Some Recent Updates on Causal Inference Methods

ZHANG CHUANCHUAN

ccz.zhang@gmail.com

DiD and Event Study

1. Basic Ideas and Model Setup

- Origin of the DID
- Modern applications
- Model setup and assumptions

2. Extensions

- G*T; Cohort DID; DDD; ES; SC&DID; Bartik IV

3. Heterogeneity

- Interpretation of the TWFE estimators
- Problems of the TWFE estimators with heterogeneous treatment effects
- Alternative estimators

1. Basic Ideas and Model Setup

- **1.1 Origin of the DID**
  - John Snow's Cholera Hypothesis
    - Cholera was transmitted by water, not air (Snow 1855)

- **1.1 Origin of the DID**
  - Snow collected data on household enrollment in water supply companies, then matched those data with the city's data on the cholera death rates at the household level

- **1.1 Origin of the DID**

  - In 1849, there were 135 cases of cholera per 10,000 households at Southwark and Vauxhall and 85 for Lambeth. But in 1854, there were 147 per 100,000 in Southwark and Vauxhall, whereas Lambeth's cholera cases per 10,000 households fell to 19.

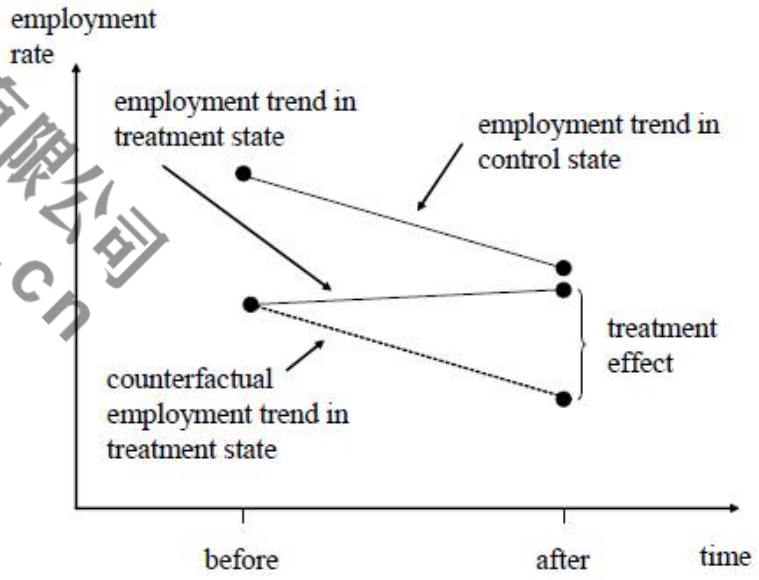
Company name	1849	1854
Southwark and Vauxhall	135	147
Lambeth	85	19

- **1.2 Modern applications**
  - Card and Krueger(1994)

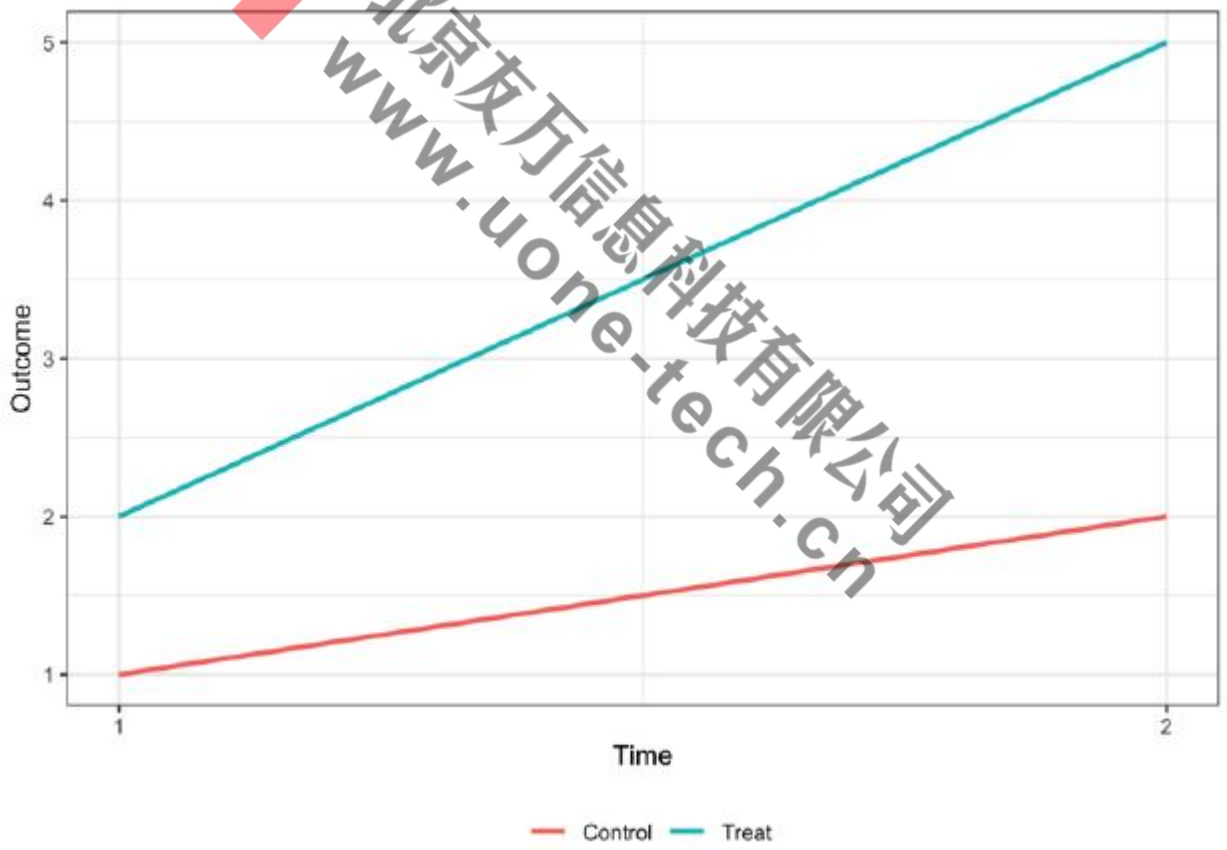
Average employment per store before and after the New Jersey minimum wage increase

