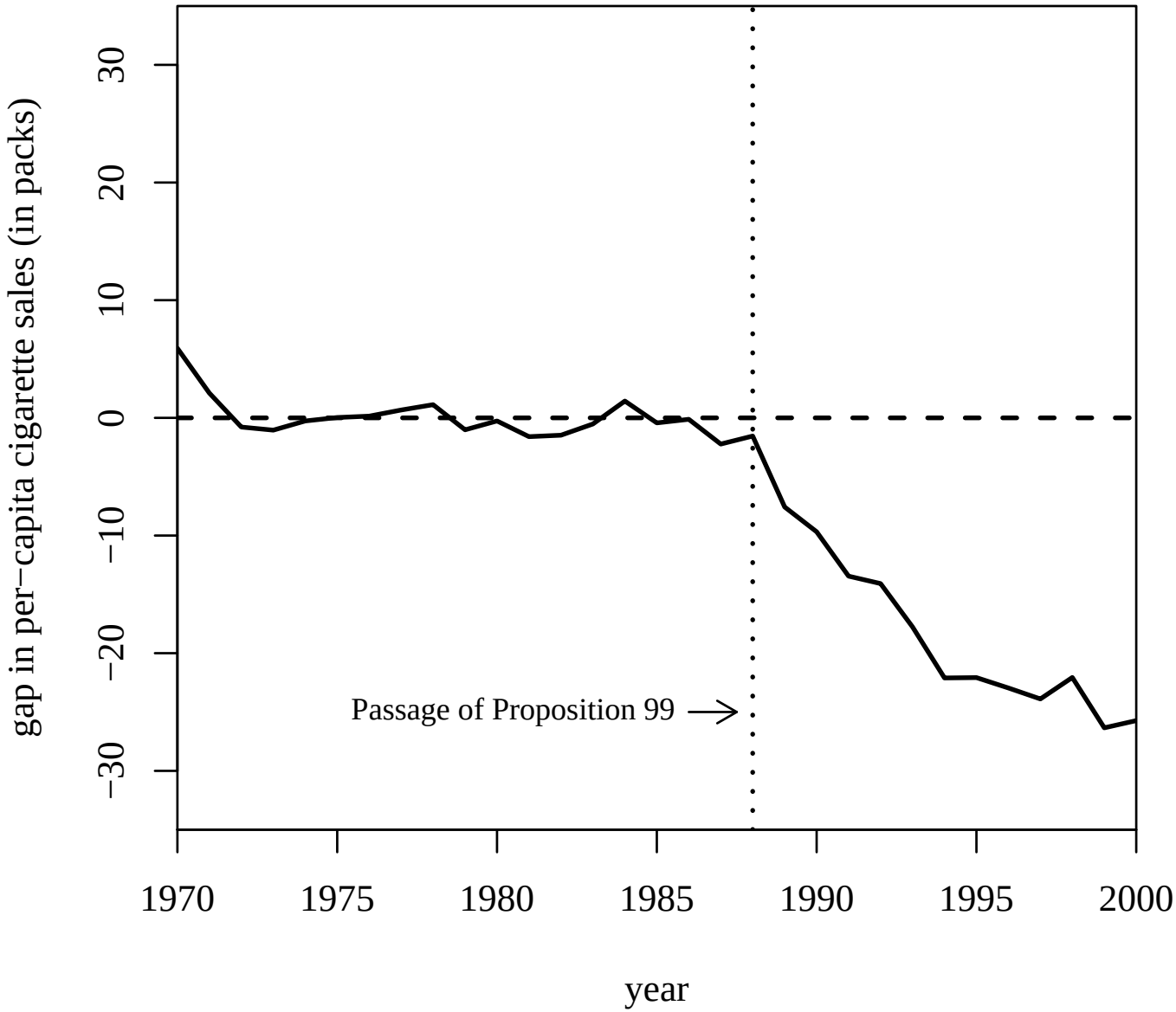


# Predictor Means: Actual vs. Synthetic California

| | California | | Average of |
|---|---|---|---|
| Variables | Real | Synthetic | 38 control states |
| 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).

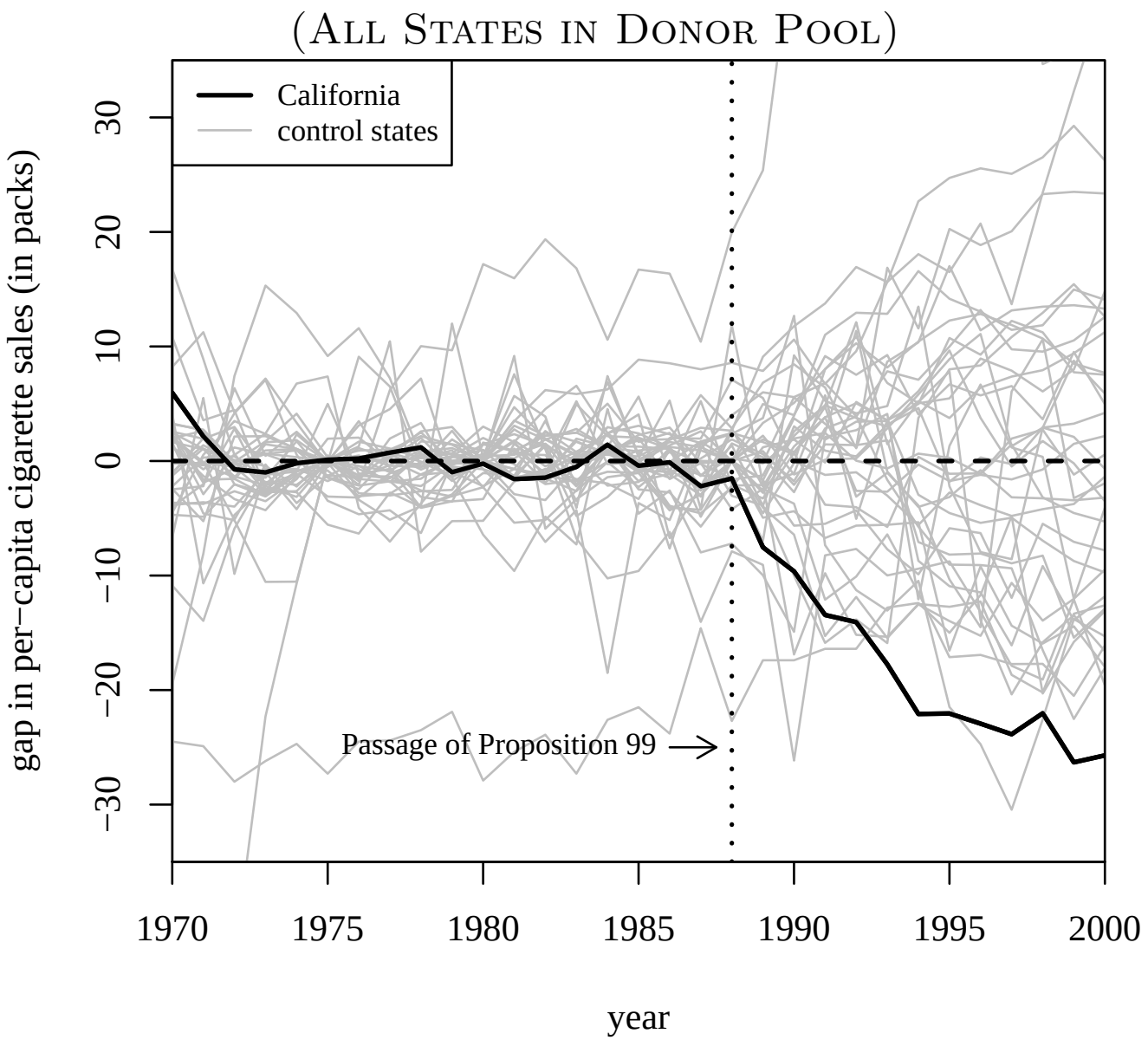
# Smoking Gap Between CA and synthetic CA



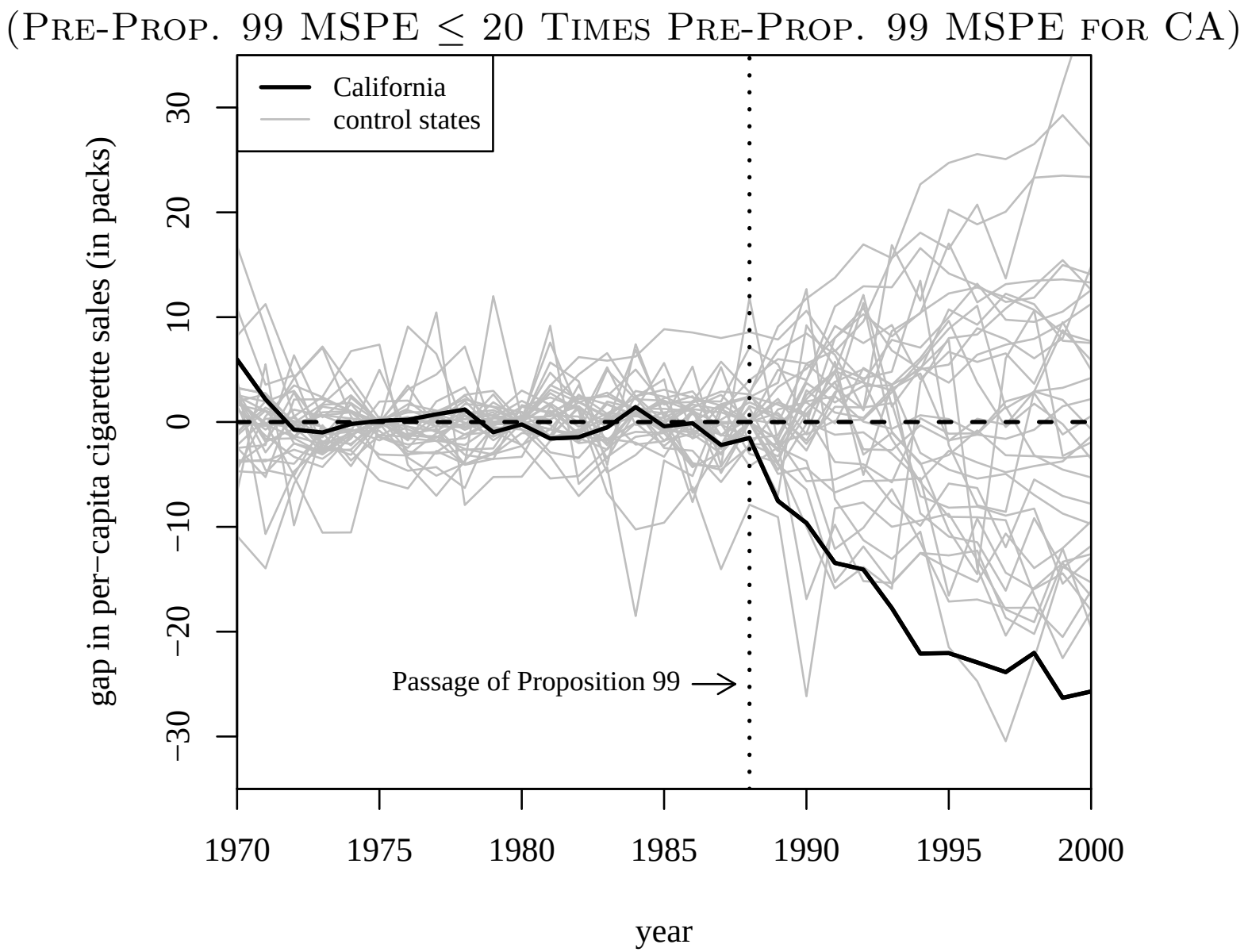
# Inference

- Iteratively apply the synthetic method to each state in the "donor pool" and obtain a distribution of placebo effects
- Compare the gap for California to the distribution of the placebo gaps.
- Question is whether the effect estimated by the synthetic control for the unit affected by the intervention is large relative to the effect estimated for a unit chosen at random.
- Valid inference regardless of the number of available comparison units, time periods, and whether the data are individual or aggregate.

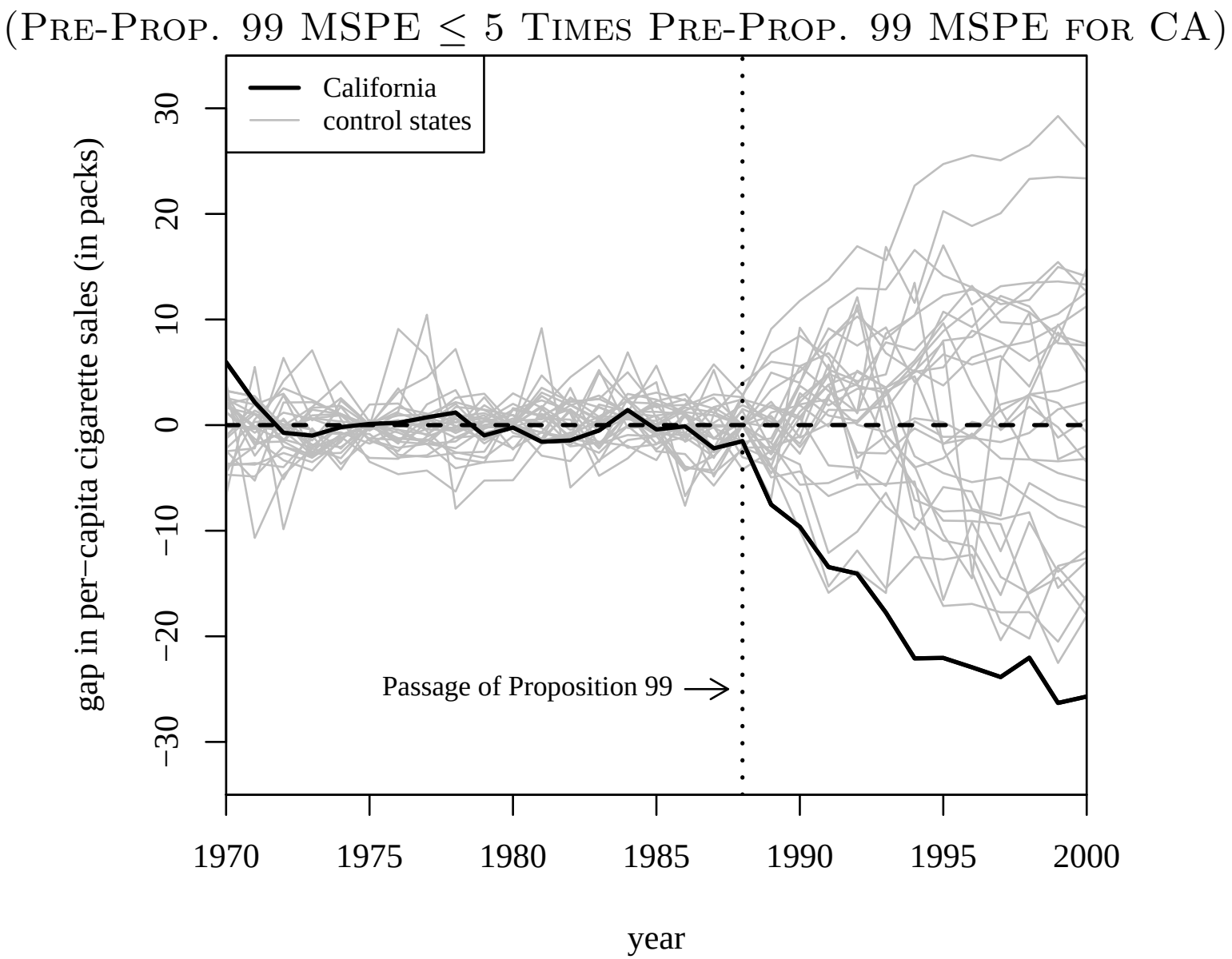
# Smoking Gap for CA and 38 control states