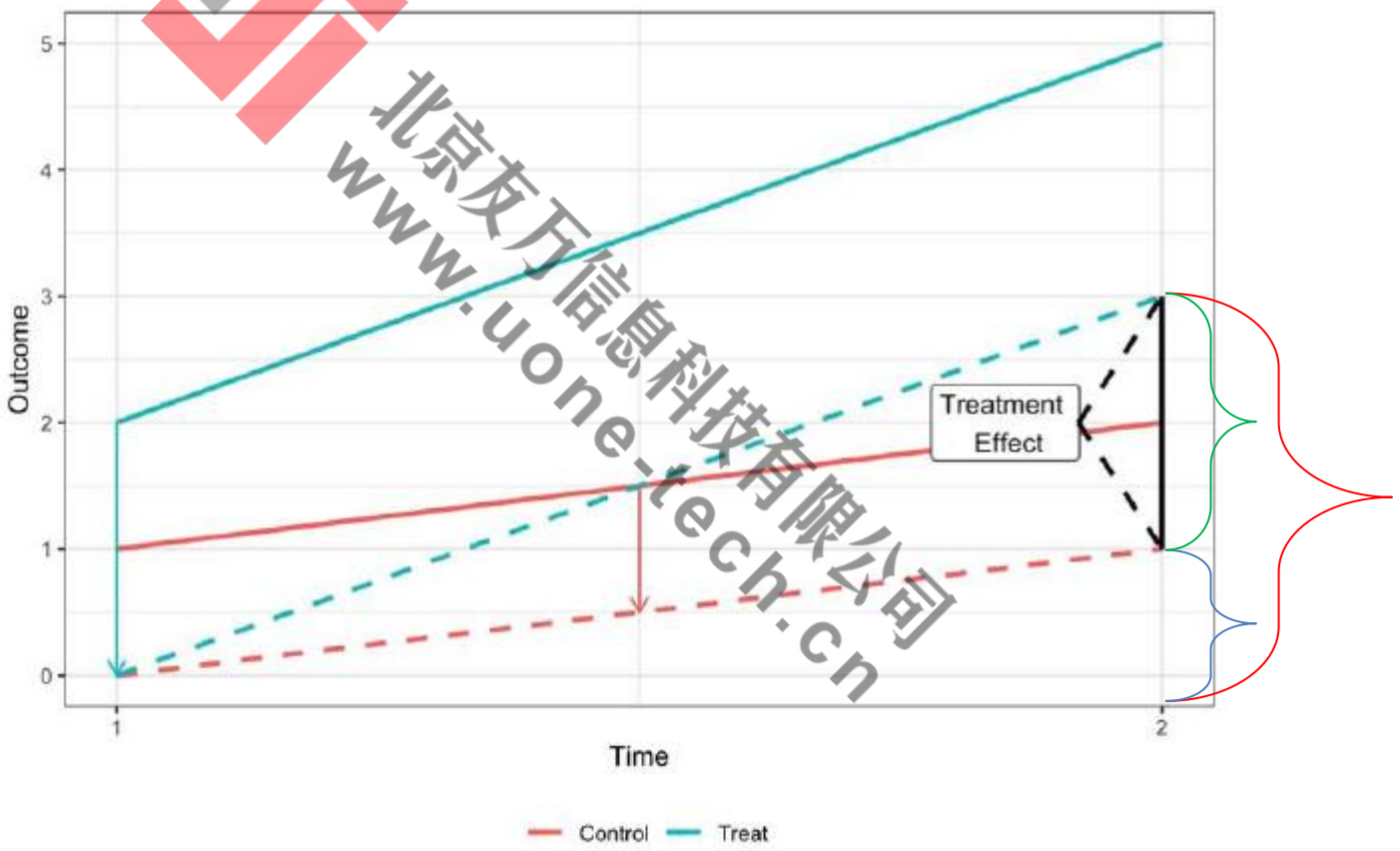
Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses



- **1.3 Model setup and assumptions**





- It can be shown that the treatment effect shown in the above plot is equivalent to the coefficient of the interaction between *Post* and *Treat* in the following regression:

$$Y_{it} = \gamma_t POST_t + \gamma_i TREAT_i + \delta POST_t \times TREAT_i + \epsilon_{it}$$

- **Assumptions**
  - Common trends: The **changes** in time trend is **independent of the treatment status** (or the differences by treatment status is independent of the time effect).
    - Two time periods, two groups

$$\Delta Y_i = \lambda + \tau Treated_i + \Delta\epsilon_i$$

- **Assumptions (cont.)**
  - No confounding policies
  - No spillover effects (implied by STUVA)
  - Functional form: An additive structure for potential outcome in the control group

$$E(Y_{0ist}|s,t) = \gamma_s + \lambda_t$$

- *Note: Comparing with the individual fixed effects estimation, the regressor of interest in DiD setup varies only at a more aggregate level such as region or cohort. This implies that DiD doesn't require individual-level panel data.*

- Regression DD

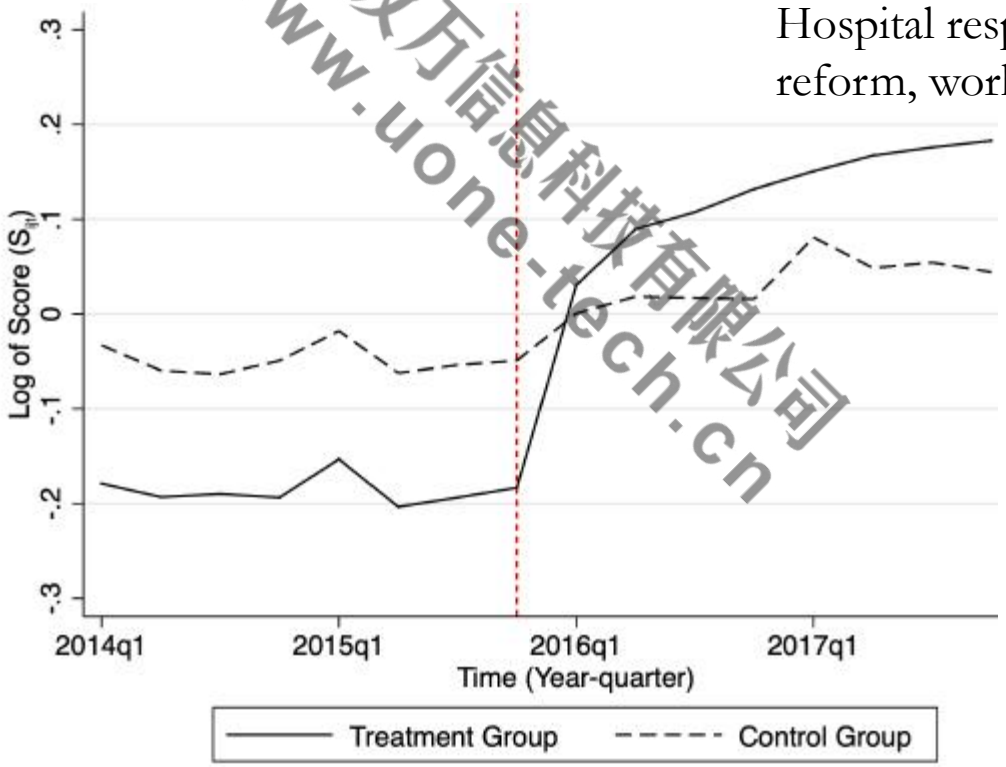
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta(NJ_s \cdot d_t) + \varepsilon_{ist}$$

- The link between the parameters in regression equation and those conditional means in the DD model illustrated by potential outcomes,

$$\begin{aligned}
\alpha &= E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb} \\
\gamma &= E(Y_{ist}|s = NJ, t = Feb) - E(Y_{ist}|s = PA, t = Feb) = \gamma_{NJ} - \gamma_{PA} \\
\lambda &= E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb} \\
\beta &= \{E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)\} \\
&\quad -\{E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)\}.
\end{aligned}$$

- (**Indirect/Supportive**) Test for assumption of common trends (fundamentally untestable)

Yan, Yi and Zhang, 2021, Hospital responses to DIP reform, working paper



2. Extensions

2.1 Multiple time periods or multiple groups

2.2 DiD using cross-sectional data

2.3 Triple differences, DDD

2.4 An event study approach

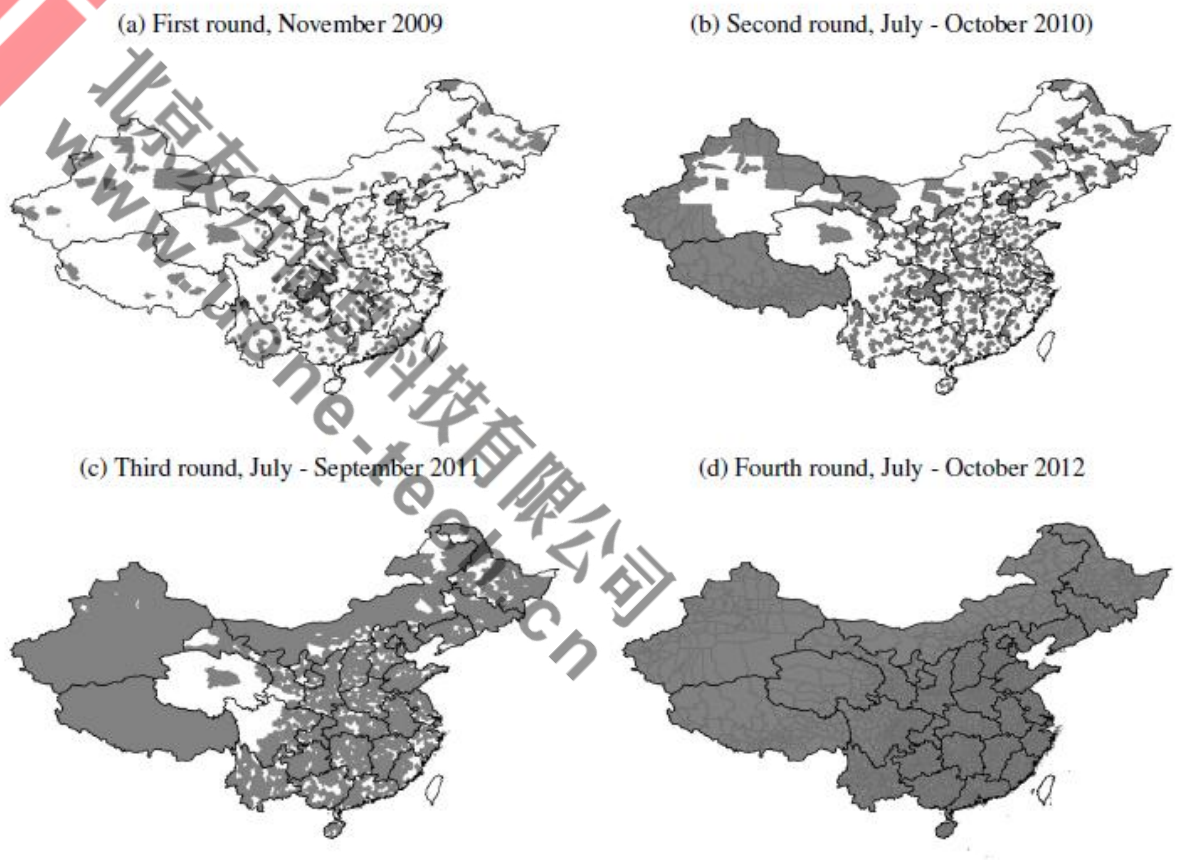
2.5 One related approach: Synthetic control