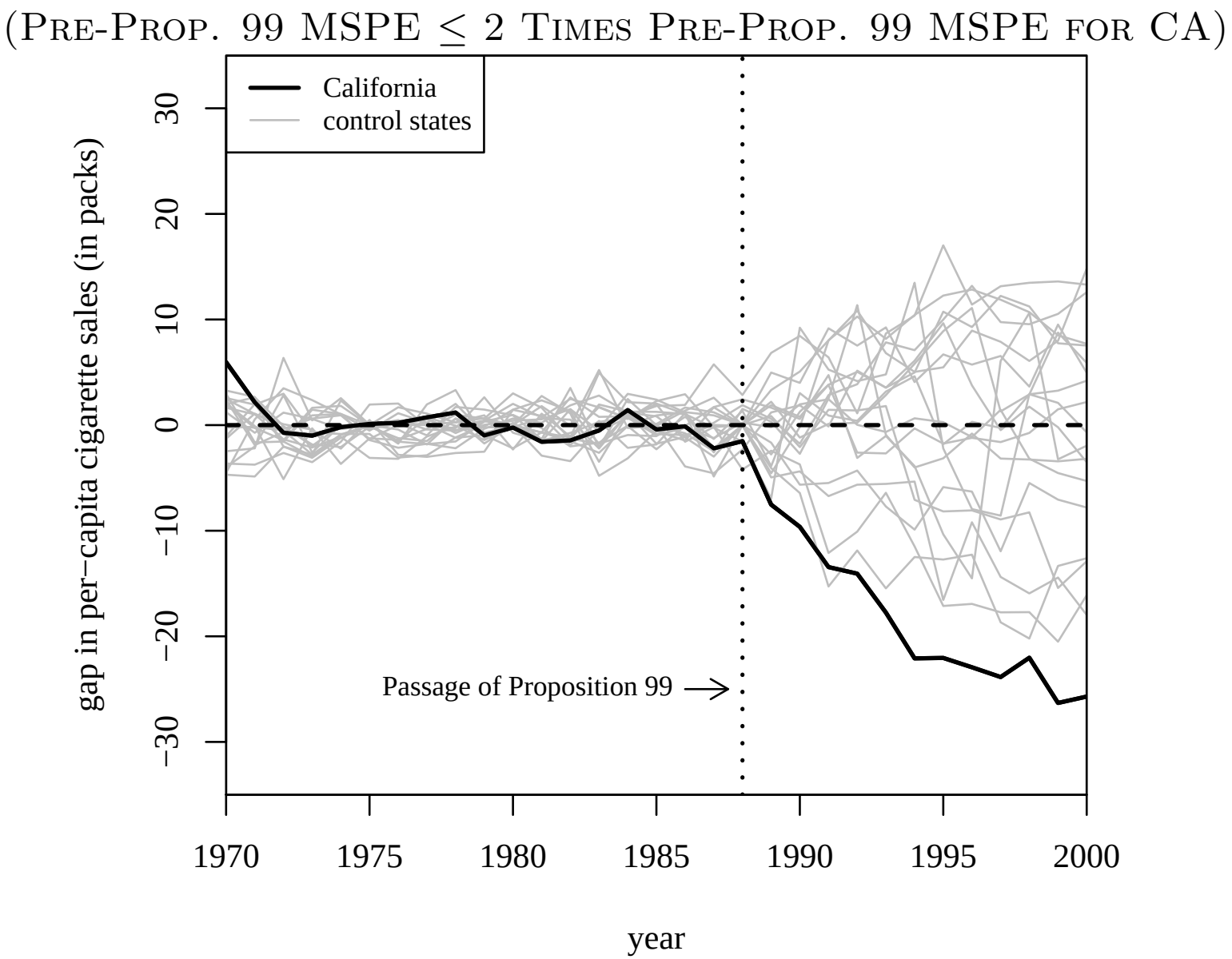
# Smoking Gap for CA and 34 control states



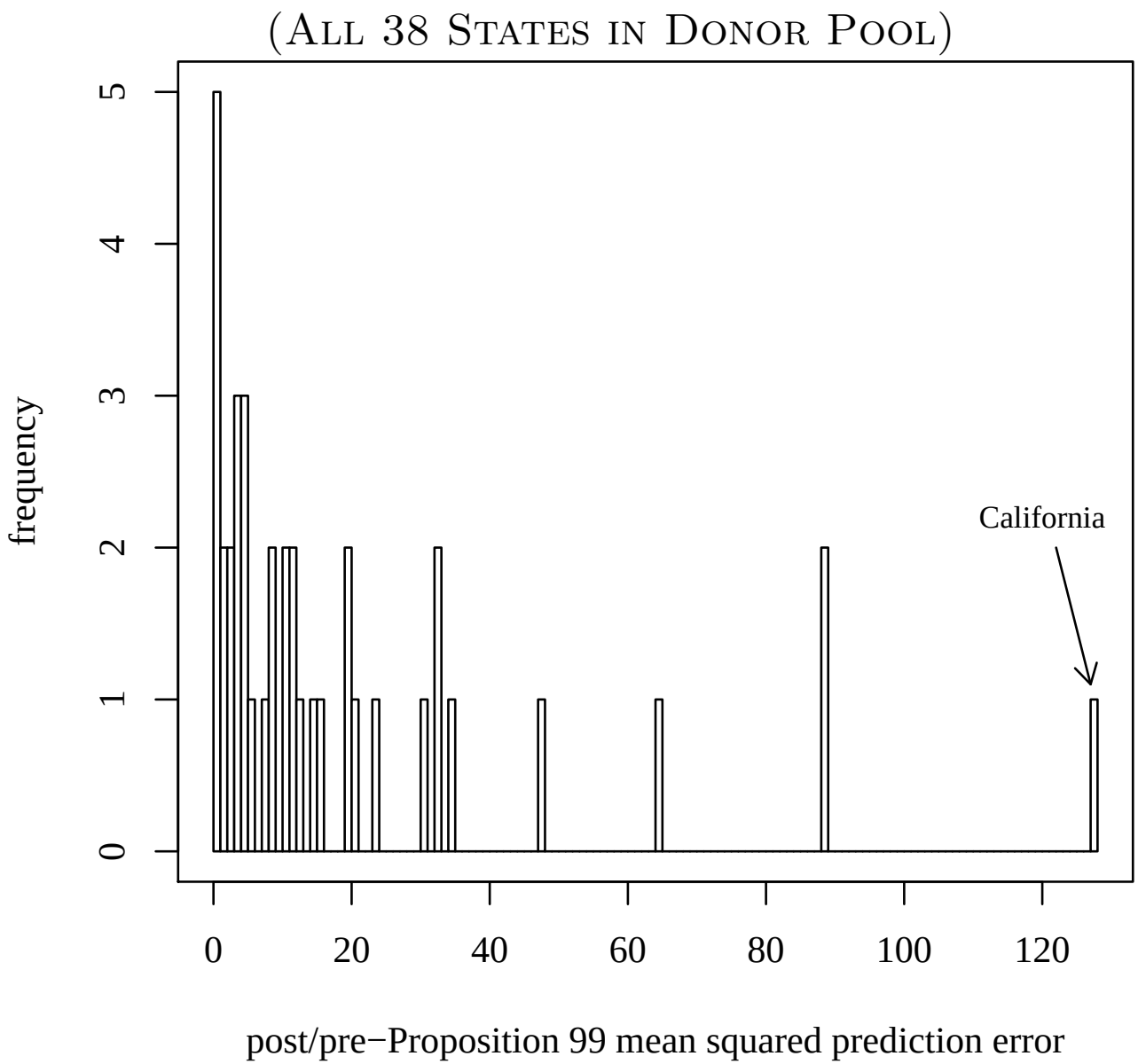
# Smoking Gap for CA and 29 control states



# Smoking Gap for CA and 19 control states



# Ratio Post-Prop. 99 MSPE to Pre-Prop. 99 MSPE



# An Application to Cross-Country Data

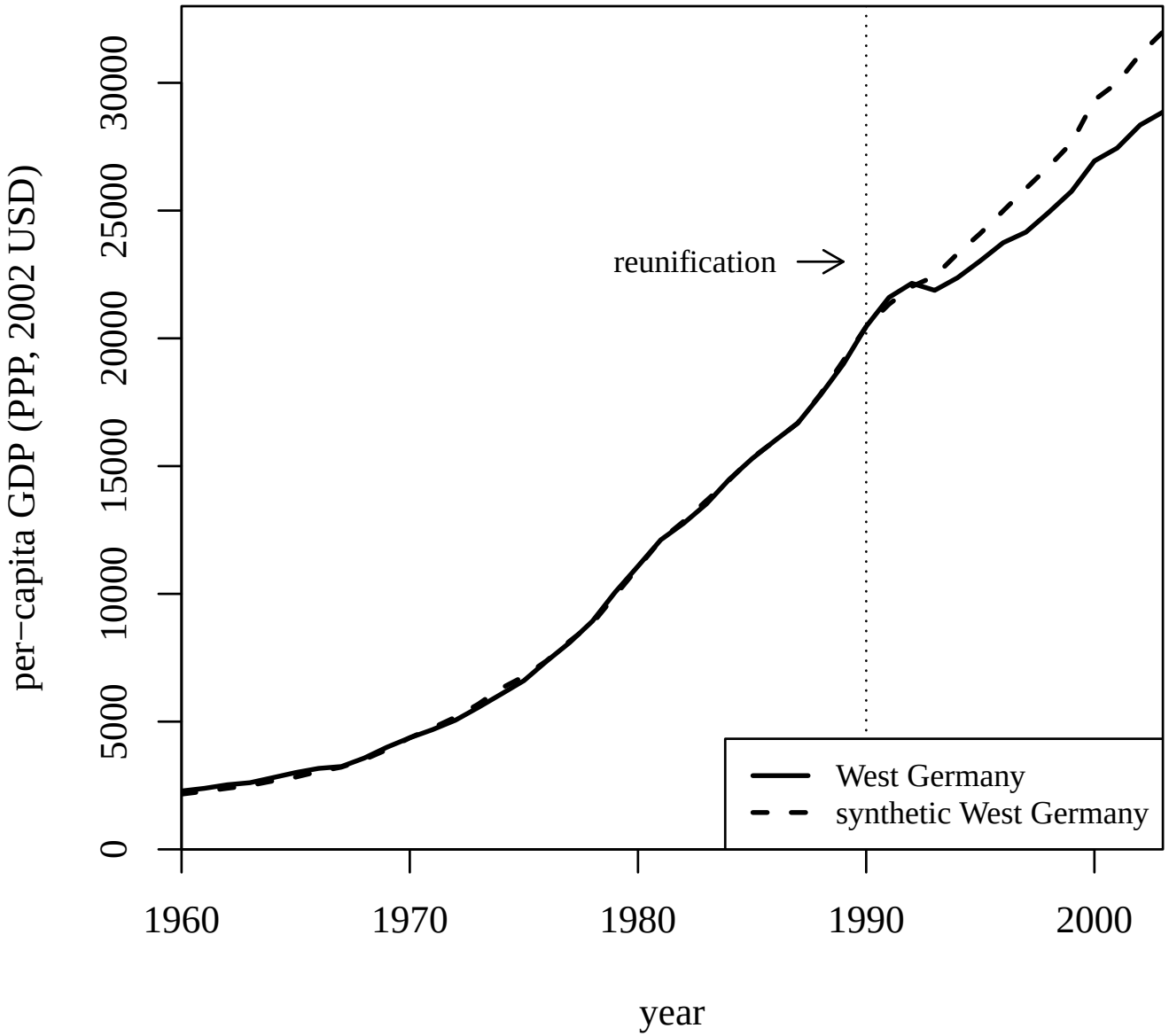
- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics.
- The synthetic control method provides an appealing data-driven procedure to study the effects of events or interventions that take place at the country level.
- Application: the economic impact of the 1990 German unification in West Germany.
- Donor pool is restricted to 21 OECD countries.

# Economic Growth Predictors Means

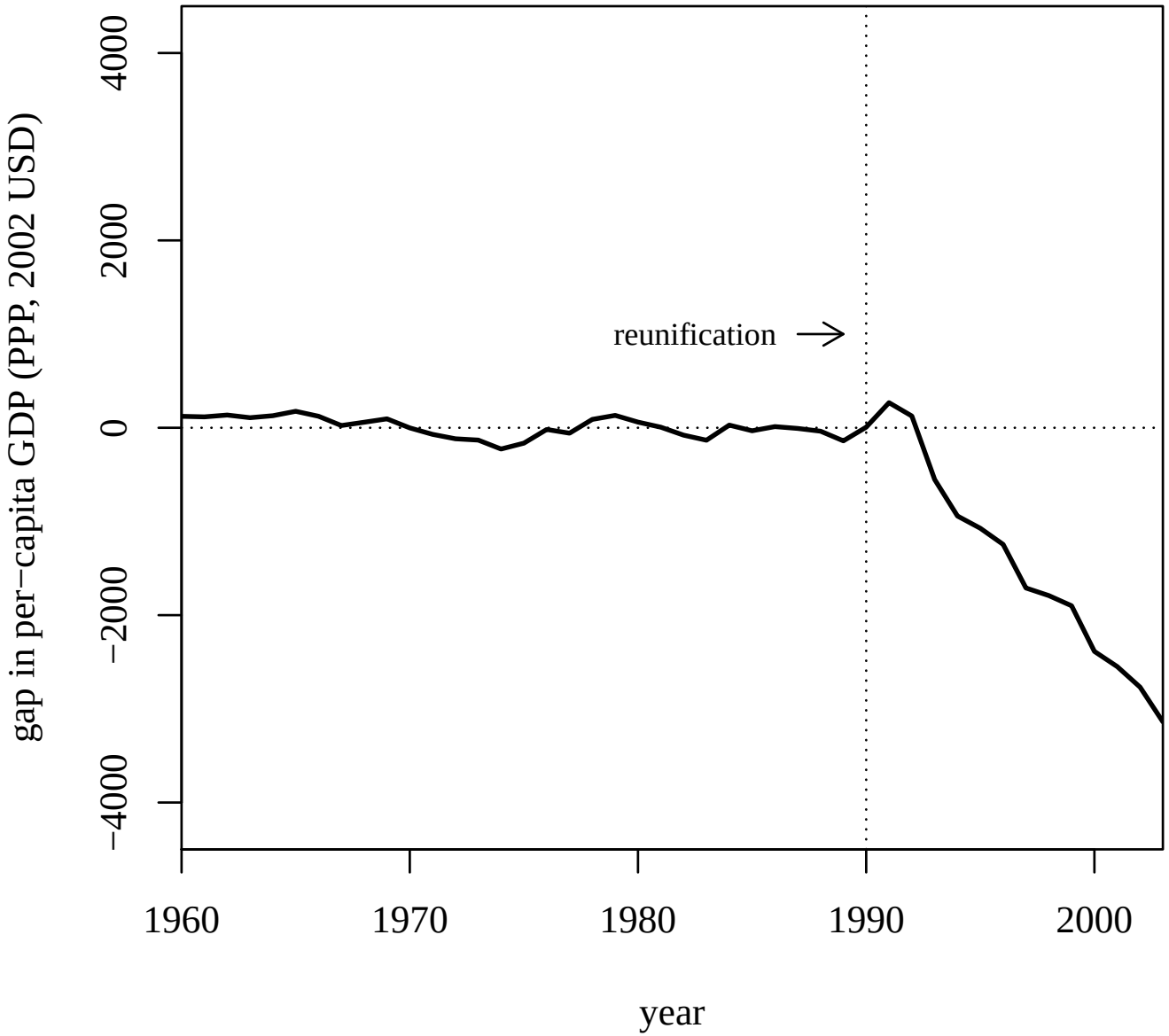
| | West Germany | Synthetic West Germany | OECD Sample excl. West Germany |
|---|---|---|---|
| GDP per-capita | 8169.8 | 8163.1 | 8049.3 |
| Trade openness | 45.8 | 54.4 | 32.6 |
| Inflation rate | 3.4 | 4.7 | 7.3 |
| Industry share | 34.7 | 34.7 | 34.3 |
| Schooling | 55.5 | 55.6 | 43.8 |
| Investment rate | 27.0 | 27.1 | 25.9 |

*Note:* GDP, inflation rate, and trade openness are averaged for the 1960–1989 period. Industry share is averaged for the 1980–1989 period. Investment rate and schooling are averaged for the 1980–1985 period.