2.6 Another related approach: Bartik Instruments

- **2.1 Multiple time periods or multiple groups**
  - No variation in treatment across time/group
  - **Variation in treatment timing**: Staggered DiD
  - **Variation in treatment intensity** (multiple-values/continuous treatment)

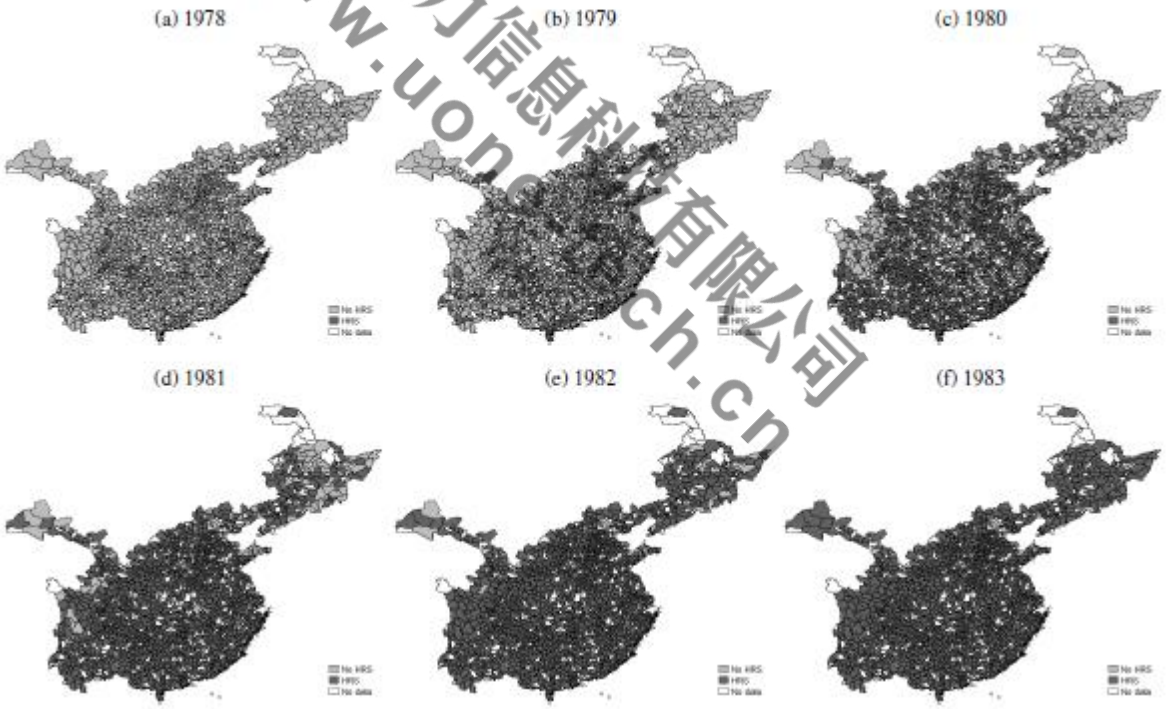
- Staggered DiD
  - Huang and Zhang (2021, AEJ)

Figure 1: County-by-County Roll-Out of the NRPS over Time



- Staggered DiD
  - Hu, Zhi-an et al., Short-term Gains, Long-term Loss: Unintended Effects of China's Land Reform on Education and Labor Market Outcomes (December 16, 2019). Available at SSRN: https://ssrn.com/abstract=3504733

Figure 2: Roll-out the HRS across Counties between 1978 and 1983



- DiD with continuous treatment
  - Nunn and Qian (2009, QJE)

$$y_{it} = \beta \ln Potato\ Area_i \cdot I_t^{Post} + \sum_{j=1100}^{1900} X'_i \Gamma_t^j \Phi_j + \sum_c \gamma_c I_i^c + \sum_{j=1100}^{1900} \rho_j I_t^j + \varepsilon_{it}$$

$$y_{it} = \sum_{j=1100}^{1900} \beta_j \ln Potato\ \ Area_i \cdot I_t^j + \sum_{j=1100}^{1900} X'_i \Gamma_t^j \Phi_j + \sum_c \gamma_c I_i^c + \sum_{j=1100}^{1900} \rho_j I_t^j + \varepsilon_{it}$$

- Conventional model specification under multiple time periods or multiple groups
  - Static specification: TWFE

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau^{\text{static}} D_{it} + \tilde{\varepsilon}_{it}$$

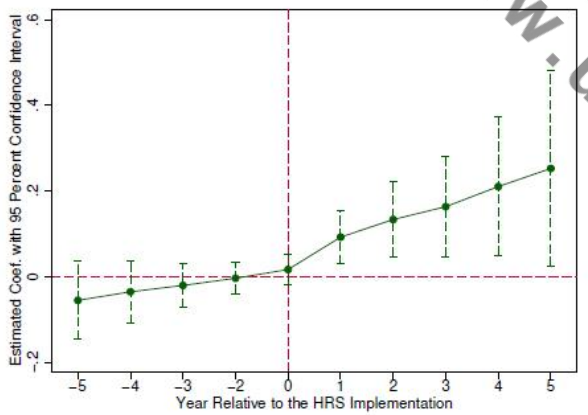
  - Dynamic specification: TWFE with lags and leads

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\substack{h=-a \\ h \neq -1}}^{b-1} \tau_h \mathbf{1}\left[K_{it} = h\right] + \tau_{b+} \mathbf{1}\left[K_{it} \geq b\right] + \tilde{\varepsilon}_{it}$$

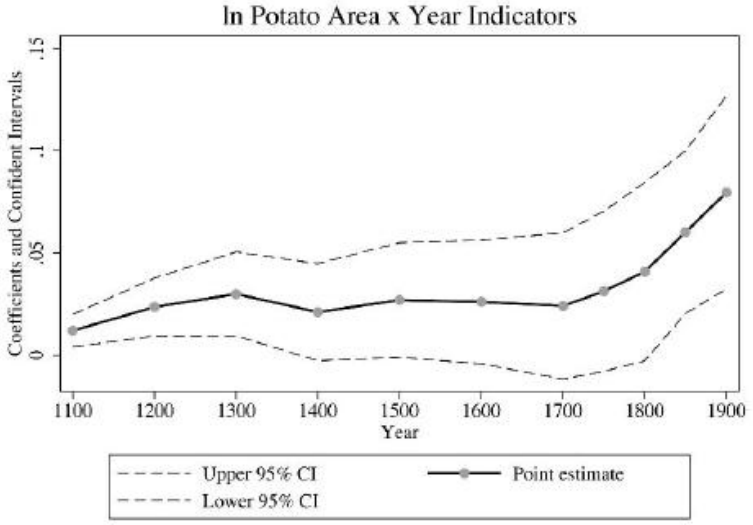
- Notes on the dynamic specification
  - a >=0 and b >= 0 are the numbers of included "leads" and "lags" of the event indicator, respectively.
  - The first lead, 1 [$K_{it} = 1$], is often excluded as a normalization, while the coefficients on the other leads (if present) are interpreted as measures of "pre-trends".
  - The coefficients on the lags are interpreted as a **dynamic path of causal effects.**

- Testing common trends assumption and visualization

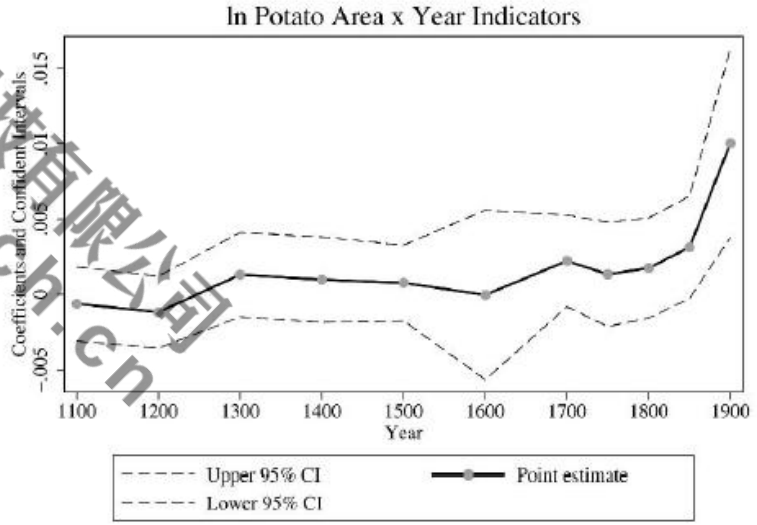
Figure 5: Effects of the HRS on Grain Output per Unit Area



Notes: Data on the timing of the HRS implementation are collected from county gazettes; the grain output data are from the Ministry of Agriculture and Rural Affairs of the People's Republic of China. Plotted are coefficients of the HRS variable and its lagged and forwarded terms. Specifically, we estimate the following equation: $Grain_{ct} = \gamma_0 + \sum_l \beta_l D_{ctl}^{HRS} + \lambda_c + \sigma_t + \varepsilon_{ct}$*, where* $Grain_{ct}$ *denotes the grain output per unit area (in log) in county c and year t;* D_{ctl}^{HRS} *is a set of dummies that denote the number of years relative to the HRS implementation in county c (e.g., the variable* D_{ct1}^{HRS} *denotes whether it is the year after the HRS, while* D_{ct0}^{HRS} *denotes whether it is the year when the HRS was implemented). Plotted are coefficients,* β_l*, with corresponding 95 percent confidence intervals.*



(a) ln Total Population



(b) City Population Share

FIGURE IV
Flexible Estimates of the Relationship between Potato-Suitable Land and Either Total Population or City Population Share

- **2.2 DiD using cross-sectional data, e.g. cohort DiD**
  - Duflo(2001, AER)

TABLE 3—MEANS OF EDUCATION AND LOG(WAGE) BY COHORT AND LEVEL OF PROGRAM CELLS

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Notes: The sample is made of the individuals who earn a wage. Standard errors are in parentheses.