# West Germany and synthetic West Germany



# GDP Gap: West Germany and synthetic West Germany



# Country Weights in the Synthetic West Germany

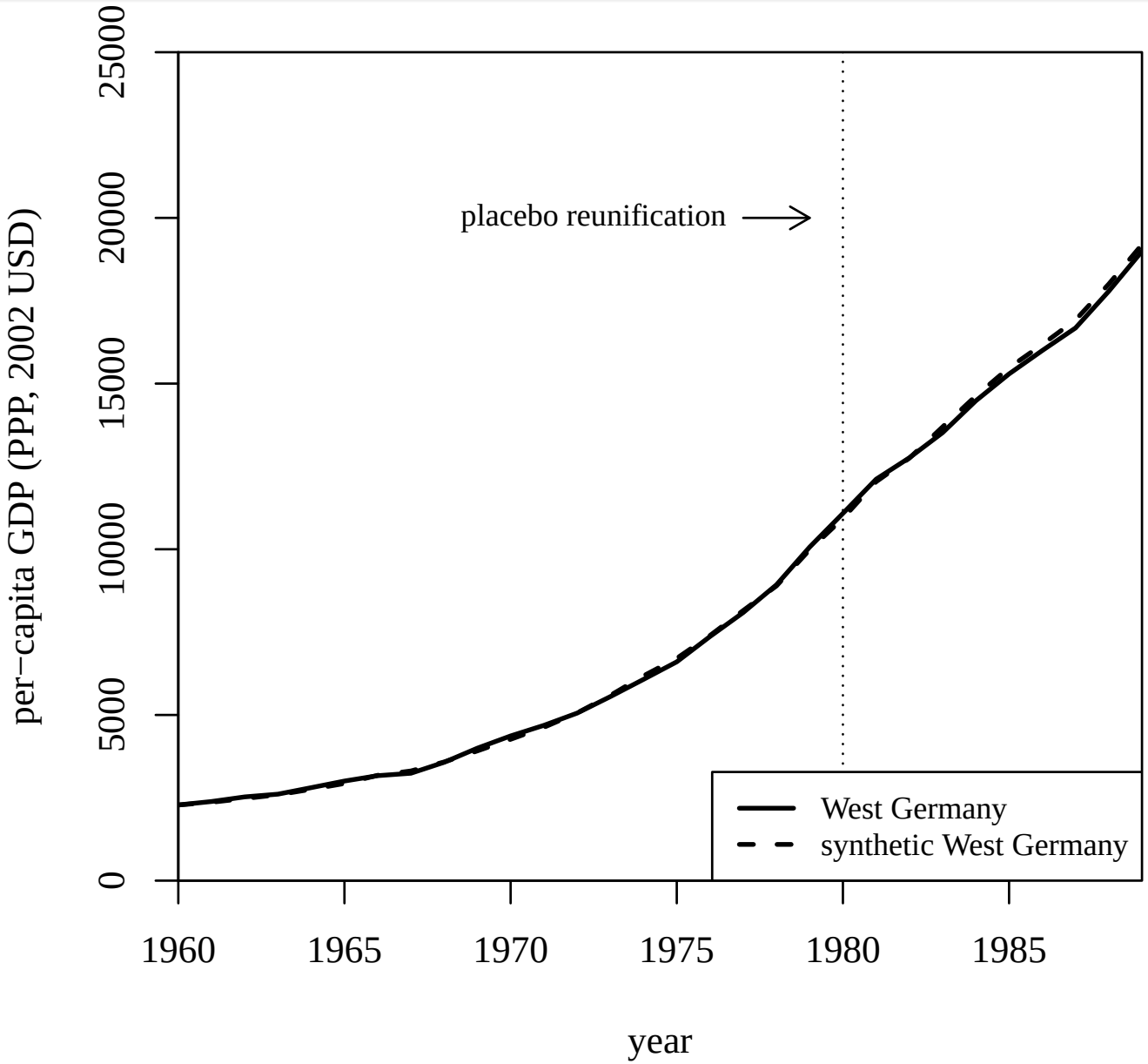
| Country | Weight | Country | Weight |
|---|---|---|---|
| Australia | 0 | Netherlands | 0.11 |
| Austria | 0.47 | New Zealand | 0.11 |
| Belgium | 0 | Norway | 0 |
| Canada | 0 | Portugal | 0 |
| Denmark | 0 | Spain | 0 |
| France | 0 | Sweden | 0 |
| Greece | 0 | Switzerland | 0 |
| Ireland | 0 | United Kingdom | 0.17 |
| Italy | 0 | United States | 0 |
| Japan | 0 | | 0.14 |

# Country Weights in the Synthetic West Germany

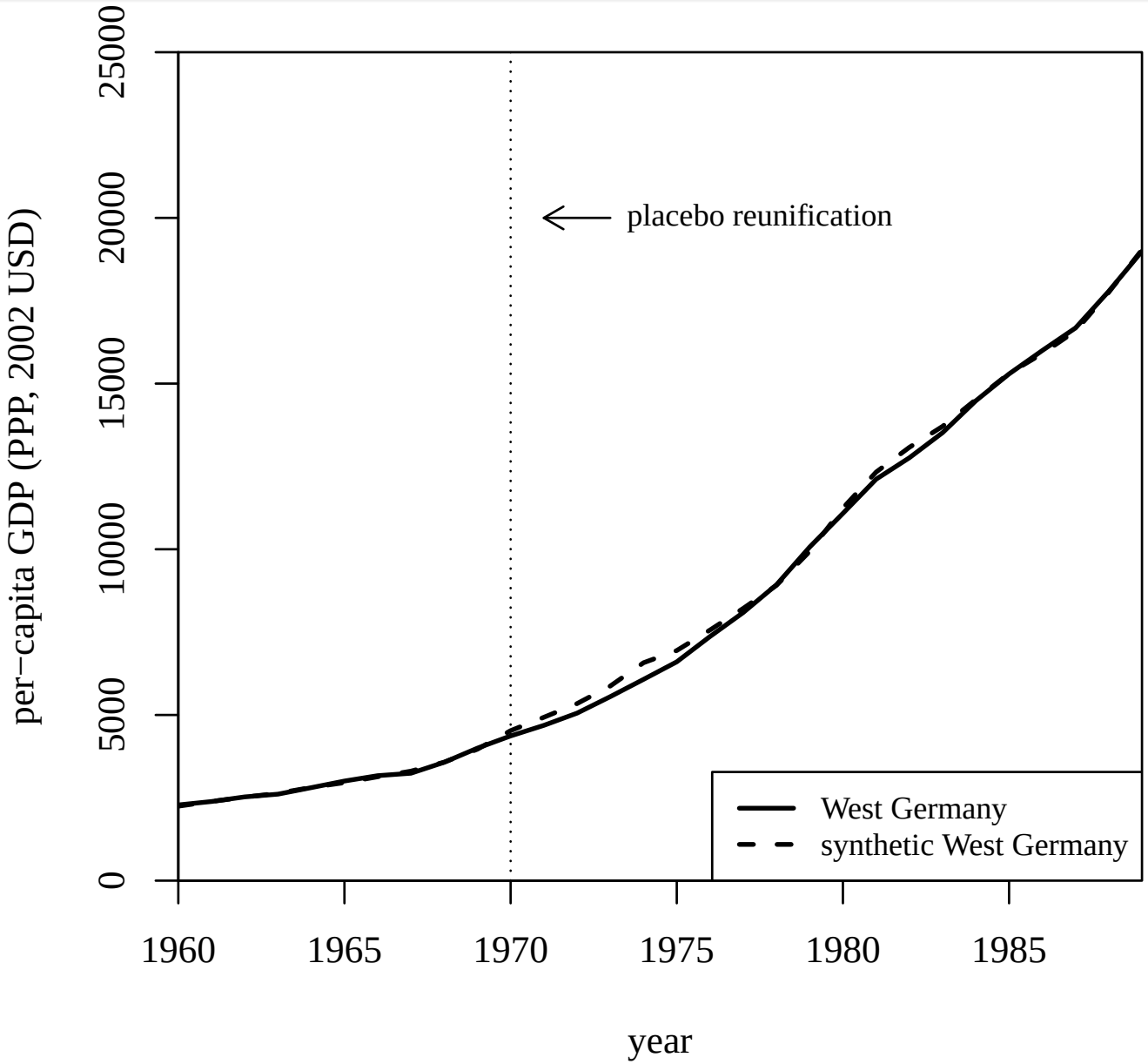
| Country | Synthetic Weight | Regression Weight | Country | Synthetic Weight | Regression Weight |
|---|---|---|---|---|---|
| Australia | 0 | 0.1 | Netherlands | 0.11 | 0.18 |
| Austria | 0.47 | 0.33 | New Zealand | 0 | -0.08 |
| Belgium | 0 | 0.1 | Norway | 0 | -0.07 |
| Canada | 0 | 0.09 | Portugal | 0 | -0.14 |
| Denmark | 0 | 0.04 | Spain | 0 | 0 |
| France | 0 | 0.16 | Switzerland | 0.17 | -0.06 |
| Greece | 0 | 0.02 | United Kingdom | 0 | -0.04 |
| Italy | 0 | -0.17 | United States | 0.14 | 0.21 |
| Japan | 0.11 | 0.32 | | | |

*Note:* Synthetic Weight: Unit weight assigned by the synthetic control method. Regression Weight: Unit weight assigned by linear regression.