- Model specification of cohort DiD
  - Static specification

$$(1) \quad S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + (P_j T_i)\gamma_1 + (\mathbf{C_j} T_i)\delta_1 + \varepsilon_{ijk}$$

  - Dynamic specification

$$(2) \quad S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + \sum_{l=2}^{23} (P_j \times d_{il})\gamma_{1l} + \sum_{l=2}^{23} (\mathbf{C_j} \times d_{il})\delta_{1l} + \varepsilon_{ijk}$$

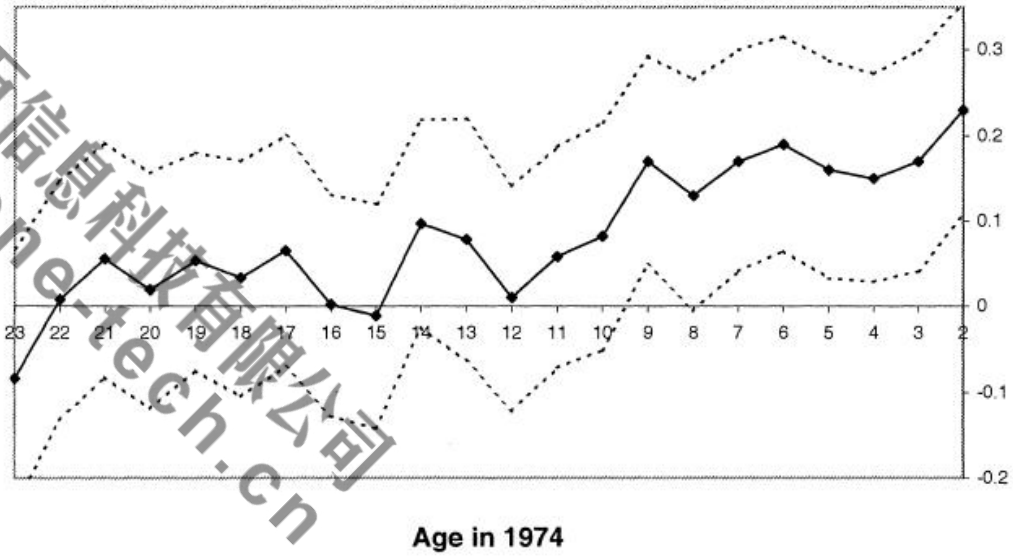


FIGURE 1. COEFFICIENTS OF THE INTERACTIONS AGE IN 1974* PROGRAM INTENSITY IN THE REGION OF BIRTH IN THE EDUCATION EQUATION

- **2.3 Triple differences**
  - Gruber (1994, AER), treatment relies on state (j), year (t), and demographic group (women aged 20-40, i)

$$
\begin{aligned}
(1)\quad W_{ijt} = {} & \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j \\
& + \beta_4 \mathrm{TREAT}_i + \beta_5 (\delta_j \times \tau_t) \\
& + \beta_6 (\tau_t \times \mathrm{TREAT}_i) \\
& + \beta_7 (\delta_j \times \mathrm{TREAT}_i) \\
& + \beta_8 (\delta_j \times \tau_t \times \mathrm{TREAT}_i).
\end{aligned}
$$

- Gruber (1994, AER)

$$(1)\quad W_{ijt} = \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 \mathrm{TREAT}_i + \beta_5(\delta_j \times \tau_t) + \beta_6(\tau_t \times \mathrm{TREAT}_i) + \beta_7(\delta_j \times \mathrm{TREAT}_i) + \beta_8(\delta_j \times \tau_t \times \mathrm{TREAT}_i).$$

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

Notes: Cells contain mean log hourly wage for the group identified. Standard errors are given in parentheses; sample sizes are given in square brackets. Years before/after law change, and experimental/nonexperimental states, are defined in the text. Difference-in-difference-in-difference (DDD) is the difference-in-difference from the upper panel minus that in the lower panel.

- DDD: 2 DiDs; DDDD: 4 DiDs

Table 3: Effects of the NRPS on Receipt of Private Transfers and Household Expenditures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Rural *hukou*				Urban *hukou*		
VARIABLES	Received private transfer (yes = 1)	Log(Received private transfer)	Log(HH total exp)	Log(HH food exp)	Received private transfer (yes = 1)	Log(Received private transfer)	Log(HH total exp)	Log(HH food exp)
Panel A: Age-eligible group (60+)								
Mean of Y	0.39	6.68	9.55	8.55	0.38	7.43	10.2	9.3
$NRPS_{ct}$	0.005	0.170	0.058	0.096*	0.012	0.175	-0.000	0.044
	(0.031)	(0.110)	(0.043)	(0.058)	(0.031)	(0.157)	(0.039)	(0.039)
Observations	15,833	6,098	15,429	15,906	6,812	2,582	6,633	6,923
R-squared	0.164	0.226	0.204	0.262	0.221	0.362	0.275	0.310
F-statistic	–	–	–	–	0.03	0.00	1.20	0.65
P-value	–	–	–	–	0.86	0.97	0.27	0.42
Panel B: Age-ineligible group (45-59)								
Mean of Y	0.43	7.00	9.91	8.76	0.36	7.54	10.39	9.30
$NRPS_{ct}$	0.001	-0.001	-0.003	0.036	0.031	-0.026	-0.012	0.009
	(0.027)	(0.103)	(0.032)	(0.051)	(0.023)	(0.153)	(0.036)	(0.039)
Observations	22,456	9,697	22,151	22,702	8,302	3,001	8,275	8,477
R-squared	0.278	0.254	0.206	0.284	0.273	0.345	0.278	0.294
F-statistic	0.04	2.20	2.25	1.60	–	–	–	–
P-value	0.85	0.14	0.13	0.21	–	–	–	–

Note: The data are from the CHARLS and CFPS for individuals ages 45 years and older. The covariates in the regressions in each column include age and its square, and dummies for gender, education level, survey year, and county. All the standard errors are clustered at the county level.