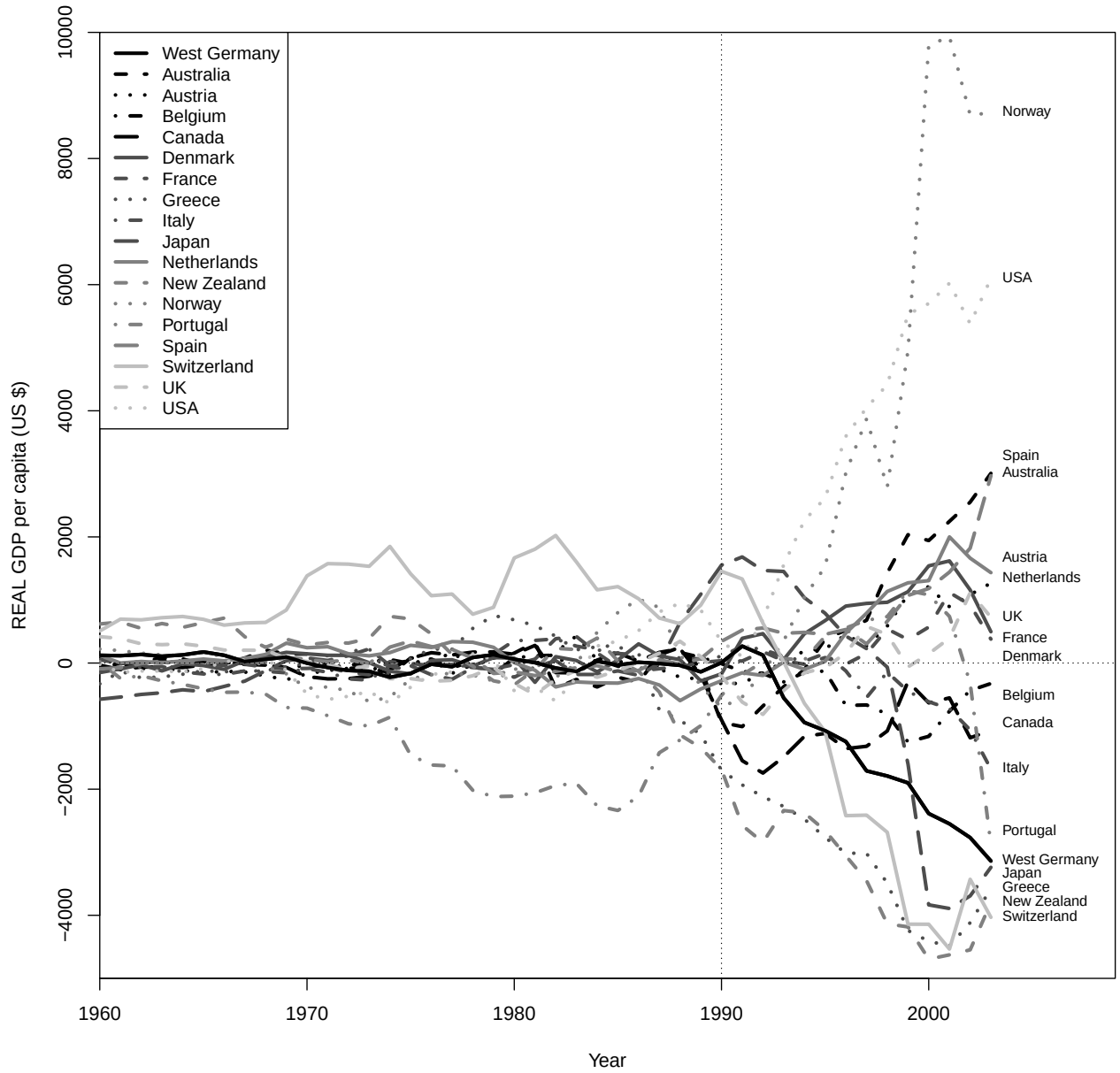
# Placebo Reunification 1980



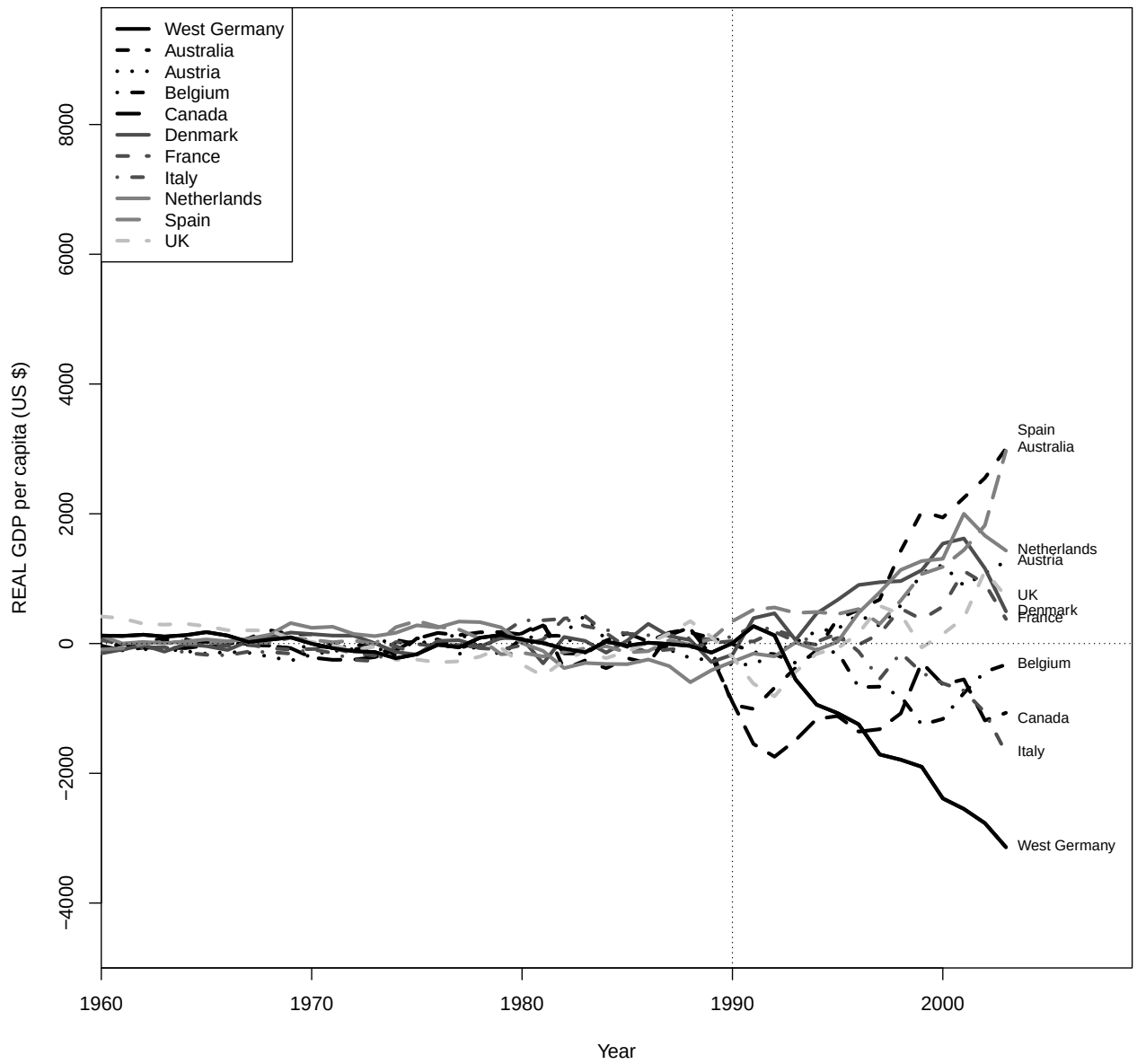
# Placebo Reunification 1970



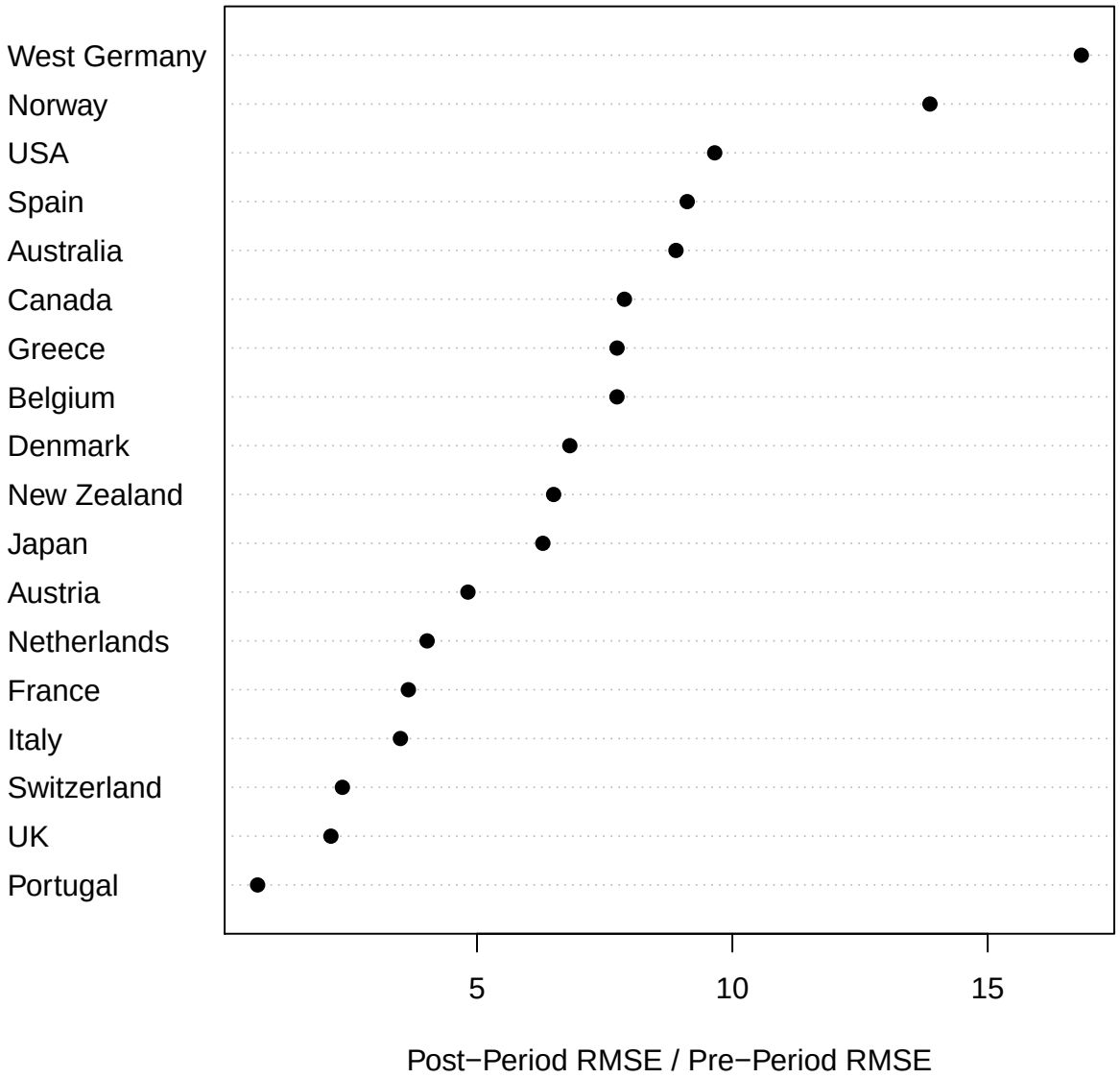
# Per-Capita GDP gaps in West Germany and placebo gaps



# Per-Capita GDP gaps in West Germany and placebo gaps



# Ratio of post- and pre-reunification MSPE



# Panel Setup

- Let $y$ and $\mathbf{x} \equiv (x_1, x_2, ..., x_K)$ be observable random variables and $c$ be an unobservable random variable
- We are interested in the partial effects of variable $x_j$ in the population regression function

$$E[y|x_1, x_2, ..., x_K, c]$$

- We observe a sample of $i = 1, 2, ..., N$ cross-sectional units for $t = 1, 2, ...T$ time periods (a balanced panel)
  - For each unit $i$, we denote the observable variables for all time periods as $\{(y_{it}, \mathbf{x}_{it}) : t = 1, 2, ..., T\}$
  - $\mathbf{x}_{it} \equiv (x_{it1}, x_{it2}, ..., x_{itK})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{\mathbf{y}_i, \mathbf{x}_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
  - $\mathbf{y}_i \equiv (y_{i1}, y_{i2}, ..., y_{iT})'$ and $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$
  - Consider asymptotic properties with $T$ fixed and $N \to \infty$

# Panel Setup

Single unit:

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \qquad \mathbf{X}_i = \begin{pmatrix} x_{i,1,1} & x_{i,1,2} & x_{i,1,j} & \cdots & x_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{i,t,1} & x_{i,t,2} & x_{i,t,j} & \cdots & x_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{i,T,1} & x_{i,T,2} & x_{i,T,j} & \cdots & x_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_i \\ \vdots \\ \mathbf{y}_N \end{pmatrix}_{NT \times 1} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_N \end{pmatrix}_{NT \times K}$$

# Unobserved Effects Model: Farm Output

- For a randomly drawn cross-sectional unit $i$, the model is given by

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - $y_{it}$: output of farm $i$ in year $t$
  - $\mathbf{x}_{it}$: $1 \times K$ vector of variable inputs for farm $i$ in year $t$, such as labor, fertilizer, etc. plus an intercept
  - $\boldsymbol{\beta}$: $K \times 1$ vector of marginal effects of variable inputs
  - $c_i$: farm effect, i.e. the sum of all time-invariant inputs known to farmer $i$ (but unobserved for the researcher), such as soil quality, managerial ability, etc.
    - often called: unobserved effect, unobserved heterogeneity, etc.
  - $\varepsilon_{it}$: time-varying unobserved inputs, such as rainfall, unknown to the farmer at the time the decision on the variable inputs $\mathbf{x}_{it}$ is made
    - often called: idiosyncratic error
- What happens when we regress $y_{it}$ on $\mathbf{x}_{it}$?

# Pooled OLS

- When we ignore the panel structure and regress $y_{it}$ on $\mathbf{x}_{it}$ we get

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}, \qquad t = 1, 2, ..., T$$

with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- Main assumption to obtain consistent estimates for $\boldsymbol{\beta}$ is:
  - $E[v_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}] = E[v_{it}|\mathbf{x}_{it}] = 0$ for $t = 1, 2, ..., T$
    - $\mathbf{x}_{it}$ are strictly exogenous: the composite error $v_{it}$ in each time period is uncorrelated with the past, current, and future regressors
    - But: labour input $\mathbf{x}_{it}$ likely depends on soil quality $c_i$ and so we have omitted variable bias and $\hat{\boldsymbol{\beta}}$ is not consistent
  - No correlation between $\mathbf{x}_{it}$ and $v_{it}$ implies no correlation between unobserved effect $c_i$ and $\mathbf{x}_{it}$ for all $t$
    - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)

# Unobserved Effects Model: Program Evaluation

- Program evaluation model:

$$y_{it} = prog_{it}\,\beta + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - $y_{it}$: log wage of individual $i$ in year $t$
  - $prog_{it}$: indicator coded 1 if individual $i$ participants in training program at $t$ and 0 otherwise
  - $\beta$: effect of program
  - $c_i$: sum of all time-invariant unobserved characteristics that affect wages, such as ability, etc.
- What happens when we regress $y_{it}$ on $prog_{it}$? $\hat{\beta}$ not consistent since $prog_{it}$ is likely correlated with $c_i$ (e.g. ability)
- Always ask: Is there a time-constant unobserved variable ($c_i$) that is correlated with the regressors? If yes, pooled OLS is problematic
- Additional problem: $v_{it} \equiv c_i + \varepsilon_{it}$ are serially correlated for same $i$ since $c_i$ is present in each $t$ and thus pooled OLS standard errors are invalid

# Fixed Effect Regression

- Our unobserved effects model is:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- If we have data on multiple time periods, we can think of $c_i$ as fixed effects or "nuisance parameters" to be estimated

- OLS estimation with fixed effects yields:

$$(\widehat{\boldsymbol{\beta}}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{\boldsymbol{b}, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \mathbf{x}_{it}\boldsymbol{b} - m_i)^2$$

this amounts to including $N$ farm dummies in regression of $y_{it}$ on $\mathbf{x}_{it}$

# Derivation: Fixed Effects Regression

$$(\widehat{\boldsymbol{\beta}}, \widehat{c}_1, \ldots, \widehat{c}_N) = \underset{\boldsymbol{b}, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \mathbf{x}_{it}\boldsymbol{b} - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}' (y_{it} - \mathbf{x}_{it}\widehat{\boldsymbol{\beta}} - \widehat{c}_i) = 0$$

and

$$\sum_{t=1}^{T} (y_{it} - \mathbf{x}_{it}\widehat{\boldsymbol{\beta}} - \widehat{c}_i) = 0$$

for $i = 1, \ldots, N$.

# Derivation: Fixed Effects Regression

Therefore, for $i = 1, \ldots, N$,

$$\widehat{c}_i = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}\widehat{\boldsymbol{\beta}}) = \bar{y}_i - \bar{\mathbf{x}}_i\widehat{\boldsymbol{\beta}},$$

where

$$\bar{\mathbf{x}}_i \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}, \qquad \bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T}y_{it}.$$

Plug this result into the first FOC to obtain:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'(y_{it} - \bar{y}_i)\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right)$$

with time-demeaned variables $\ddot{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$.

# Fixed Effects Regression

Running a regression with the time-demeaned variables $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_i$ is numerically equivalent to a regression of $y_{it}$ on $\mathbf{x}_{it}$ and unit specific dummy variables.

Fixed effects estimator is often called the within estimator because it only uses the time variation within each cross-sectional unit.

Even better, the regression with the time-demeaned variables is consistent for $\beta$ even when $Cov[\mathbf{x}_{it}, c_i] \neq 0$, because time-demeaning eliminates the unobserved effects:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}\beta + c_i + \varepsilon_{it} \\ \bar{y}_i &= \bar{\mathbf{x}}_i\beta + c_i + \bar{\varepsilon}_i \end{aligned}$$

---

$$\begin{aligned} (y_{it} - \bar{y}_i) &= (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\beta + (c_i - c_i) + (\varepsilon_{it} - \bar{\varepsilon}_i) \\ \ddot{y}_{it} &= \ddot{\mathbf{x}}_{it}\beta + \ddot{\varepsilon}_{it} \end{aligned}$$

# Fixed Effects Regression: Main Results

- Identification assumptions:
  1. $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0, \;\; t = 1, 2, ..., T$
     - regressors are strictly exogenous conditional on the unobserved effect
     - allows $\mathbf{x}_{it}$ to be arbitrarily related to $c_i$
  2. $rank(\sum_{t=1}^{T} E[\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}]) = K$
     - regressors vary over time for at least some $i$ and are not collinear
- Fixed effects estimator:
  1. Demean and regress $\ddot{y}_{it}$ on $\ddot{\mathbf{x}}_{it}$ (need to correct degrees of freedom)
  2. Regress $y_{it}$ on $\mathbf{x}_{it}$ and unit dummies (dummy variable regression)
  3. Regress $y_{it}$ on $\mathbf{x}_{it}$ with canned fixed effects routine
     - R: `plm(y~x , model = within, data = data)`
- Properties (under assumptions 1-2):
  - $\hat{\beta}_{FE}$ is consistent: $\underset{N\to\infty}{\text{plim}}\, \hat{\beta}_{FE,N} = \beta$
  - $\hat{\beta}_{FE}$ also unbiased conditional on $\mathbf{X}$

# Fixed Effects Regression: Main Issues

- Inference:
  - Standard errors have to be "clustered" by panel unit (e.g. farm) to allow correlation in the $\varepsilon_{it}$'s for the same $i$.
    - R: `coeftest(mod, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))`
  - Yields valid inference as long as number of clusters is reasonably large

- Typically we care about $\boldsymbol{\beta}$, but unit fixed effects $c_i$ could be of interest
  - $\widehat{c}_i$ from dummy variable regression is unbiased but not consistent for $c_i$ (based on fixed $T$ and $N \to \infty$)
  - `plm` routine demeans the data before running the regression and therefore does not estimate $\widehat{c}_i$
    - intercept shows average $\widehat{c}_i$ across units.
    - we can recover $\widehat{c}_i$ using $\widehat{c}_i = \bar{y}_i - \bar{\mathbf{x}}_i \widehat{\boldsymbol{\beta}}$
    - `fixef(mod)`

# Example: Direct Democracy and Naturalizations

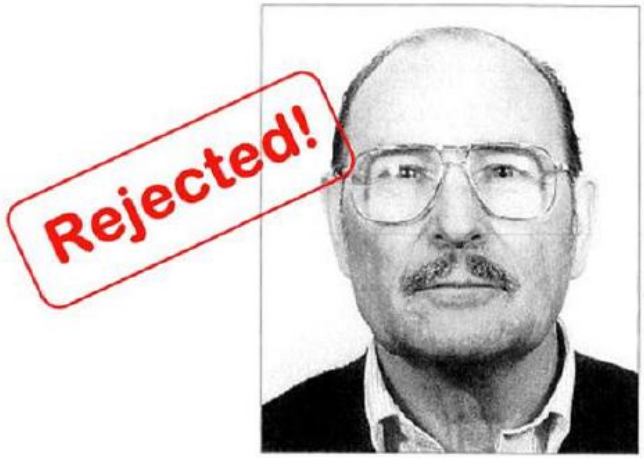
- Do minorities fare worse under direct democracy than under representative democracy?
- Hainmueller and Hangartner (2014) examine data on naturalization requests of immigrants in Switzerland, where municipalities vote on naturalization applications in:
  - referendums (direct democracy)
  - elected municipality councils (representative democracy)
- Annual panel data from 1,400 municipalities for the 1991-2009 period
  - $y_{it}$: naturalization rate = # naturalizations$_{it}$ / eligible foreign population $_{it-1}$
  - $x_{it}$: 1 if municipality used representative democracy, 0 if municipality used direct democracy in year $t$

# Naturalization Referenda

***Cardone Giuseppa,*** *italienische Staatsangehörige, Gerliswilstrasse 26, 6020 Emmenbrücke*


| | |
|---|---|
| **Geburtsort:** | Pietrelcina (I) |
| **Geburtsdatum:** | 9. Dezember 1939 |
| **Zivilstand:** | geschieden |
| **Ausbildung:** | Volksschule |
| **Bisherige Tätigkeiten:** | Mitarbeit auf elterlichem Bauerngut, Lingerie-Mitarbeiterin in Hotels |
| **Jetzige Tätigkeit:** | IV-Rentnerin seit 1997 |
| **Arbeitgeber:** | – |
| **Einreise in die Schweiz:** | 15. Oktober 1962 |
| **Zuzug nach Emmen:** | 23. September 1970 |
| **Hobbys:** | – |
| **Steuern:** | Steuerbares Einkommen Fr. 33 900.– Steuerbares Vermögen Fr. 28 000.– |
| **Kinder:** | – |
| **Einbürgerungstaxe:** | Fr. 123.– |
| **Einbürgerungsgebühr:** | Fr. 500.– |

***Deak Janos,*** *ungarischer Staatsangehöriger, Ghürschweg 13, 6020 Emmenbrücke*




| | |
|---|---|
| **Geburtsort:** | Bucsa (H) |
| **Geburtsdatum:** | 14. Mai 1936 |
| **Zivilstand:** | geschieden |
| **Ausbildung:** | Volksschule, Lehre als Mineur und Sprengmeister, Zusatzausbildung als Maler |
| **Bisherige Tätigkeiten:** | Bau-Hilfsarbeiter, selbstständiger Maler |
| **Jetzige Tätigkeit:** | IV-Rentner seit 1987 (Verkehrsunfall) |
| **Arbeitgeber:** | – |
| **Einreise in die Schweiz:** | 17. November 1956 |
| **Zuzug nach Emmen:** | 26. Juni 1991 |
| **Hobbys:** | Fischen, Pilze sammeln, Modellflugzeuge basteln |
| **Steuern:** | Steuerbares Einkommen Fr. 28 400.– Steuerbares Vermögen Fr. 0.– |
| **Kinder:** | – |
| **Einbürgerungstaxe:** | Fr. 100.– |
| **Einbürgerungsgebühr:** | Fr. 500.– |

# Naturalization Panel Data Long Format

```
> d <- read.dta("Swiss_Panel_long.dta")
> print(d[30:40,],digits=2)
   muniID      muni_name year nat_rate repdem
30      2 Affoltern A.A. 2001     3.21      0
31      2 Affoltern A.A. 2002     4.64      0
32      2 Affoltern A.A. 2003     4.84      0
33      2 Affoltern A.A. 2004     5.62      0
34      2 Affoltern A.A. 2005     4.39      0
35      2 Affoltern A.A. 2006     8.12      1
36      2 Affoltern A.A. 2007     7.07      1
37      2 Affoltern A.A. 2008     8.98      1
38      2 Affoltern A.A. 2009     6.12      1
39      3     Bonstetten 1991     0.83      0
40      3     Bonstetten 1992     0.84      0
```

# Pooled OLS

```
> summary(lm(nat_rate~repdem,data=d))

Call:
lm(formula = nat_rate ~ repdem, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-4.726 -2.223 -1.523  1.411 21.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.22268    0.06904   32.19   <2e-16 ***
repdem       2.50332    0.12907   19.39   <2e-16 ***
```

# Time-Demeaning for Fixed Effects: $y_{it} \rightarrow \ddot{y}_{it}$

```
> library(plyr)
> d <- ddply(d, .(muniID), transform,
+                         nat_rate_demean = nat_rate - mean(nat_rate),
+                         nat_rate_mean   = mean(nat_rate),
+                         repdem_demean   = repdem - mean(repdem))
>
> print(d[20:38,
+       c("muniID","muni_name","year","nat_rate","nat_rate_mean","nat_rate_demean","repdem","repdem_demean")
+       ],digits=2)
   muniID      muni_name year nat_rate nat_rate_mean nat_rate_demean repdem repdem_demean
20      2 Affoltern A.A. 1991     0.22           3.6           -3.38      0         -0.21
21      2 Affoltern A.A. 1992     0.95           3.6           -2.65      0         -0.21
22      2 Affoltern A.A. 1993     1.05           3.6           -2.55      0         -0.21
23      2 Affoltern A.A. 1994     0.83           3.6           -2.76      0         -0.21
24      2 Affoltern A.A. 1995     2.00           3.6           -1.59      0         -0.21
25      2 Affoltern A.A. 1996     1.78           3.6           -1.82      0         -0.21
26      2 Affoltern A.A. 1997     1.86           3.6           -1.73      0         -0.21
27      2 Affoltern A.A. 1998     2.05           3.6           -1.54      0         -0.21
28      2 Affoltern A.A. 1999     2.40           3.6           -1.19      0         -0.21
29      2 Affoltern A.A. 2000     2.20           3.6           -1.40      0         -0.21
30      2 Affoltern A.A. 2001     3.21           3.6           -0.39      0         -0.21
31      2 Affoltern A.A. 2002     4.64           3.6            1.04      0         -0.21
32      2 Affoltern A.A. 2003     4.84           3.6            1.25      0         -0.21
33      2 Affoltern A.A. 2004     5.62           3.6            2.03      0         -0.21
34      2 Affoltern A.A. 2005     4.39           3.6            0.79      0         -0.21
35      2 Affoltern A.A. 2006     8.12           3.6            4.52      1          0.79
36      2 Affoltern A.A. 2007     7.07           3.6            3.47      1          0.79
37      2 Affoltern A.A. 2008     8.98           3.6            5.38      1          0.79
38      2 Affoltern A.A. 2009     6.12           3.6            2.52      1          0.79
```

# Fixed Effects Regression with Demeaned Data

```
> summary(lm(nat_rate_demean~repdem_demean,data=d))

Call:
lm(formula = nat_rate_demean ~ repdem_demean, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4712 -2.0883 -0.5978  1.0841 21.3076

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.266e-16  5.279e-02    0.00        1
repdem_demean 3.023e+00  1.293e-01   23.39   <2e-16 ***
```

# Fixed Effects Regression with Canned Routine

```
> library(plm)
> library(lmtest)
> d <- plm.data(d, indexes = c("muniID", "year"))
> mod_fe <- plm(nat_rate~repdem,data=d,model="within")
> coeftest(mod_fe,
vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

       Estimate Std. Error t value  Pr(>|t|)
repdem  3.02280    0.18525  16.318 < 2.2e-16 ***
---
```

# Fixed Effects Regression with Dummies

```
> mod_du <- plm(nat_rate~repdem+as.factor(muniID),data=d,model="pooling")
> coeftest(mod_du, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                        Estimate  Std. Error     t value  Pr(>|t|)
(Intercept)           1.5922e+00  4.0068e-02  3.9737e+01 < 2.2e-16 ***
repdem                3.0228e+00  1.9032e-01  1.5883e+01 < 2.2e-16 ***
as.factor(muniID)2    1.3674e+00  1.4249e-08  9.5960e+07 < 2.2e-16 ***
as.factor(muniID)3    1.2923e+00  1.4283e-08  9.0472e+07 < 2.2e-16 ***
as.factor(muniID)9    1.2847e+00  1.3404e-08  9.5837e+07 < 2.2e-16 ***
as.factor(muniID)10   1.2718e+00  1.4182e-08  8.9675e+07 < 2.2e-16 ***
as.factor(muniID)13   3.2655e-01  1.2597e-08  2.5922e+07 < 2.2e-16 ***
as.factor(muniID)25   5.6413e-02  3.0051e-02  1.8772e+00 0.0605523 .
as.factor(muniID)26   3.1257e+00  1.0017e-02  3.1204e+02 < 2.2e-16 ***
as.factor(muniID)29   3.1797e+00  3.0051e-02  1.0581e+02 < 2.2e-16 ***
as.factor(muniID)33   3.2293e+00          NA          NA        NA
as.factor(muniID)34   1.7467e+00  3.0051e-02  5.8123e+01 < 2.2e-16 ***
.
```

# Applying Fixed Effects

- We can use fixed effects for other data structures to restrict comparisons to within unit variation
  - Matched pairs
    - Twin fixed effects to control for unobserved effects of family background
  - Cluster fixed effects in hierarchical data
    - School fixed effects to control for unobserved effects of school

# Problems that (even) fixed effects do not solve

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Where $y_{it}$ is murder rate and $x_{it}$ is police spending per capita
- What happens when we regress $y$ on $x$ and city fixed effects?
  - $\hat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on $c_i$ holds
    - $E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, c_i] = 0, \ t = 1, 2, ..., T$
    - implies $\varepsilon_{it}$ uncorrelated with past, current, and future regressors
- Most common violations:
  1. Time-varying omitted variables
     - economic boom leads to more police spending and less murders
     - can include time-varying controls, but avoid post-treatment bias
  2. Simultaneity
     - if city adjusts police based on past murder rate, then spending$_{t+1}$ is correlated with $\varepsilon_t$ (since higher $\varepsilon_t$ leads to higher murder rate at $t$)
     - strictly exogenous $x$ cannot react to what happens to $y$ in the past or the future!
- Fixed effects do not obviate need for good research design!

# Random Effects

- Reconsider our unobserved effects model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Cannot use the fixed effects regression to estimate effects of time-constant regressors in $\mathbf{x}_{it}$ (e.g. soil quality, farm location, etc.)
    - Since fixed effect estimator allows $c_i$ to be correlated with $\mathbf{x}_{it}$, we cannot distinguish the effects of time-invariant regressors from the time-invariant unobserved effect $c_i$
    - If a regressor does not change much in time, the standard errors of the coefficients in the fixed effects regression will be large (because there is little variation in the demeaned regressor $\ddot{\mathbf{x}}_{it} \equiv \mathbf{x}_{it} - \bar{\mathbf{x}}_i i$)
  - Need orthogonality assumption: $Cov[\mathbf{x}_{it}, c_i] = 0, \quad t = 1, ..., T$
    - Strong assumption: Unobserved effects $c_i$ are uncorrelated with each explanatory variable in $\mathbf{x}_{it}$ in each time period.
    - For example, if we include soil quality in $\mathbf{x}_{it}$, we have to assume it is uncorrelated with all other time-invariant inputs.

# Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

1. $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0, \ t = 1, 2, ..., T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|\mathbf{x}_i] = E[c_i] = 0$: unobserved effects $c_i$ are uncorrelated with regressors
3. $rank\, E[\mathbf{X}_i'\Omega\mathbf{X}_i] = K$: no collinearity among regressors
   - $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that $\Omega$ takes a special form:
   - $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{x}_i, c_i] = \sigma_\varepsilon^2\boldsymbol{I}_T$: idiosyncratic errors are homoscedastic for all $t$ and serially uncorrelated
   - $E[c_i^2|\mathbf{x}_i] = \sigma_c^2$: unobserved effect $c_i$ is homoscedastic

# Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

1. $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0, \ t = 1, 2, ..., T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|\mathbf{x}_i] = E[c_i] = 0$: unobserved effects $c_i$ are uncorrelated with regressors
3. $rank\, E[\mathbf{X}_i'\Omega\mathbf{X}_i] = K$: no collinearity among regressors
   - $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that $\Omega$ takes a special form:
   - $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{x}_i, c_i] = \sigma_\varepsilon^2\boldsymbol{I}_T$: idiosyncratic errors are homoscedastic for all $t$ and serially uncorrelated
   - $E[c_i^2|\mathbf{x}_i] = \sigma_c^2$: unobserved effect $c_i$ is homoscedastic

Assumption 4 implies $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i'|\mathbf{x}_i] = \begin{pmatrix} \sigma_c^2 + \sigma_\varepsilon^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_\varepsilon^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & & \sigma_c^2 + \sigma_\varepsilon^2 \end{pmatrix}_{T\times T}$

# Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

1. $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0, \ t = 1, 2, ..., T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|\mathbf{x}_i] = E[c_i] = 0$: unobserved effects $c_i$ are uncorrelated with regressors
3. $rank\, E[\mathbf{X}_i'\Omega\mathbf{X}_i] = K$: no collinearity among regressors
   - $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that $\Omega$ takes a special form:
   - $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{x}_i, c_i] = \sigma_\varepsilon^2\boldsymbol{I}_T$: idiosyncratic errors are homoscedastic for all $t$ and serially uncorrelated
   - $E[c_i^2|\mathbf{x}_i] = \sigma_c^2$: unobserved effect $c_i$ is homoscedastic

- Given assumptions 1-3, pooled OLS is consistent, since composite error $v_{it}$ is uncorrelated with $\mathbf{x}_{it}$ for all $t$
- However, pooled OLS ignores the serial correlation in $v_{it}$

# Random Effects Assumptions

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

1. $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0, \ t = 1, 2, ..., T$: explanatory variables are strictly exogenous conditional on the unobserved effect
2. $E[c_i|\mathbf{x}_i] = E[c_i] = 0$: unobserved effects $c_i$ are uncorrelated with regressors
3. $rank\, E[\mathbf{X}_i'\Omega\mathbf{X}_i] = K$: no collinearity among regressors
   - $\Omega = E[\boldsymbol{v}_i\boldsymbol{v}_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
4. We typically also assume that $\Omega$ takes a special form:
   - $E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{x}_i, c_i] = \sigma_\varepsilon^2\boldsymbol{I}_T$: idiosyncratic errors are homoscedastic for all $t$ and serially uncorrelated
   - $E[c_i^2|\mathbf{x}_i] = \sigma_c^2$: unobserved effect $c_i$ is homoscedastic

- Random effects estimator $\hat{\boldsymbol{\beta}}_{RE}$ exploits this serial correlation in a generalized least squares (GLS) framework
  - $\hat{\boldsymbol{\beta}}_{RE}$ is consistent under assumptions 1-3: $\underset{N\to\infty}{\text{plim}}\ \hat{\boldsymbol{\beta}}_{RE,N} = \boldsymbol{\beta}$
  - $\hat{\boldsymbol{\beta}}_{RE}$ is asymptotically efficient given assumption 4 (in the class of estimators consistent under $E[\mathbf{v}_i|\mathbf{x}_i] = \mathbf{0}$)

# Random Effects Estimator

- Consider the transformation parameter:

$$\lambda = 1 - \left( \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_c^2} \right)^{1/2} \quad \text{with } 0 \leq \lambda \leq 1$$

  - $\sigma_\varepsilon^2 = Var[\varepsilon_{it}]$: variance of idiosyncratic error
  - $\sigma_c^2 = Var[c_i]$: variance of unobserved effect

- $\hat{\beta}_{RE}$ is equivalent to pooled OLS on quasi-demeaned data:

$$\begin{aligned} y_{it} - \lambda \bar{y}_i &= (\mathbf{x}_{it} - \lambda \bar{x}_i)\boldsymbol{\beta} + (v_{it} - \lambda \bar{v}_i), \quad \forall\ i, t \\ \tilde{y}_{it} &= \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \tilde{v}_{it} \end{aligned}$$

  - As $\lambda \to 1$, $\hat{\beta}_{RE} \to \hat{\beta}_{FE}$
  - As $\lambda \to 0$, $\hat{\beta}_{RE} \to \hat{\beta}_{Pooled\ OLS}$
    - $\lambda \to 1$ as $T \to \infty$ or if variance of $c_i$ is large relative to variance of $\varepsilon_{it}$

- $\lambda$ can be estimated from data $\widehat{\lambda} = 1 - (\widehat{\sigma}_\varepsilon^2/(\widehat{\sigma}_\varepsilon^2 + T\widehat{\sigma}_c^2))^{1/2}$
- Usually wise to cluster the standard errors since assumption 4 is strong

# Random Effects Regression

```
> mod_re <- plm(nat_rate~repdem,data=d,model="random")
> coeftest(mod_re, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 2.120793   0.097108  21.840 < 2.2e-16 ***
repdem      2.859397   0.189008  15.129 < 2.2e-16 ***
---
```

# Summary: Fixed Effects, Random Effects, Pooled OLS

- Main assumptions:
  1. Regressors are strictly exogenous conditional on the time-invariant unobserved effects
  2. Regressors are uncorrelated with the time-invariant unobserved effects

- Results:
  - Fixed effects estimator is consistent given assumption 1, but rules out time-invariant regressors
  - Random effects estimator and pooled OLS are consistent under assumptions 1-2, and allow for time-invariant regressors
  - Given homoscedasticity assumptions (random effects assumption 4), the random effects estimator is asymptotically efficient

- Assumption 2 is strong so fixed effects are typically more credible
  - Often the main reason for using panel data is to rule out all time-invariant unobserved confounders!

# Hausman Test

Given the homoskedastic model (RE assumptions 1-4):

| | $\widehat{\beta}_{RE}$ | $\widehat{\beta}_{FE}$ |
|---|---|---|
| $H_0 : \text{Cov}[\mathbf{x}_{it}, c_i] = 0$ | consistent and efficient | consistent |
| $H_1 : \text{Cov}[\mathbf{x}_{it}, c_i] \neq 0$ | inconsistent | consistent |

Then,

- Under $H_0$, $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be close to zero.
- Under $H_1$, $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be different from zero.
- It can be shown that in large samples, under $H_0$, the test statistic

$$(\widehat{\beta}_{FE} - \widehat{\beta}_{RE})'(\widehat{\text{Var}}[\widehat{\beta}_{FE}] - \widehat{\text{Var}}[\widehat{\beta}_{RE}])^{-1}(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}) \xrightarrow{d} \chi^2_k$$

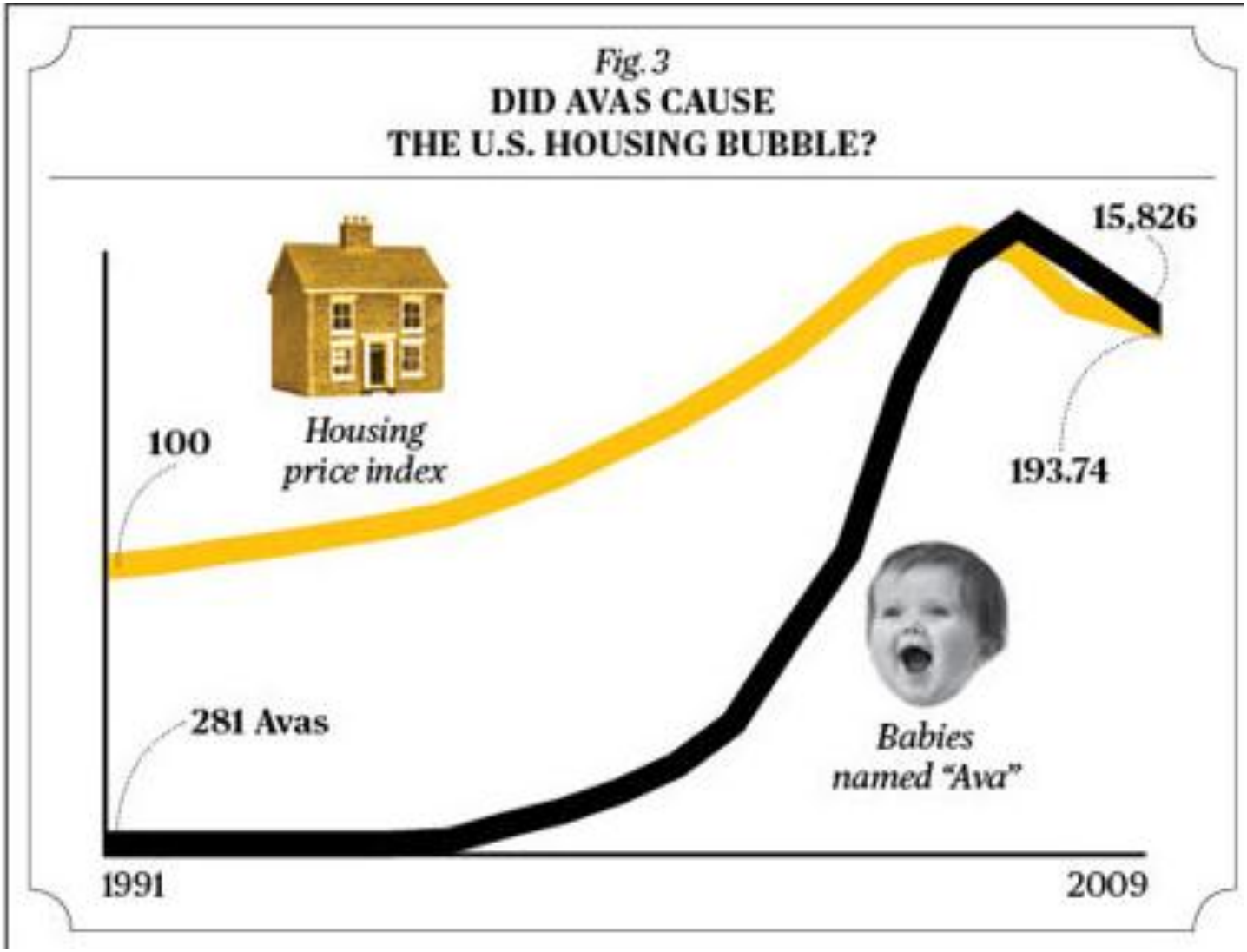
  where $k$ is number of time-varying regressors.
- We may reject the null hypothesis of "random effects" and stick with the less efficient, but consistent fixed effect specification.

# Hausman Test

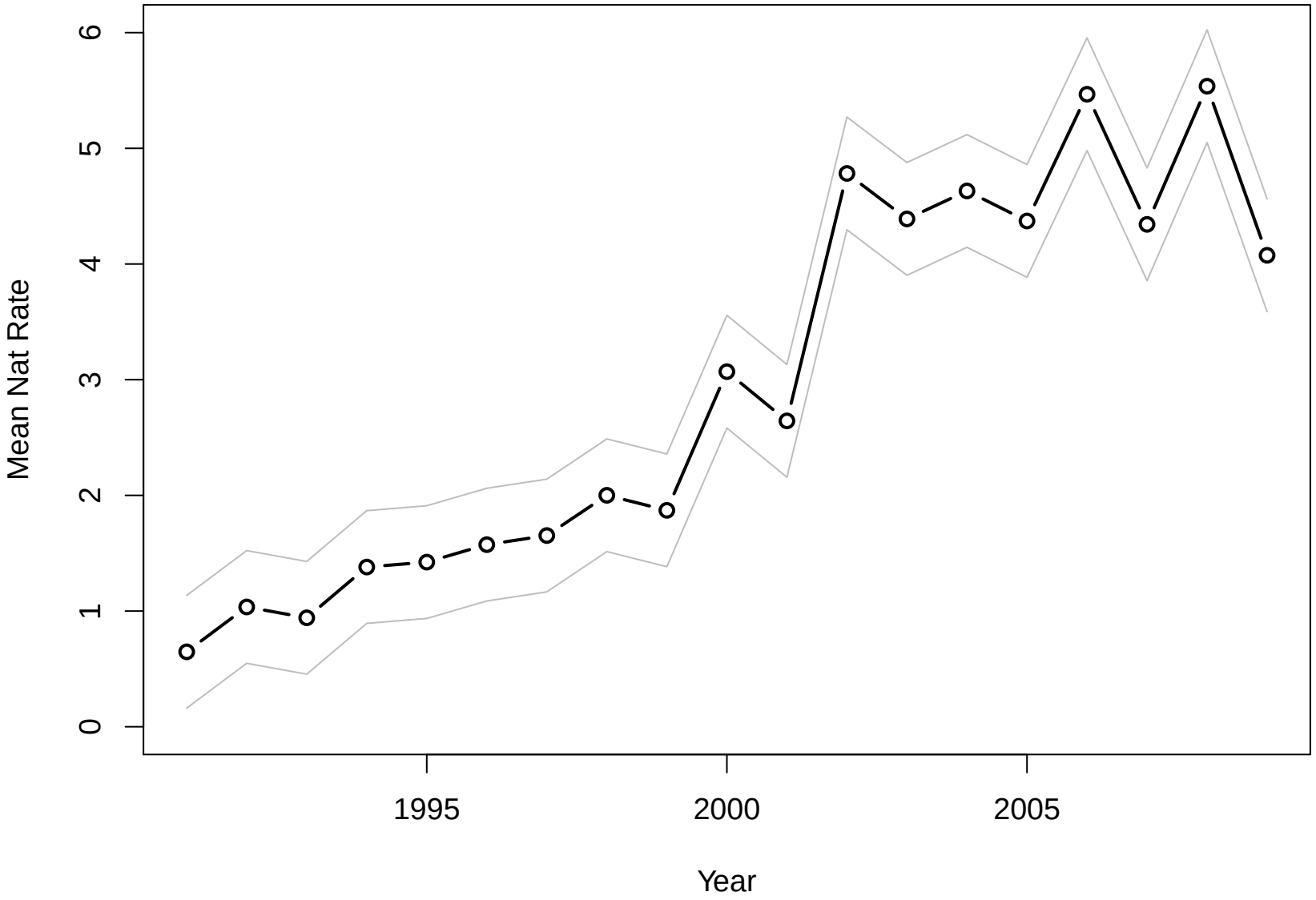
```
> phtest(mod_fe, mod_re)

Hausman Test

data:  nat_rate ~ repdem
chisq = 28.7935, df = 1, p-value = 8.052e-08
alternative hypothesis: one model is inconsistent
```

- Hausman test does not test if the fixed effects model is correct, the test assumes that the fixed effects estimator is consistent!
- Conventional Hausman test assumes homoscedastic model and does not allow for clustering