**** p<0.01, ** p<0.05, * p<0.1.*

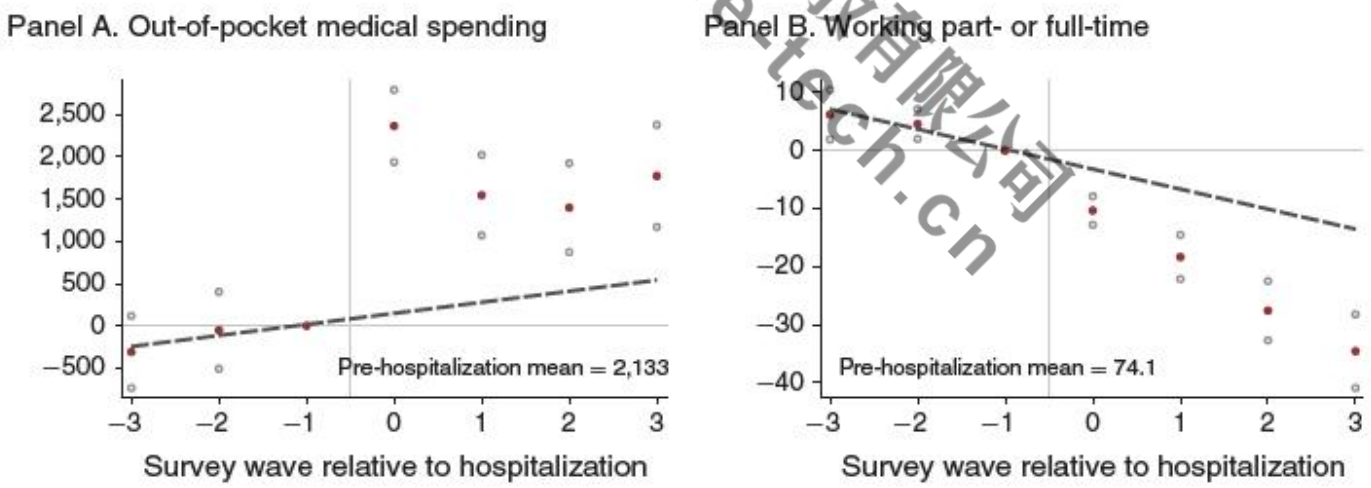
- **2.4 An event study approach**
  - An event study design is a staggered adoption design where units are treated at different times, and there may or may not be never treated units. It also nests a difference-in-differences design, where units are either first treated at time t_0 or never treated (Sun and Abraham, 2021)

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell=-K}^{-2} \mu_\ell D_{i,t}^{\ell} + \sum_{\ell=0}^{L} \mu_\ell D_{i,t}^{\ell} + \upsilon_{i,t}$$

- Dobkin et al.(2018, AER)

$$(3) \qquad y_{it} = \gamma_t + X_{it}\alpha + \sum_{r=S}^{-2}\mu_r + \sum_{r=0}^{F}\mu_r + \varepsilon_{it}$$

  - Identifying assumption: The **timing of the event** (hospital admission) is uncorrelated with the outcome, conditional on having a hospital admission during observation window and the included controls.
  - An admission that is preceded by deteriorating health, or an admission caused by the adverse health effects of job loss would violate this assumption.



- Standard TWFE in event study (Sun and Abraham, 2021)

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell=-K}^{-2} \mu_\ell D_{i,t}^\ell + \sum_{\ell=0}^{L} \mu_\ell D_{i,t}^\ell + \upsilon_{i,t}$$

  - You need to exclude some relative periods from the 'fully dynamic' specification to avoid multi-collinearity either among the relative period indicators, or with the unit and time fixed effects.
  - When there are no never treated units but with a panel balanced in calendar time, we need to exclude **at least two relative period indicators**.
    - One multi-collinearity comes from the relative period indicators summing to one for every unit
    - The other multi-collinearity comes from the linear relationship between two-way fixed effects and the relative period indicators
  - Excluding relative periods close to the initial treatment is common in practice. Normalizing relative to the period prior to treatment is the most common.

- **2.5 One related approach: Synthetic control (Abadie et al. 2015, APSR)**

FIGURE 1 Trends in per Capita GDP: West Germany versus Rest of the OECD Sample

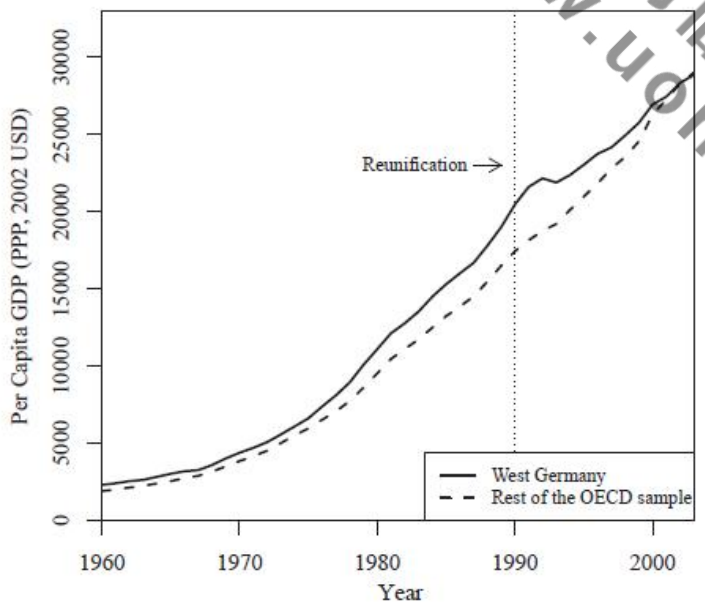
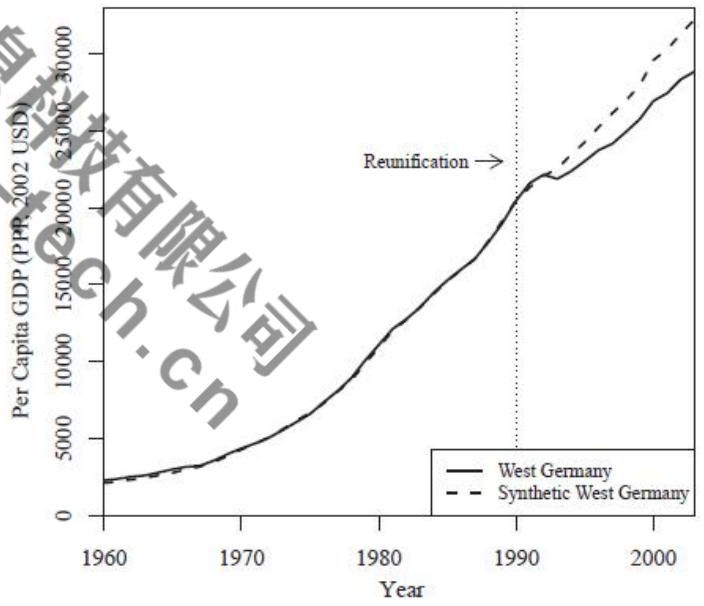