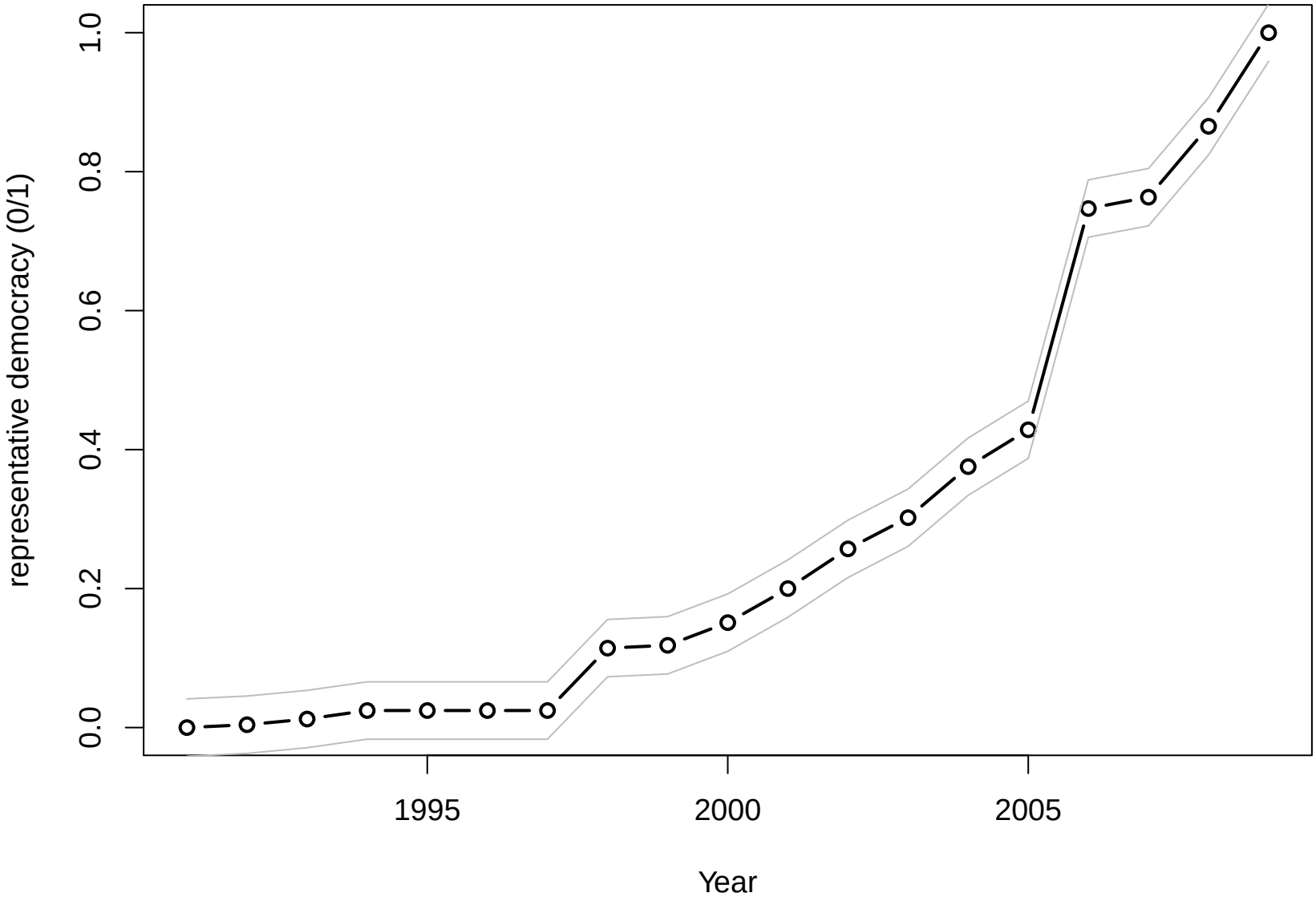
# Common Shocks or Causation?



# Naturalization Rate Over Time



# Representative Democracy Over Time



# Adding Time Effects

- Reconsider our unobserved effects model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Fixed effects assumption: $E[\varepsilon_{it}|\mathbf{x}_i, c_i] = 0, \ t = 1, 2, ..., T$: regressors are strictly exogenous conditional on the unobserved effect
  - Typical violation: Common shocks that affect all units in the same way and are correlated with $\mathbf{x}_{it}$.
    - Trends in farming technology or climate affect productivity
    - Trends in immigration inflows affect naturalization rates
  - We can allow for such common shocks by including time effects into the model

# Fixed Effects Regression: Adding Time Effects

- Linear time trend:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Linear time trend common to all units

- Time fixed effects:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + t_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

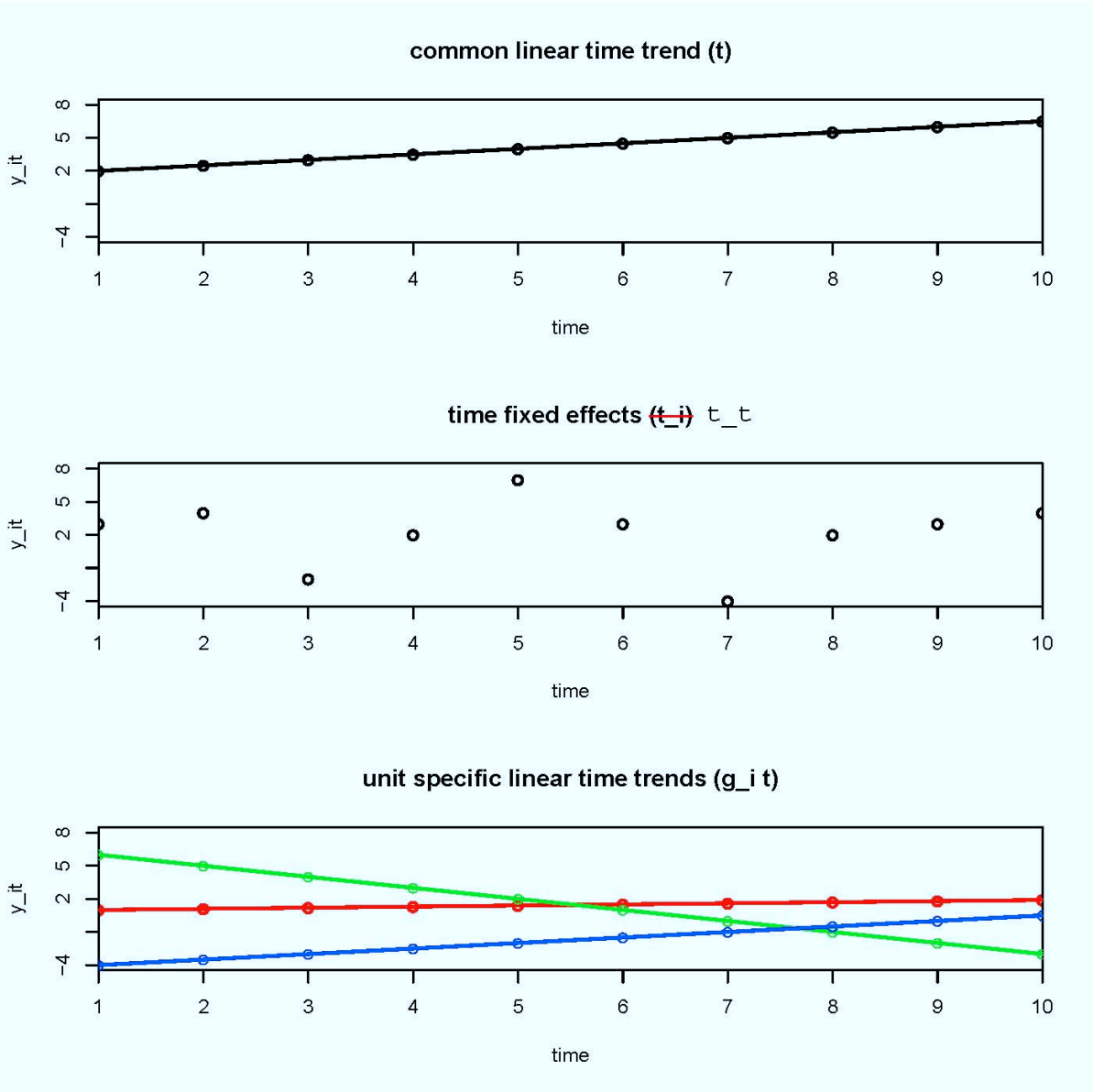
  - Common shock in each time period
  - Generalized difference-in-differences model

- Unit specific linear time trends:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + g_i \cdot t + t_t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Linear time trends that vary by unit

# Modeling Time Effects



# Modeling Time Effects

```
> d$time <- as.numeric(d$year)
> d[1:21,c("muniID","muni_name","year","time")]
   muniID      muni_name year time
1       1    Aeugst A.A. 1991    1
2       1    Aeugst A.A. 1992    2
3       1    Aeugst A.A. 1993    3
4       1    Aeugst A.A. 1994    4
5       1    Aeugst A.A. 1995    5
6       1    Aeugst A.A. 1996    6
7       1    Aeugst A.A. 1997    7
8       1    Aeugst A.A. 1998    8
9       1    Aeugst A.A. 1999    9
10      1    Aeugst A.A. 2000   10
11      1    Aeugst A.A. 2001   11
12      1    Aeugst A.A. 2002   12
13      1    Aeugst A.A. 2003   13
14      1    Aeugst A.A. 2004   14
15      1    Aeugst A.A. 2005   15
16      1    Aeugst A.A. 2006   16
17      1    Aeugst A.A. 2007   17
18      1    Aeugst A.A. 2008   18
19      1    Aeugst A.A. 2009   19
20      2 Affoltern A.A. 1991    1
21      2 Affoltern A.A. 1992    2
```

# Fixed Effects Regression: Linear Time Trend

```
> mod_fe <- plm(nat_rate~repdem+time,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

       Estimate Std. Error t value  Pr(>|t|)
repdem  0.82479    0.25853  3.1903  0.001431 **
time    0.23137    0.01714 13.4987 < 2.2e-16 ***
```

# Fixed Effects Regression: Year Fixed Effects

```
> mod_fe <- plm(nat_rate~repdem+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x) vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

         Estimate Std. Error t value  Pr(>|t|)
repdem    1.20366    0.30253  3.9786 7.043e-05 ***
year1992  0.38292    0.17197  2.2266   0.02602 *
year1993  0.27898    0.15110  1.8463   0.06492 .
year1994  0.70341    0.16712  4.2089 2.617e-05 ***
year1995  0.74591    0.17827  4.1841 2.919e-05 ***
year1996  0.89693    0.18345  4.8892 1.049e-06 ***
year1997  0.97570    0.18661  5.2285 1.788e-07 ***
year1998  1.21550    0.22506  5.4007 6.988e-08 ***
year1999  1.08051    0.21430  5.0419 4.794e-07 ***
year2000  2.23993    0.23968  9.3457 < 2.2e-16 ***
year2001  1.75531    0.24790  7.0807 1.662e-12 ***
year2002  3.82573    0.32672 11.7096 < 2.2e-16 ***
year2003  3.37837    0.32664 10.3428 < 2.2e-16 ***
year2004  3.53176    0.34285 10.3012 < 2.2e-16 ***
year2005  3.20837    0.31097 10.3171 < 2.2e-16 ***
year2006  3.92057    0.39023 10.0468 < 2.2e-16 ***
year2007  2.77646    0.36884  7.5276 6.237e-14 ***
year2008  3.84780    0.40135  9.5872 < 2.2e-16 ***
year2009  2.22388    0.41997  5.2953 1.246e-07 ***
---
```

# Unit Specific Time Trends Often Eliminate "Results"

TABLE 5.2.3
Estimated effects of labor regulation on the performance of firms in Indian states

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Labor regulation (lagged) | −.186 | −.185 | −.104 | .0002 |
| | (.064) | (.051) | (.039) | (.020) |
| Log development expenditure per capita | | .240 | .184 | .241 |
| | | (.128) | (.119) | (.106) |
| Log installed electricity capacity per capita | | .089 | .082 | .023 |
| | | (.061) | (.054) | (.033) |
| Log state population | | .720 | 0.310 | −1.419 |
| | | (.96) | (1.192) | (2.326) |
| Congress majority | | | −.0009 | .020 |
| | | | (.01) | (.010) |
| Hard left majority | | | −.050 | −.007 |
| | | | (.017) | (.009) |
| Janata majority | | | .008 | −.020 |
| | | | (.026) | (.033) |
| Regional majority | | | .006 | .026 |
| | | | (.009) | (.023) |
| State-specific trends | No | No | No | Yes |
| Adjusted $R^2$ | .93 | .93 | .94 | .95 |

*Notes*: Adapted from Besley and Burgess (2004), table IV. The table reports regression DD estimates of the effects of labor regulation on productivity. The

"labor regulation increased in states where output was declining anyway"

# Fixed Effects Regression: Unit Specific Time Trends

```
> mod_fe <- plm(nat_rate~
repdem+muniID*time+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                  Estimate  Std. Error     t value  Pr(>|t|)
repdem           0.9865241    .322868    3.0634545  2.043e-05 ***
muniID2:time    -4.1916e-02  2.0515e-09 -2.0432e+07 < 2.2e-16 ***
muniID3:time    -8.4358e-02  2.1145e-09 -3.9896e+07 < 2.2e-16 ***
.
```

# Unit Specific Quadratic Time Trends

```
> d$time2 <- d$time^2
> mod_fe <- plm(nat_rate~repdem+
muniID*time+muniID*time^2+year,data=d,model="within")
> coeftest(mod_fe, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                 Estimate  Std. Error     t value  Pr(>|t|)
repdem          1.22272779   .3804359   3.212323  1.023e-05 ***
muniID2:time     1.37177084  1.034e-09    .0344+07 < 3.4e-16 ***
muniID2:time2   -0.07068432  2.2034e-09 -1.234e+07 < 5.6e-16 ***
.
```

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in $x$ may occur with a lag
  - effect of new tax credit for children on fertility rate

- Interpretation of coefficients:
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$, which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in $x$ may occur with a lag
  - effect of new tax credit for children on fertility rate

- Interpretation of coefficients:
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$, which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$

- $\beta_0 = y_t - y_{t-1}$ immediate change in $y$ due to temporary one-unit increase in $x$ (impact propensity)

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in $x$ may occur with a lag
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$, which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $\beta_1 = y_{t+1} - y_{t-1}$ change in $y$ one period after temporary one-unit increase in $x$

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in $x$ may occur with a lag
  - effect of new tax credit for children on fertility rate

- Interpretation of coefficients:
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$, which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$

- $\beta_2 = y_{t+2} - y_{t-1}$ change in $y$ two periods after temporary one-unit increase in $x$

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Model recognizes that effect of change in $x$ may occur with a lag
  - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$, which lasts only one period
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- $y_{t+3} = y_{t-1}$ change in $y$ is zero three periods after temporary one-unit increase in $x$

# Distributed Lag Model

$$y_{it} = x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Interpretation of coefficients:
  - Consider permanent increase in $x_{it}$ from level $m$ to $m+1$ at $t$, i.e. ($x_s = m, s < t$ and $x_s = m+1, s \geq t$)
    - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_t \quad = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t+1} = (m+1)\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
    - $y_{t+2} = (m+1)\beta_0 + (m+1)\beta_1 + (m+1)\beta_2 + c_i$
    - $y_{t+3} = (m+1)\beta_0 + (m+1)\beta_1 + (m+1)\beta_2 + c_i$
- After one period $y$ has increased by $\beta_0 + \beta_1$, after two periods, $y$ has increased by $\beta_0 + \beta_1 + \beta_2$, and there are no further increases after two periods
- Long-run increase in $y$: $\beta_0 + \beta_1 + \beta_2$ (long-run propensity)

# Lagged Effects of Direct Democracy

```
> mod_lag    <- plm(nat_rate~repdem+
  lag(repdem,1)+lag(repdem,2)+lag(repdem,3)+
+                      year,data=d,model="within")
> coeftest(mod_lag, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))
t test of coefficients:
                Estimate Std. Error t value  Pr(>|t|)
repdem          0.636480   0.358658  1.7746  0.076044 .
lag(repdem, 1)  1.201065   0.422508  2.8427  0.004498 **
lag(repdem, 2) -0.164869   0.468783 -0.3517  0.725087
lag(repdem, 3) -0.524521   0.410152 -1.2788  0.201033
year1995        0.031281   0.204172  0.1532  0.878244
year1996        0.188603   0.210878  0.8944  0.371183

# long run effect
> sum(coef(mod_lag)[1:4])
[1] 1.148156
```

# Lags and Leads Model

$$y_{it} = x_{it+1}\beta_{-1} + x_{it}\beta_0 + x_{it-1}\beta_1 + x_{it-2}\beta_2 + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Can use estimate of $\beta_{-1}$ to test for anticipation effects
  - Consider temporary increase in $x_{it}$ from level $m$ to $m+1$ at $t$
    - $y_{t-2} = \beta_{-1}m + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
    - $y_{t-1} = \beta_{-1}(m+1) + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- Anticipation effect: $\beta_{-1} = y_{t-1} - y_{t-2}$ change in $y$ in period $t-1$, the period before the temporary one-unit increase in $x$
- Placebo test: if $x$ causes $y$, but $y$ does not cause $x$, then $\beta_{-1}$ should be close to zero

# Leads and Lags

```
> d <- ddply(
+   d, .(muniID), transform,
+   lead_repdem = c( repdem[-1],NA )
+ )
> d <- plm.data(d, indexes = c("muniID", "year"))
> mod_lagleads    <- plm(nat_rate~lead_repdem+
repdem+lag(repdem,1)+lag(repdem,2)+lag(repdem,3)+
+                      year,data=d,model="within")
> coeftest(mod_lagleads, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

                Estimate Std. Error t value  Pr(>|t|)
lead_repdem     0.170768   0.320634  0.5326 0.5943479
repdem          0.697573   0.438811  1.5897 0.1119975
lag(repdem, 1)  0.872396   0.460988  1.8924 0.0585159 .
lag(repdem, 2)  0.014941   0.457426  0.0327 0.9739451
lag(repdem, 3) -0.290425   0.409985 -0.7084 0.4787574
year1995        0.032882   0.204374  0.1609 0.8721896
```

# The Autor Test

- Let $D_{it}$ be a binary indicator coded 1 if unit $i$ switched from control to treatment between $t$ and $t-1$; 0 otherwise
  - Lags: $D_{it-1}$: unit switched between $t-1$ and $t-2$
  - Leads: $D_{it+1}$: unit switches between $t+1$ and $t$
- Include lags and leads into the fixed effects model:

$$y_{it} = D_{it+2}\beta_{-2} + D_{it+1}\beta_{-1} + D_{it}\beta_0 + D_{it-1}\beta_1 + D_{it-2}\beta_2 + c_i + \varepsilon_{it}$$

- Interpretation of coefficients:
  - Leads $\beta_{-1}$, $\beta_{-2}$, etc. test for anticipation effects
  - Switch $\beta_0$ tests for immediate effect
  - Lags $\beta_1$, $\beta_2$, etc. test for long-run effects
    - highest lag dummy can be coded 1 for all post-switch years

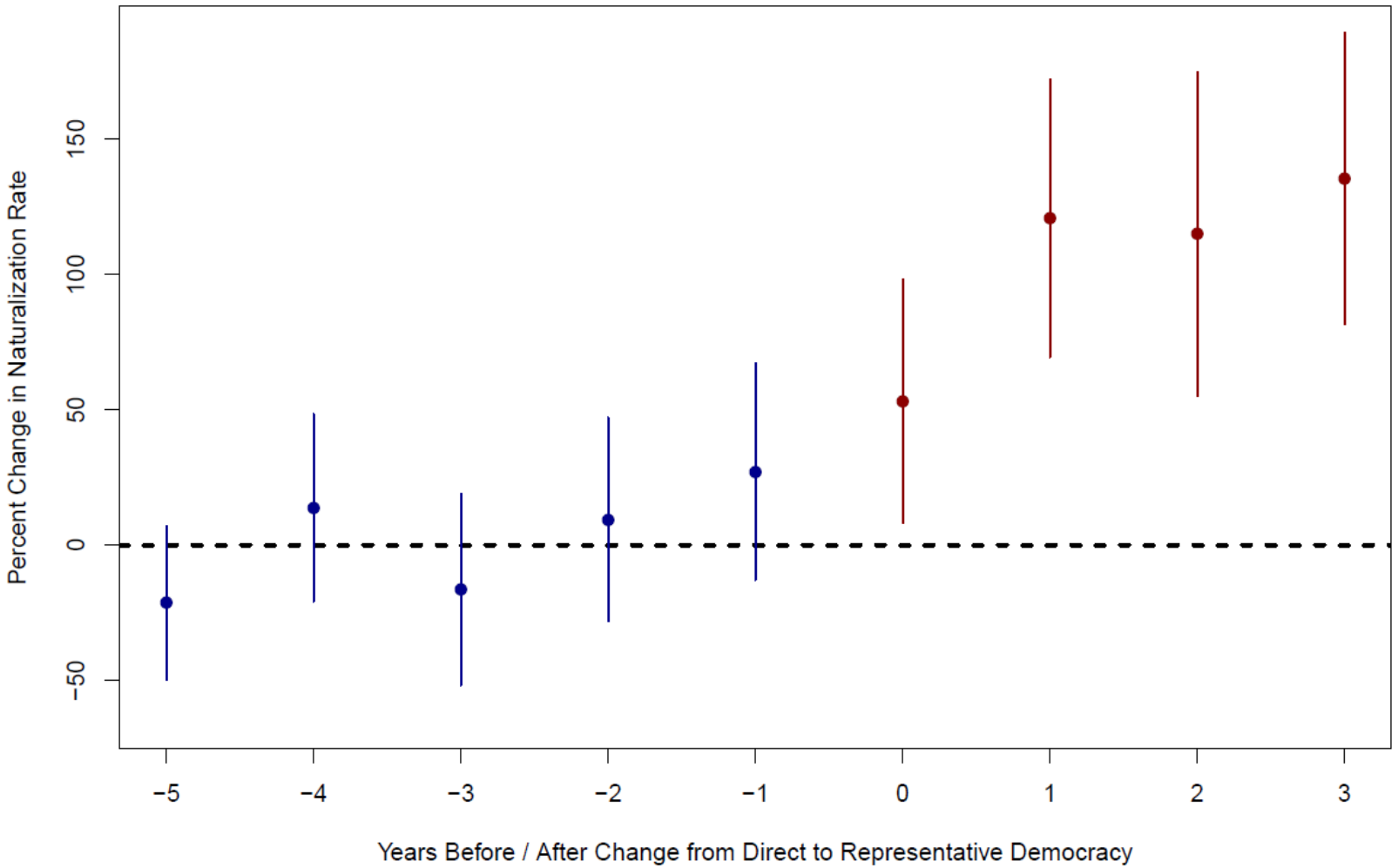
# Lags and Leads of Switch to Representative Democracy

```
> d[970:989,c(1:3,5,12:ncol(d))]
    muniID year muni_name repdem switcht lag1 lag2 lag3 lead1 lead2 lead3 lead4 lead5
970    220 1991 Hagenbuch      0       0    0    0    0     0     0     0     0     0
971    220 1992 Hagenbuch      0       0    0    0    0     0     0     0     0     0
972    220 1993 Hagenbuch      0       0    0    0    0     0     0     0     0     0
973    220 1994 Hagenbuch      0       0    0    0    0     0     0     0     0     0
974    220 1995 Hagenbuch      0       0    0    0    0     0     0     0     0     0
975    220 1996 Hagenbuch      0       0    0    0    0     0     0     0     0     0
976    220 1997 Hagenbuch      0       0    0    0    0     0     0     0     0     0
977    220 1998 Hagenbuch      0       0    0    0    0     0     0     0     0     1
978    220 1999 Hagenbuch      0       0    0    0    0     0     0     0     1     0
979    220 2000 Hagenbuch      0       0    0    0    0     0     0     1     0     0
980    220 2001 Hagenbuch      0       0    0    0    0     0     1     0     0     0
981    220 2002 Hagenbuch      0       0    0    0    0     1     0     0     0     0
982    220 2003 Hagenbuch      1       1    0    0    0     0     0     0     0     0
983    220 2004 Hagenbuch      1       0    1    0    0     0     0     0     0     0
984    220 2005 Hagenbuch      1       0    0    1    0     0     0     0     0     0
985    220 2006 Hagenbuch      1       0    0    0    1     0     0     0     0     0
986    220 2007 Hagenbuch      1       0    0    0    1     0     0     0     0     0
987    220 2008 Hagenbuch      1       0    0    0    1     0     0     0     0     0
988    220 2009 Hagenbuch      1       0    0    0    1     0     0     0     0     0
989    224 1991   Pfungen      0       0    0    0    0     0     0     0     0     0
```

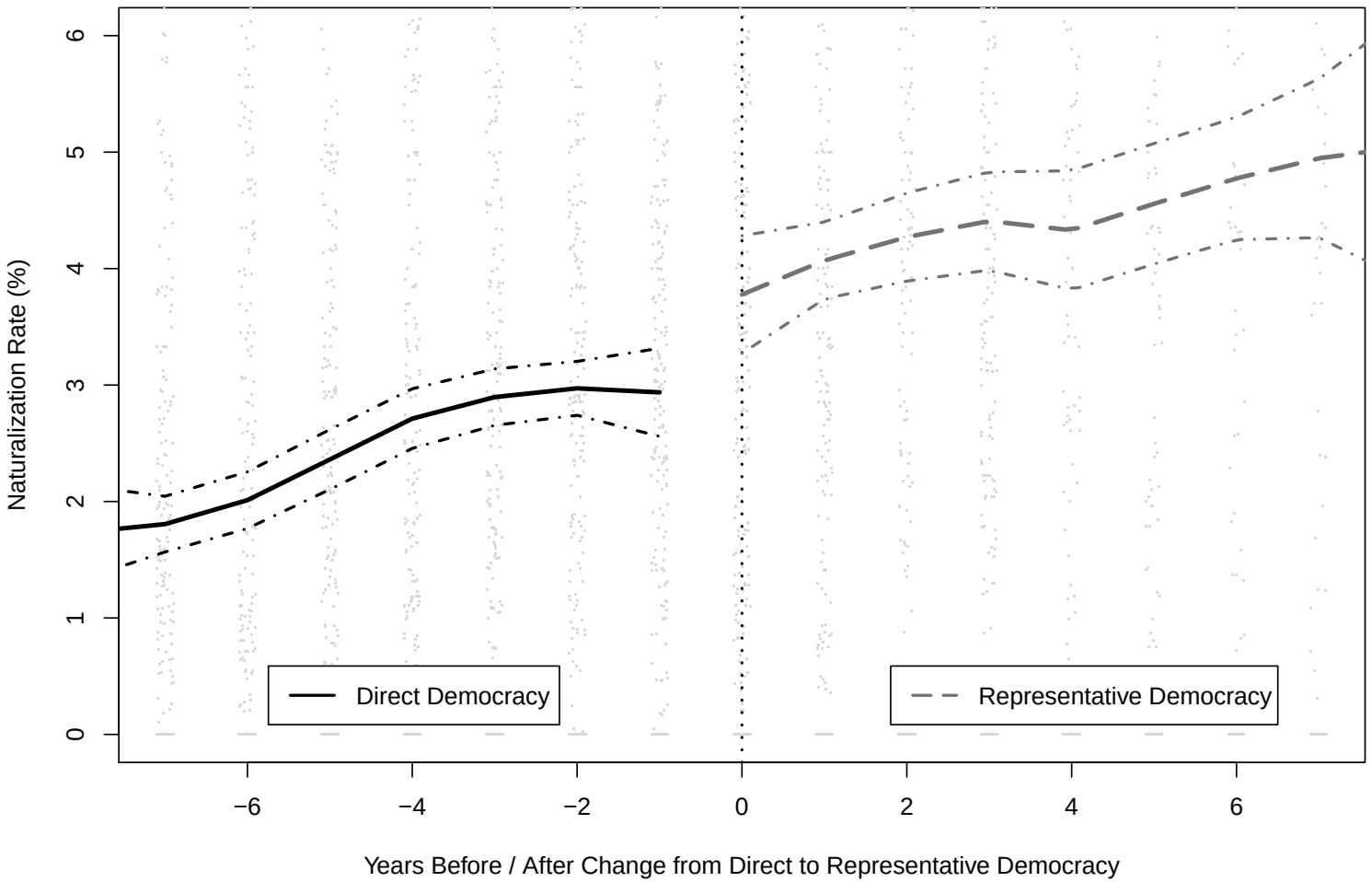
# Dynamic Effect of Switching to Representative Democracy

```
> mod_all    <- plm(nat_rate~lag3+lag2+lag1+switcht+
lead1+lead2+lead3+lead4+lead5+
+                          year,data=d,model="within")
> coeftest(mod_all, vcov=function(x)
vcovHC(x, cluster="group", type="HC1"))

t test of coefficients:

          Estimate Std. Error t value  Pr(>|t|)
lag3      1.160345   0.506989  2.2887 0.0221442 *
lag2      1.743682   0.538419  3.2385 0.0012105 **
lag1      1.881663   0.487013  3.8637 0.0001133 ***
switcht   0.756479   0.427751  1.7685 0.0770463 .
lead1     0.213876   0.389191  0.5495 0.5826635
lead2     0.084368   0.356799  0.2365 0.8130891
lead3     0.144045   0.318756  0.4519 0.6513661
lead4     0.075019   0.298425  0.2514 0.8015287
lead5    -0.094241   0.259448 -0.3632 0.7164439
year1992  0.385289   0.172172  2.2378 0.0252829 *
```

# Dynamic Effect of Switching to Representative Democracy



# Switching Plot



# Lagged Dependent Variable

$$y_{it} = \alpha\, y_{it-1} + c_i + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- $y_{it}$ could be capital stock of firm $i$ at time $t$, and $\alpha$ the capital depreciation rate
- Models with unit fixed effects and lagged $y$ do not produce consistent estimators!
  - after taking first differences to eliminate $c_i$, the differenced residual $\Delta\varepsilon_{it}$ is correlated with the lagged dependent variable $\Delta y_{it-1}$ by construction
- We might use past levels $y_{it-2}$ as an instrument for $\Delta y_{it-1}$, but this requires strong assumptions (e.g. no serial correlation in $\varepsilon_{it}$)

# Heterogeneous Treatment Effects

- So far we have assumed that the treatment effect is constant across units
- Can allow for heterogeneous treatment effects by including interaction of treatment with other regressors

$$y_{it} = treat_{it}\alpha_0 + (treat_{it} \cdot x_{it})\alpha_1 + x_{it}\beta + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

- Often the treatment is interacted with a time-invariant regressor:

$$y_{it} = treat_{it}\alpha_0 + (treat_{it} \cdot x_i)\alpha_1 + c_i + t + \varepsilon_{it}, \qquad t = 1, 2, ..., T$$

  - Note: The lower order term on the time-invariant $x_i$ is collinear with the fixed effects and drops out

# Heterogeneous Effect of Direct Democracy