FIGURE 2 Trends in per Capita GDP: West Germany versus Synthetic West Germany



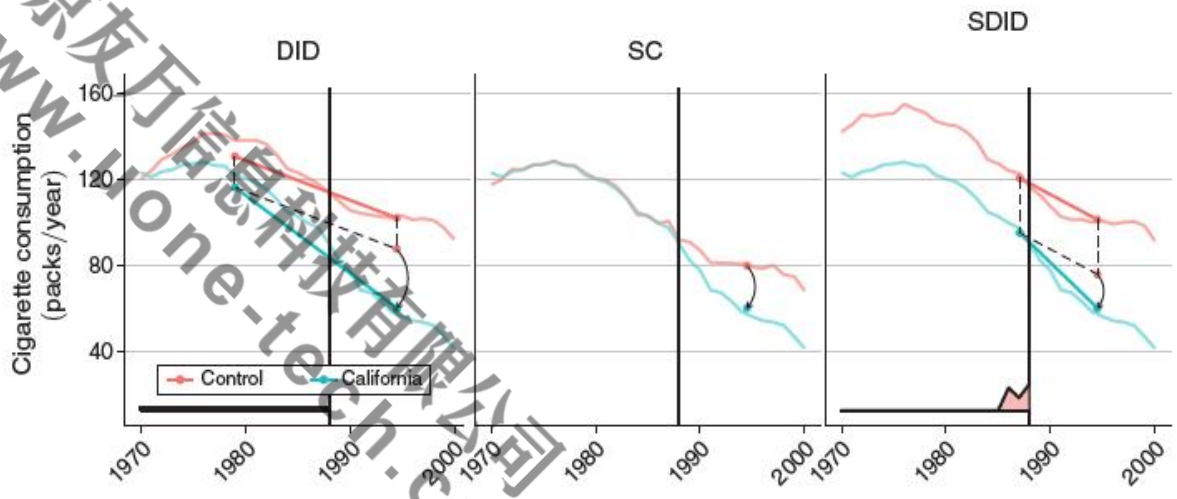
- **Synthetic control: What is new?**
  - **Construct a control group** (the synthetic control) based on preintervention observables (**including the outcome variable**), assigning different weights to different untreated units
    - DiD assigns uniform weight to the untreated units
    - Matching is usually based on X only, and uses post-treatment information.
  - Matching based on preintervention X (and Y), then conduct weighted DiD

- **Synthetic DiD** (Arkhangelsky et al., 2021, AER)

$$\hat{\delta}_i^{sc} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it},$$

$$\hat{\delta}_i^{did} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \frac{1}{T_{pre}} \sum_{t=1}^{T_{pre}} Y_{it},$$

$$\hat{\delta}_i^{sdid} = \frac{1}{T_{post}} \sum_{t=T_{pre}+1}^{T} Y_{it} - \sum_{t=1}^{T_{pre}} \hat{\lambda}_t^{sdid} Y_{it}.$$



- **2.6 Another related approach: Bartik Instruments**
  - The **Bartik instrument** (named after Bartik (1991)) is formed by interacting local industry shares and national industry growth rates.
  - It is always possible to construct a Bartik instrument.

- **Interpret the Bartik instrument**
  - Goldsmith-Pinkham , Sorkin and Swift (2020, AER)
    - Emphasizing the **exogeneity of exposure shares**.
    - The local industry shares as instruments and a weight matrix constructed from the national growth rates.
  - **Borusyak**, Hull and **Jaravel** (2021, Restud)
    - Emphasizing the **quasi-random assignment of shocks**, while exposure shares are allowed to be endogenous.
    - The outcome and treatment variables are first averaged over the level of shocks, using exposure shares as weights, to obtain shock-level aggregates. The shocks then directly instrument for the aggregated treatment.

- **Interpret the Bartik instrument**
  - The implied empirical strategy is **an exposure research design**, where the industry shares measure the differential exogenous exposure to the common shock (national industry growth)
  - **With a pretreatment period, this empirical strategy is just difference-in-differences**
    - We do not need to assume that the shares are uncorrelated with the levels of the outcome. Instead, the strategy asks whether differential exposure to common shocks leads to differential changes in the outcome
  - Questions: So what is the identification assumption? How to test?

- Autor, Dorn and Hanson. (2013, AER)
  - **Import competition per worker**, weighting import change using industrial shares, normalized by local employment size.

$$\Delta IPW_{uit} = \sum_j \frac{L_{ijt}}{L_{ujt}} \frac{\Delta M_{ocjt}}{L_{it}}.$$

  - ADH worry that there are unobservable industry shocks (e.g. technological changes) correlated with industry-level import competition. They purge their industry shocks from U.S.-specific confounders by measuring Chinese import growth outside of the U.S.

- To interpret the Bartik instrument as a DiD design, consider a case with only one or two "industries" and two time periods.
  - The research question becomes "whether locations with high shares of a particular industry experience differential changes in outcomes following shocks whose effect depends on the size of that industry."

3. Heterogeneity

- 3.1 Interpretation and problems of the TWFE estimators
  - 3.1.1 Variations in treatment timing
    - 3.1.1.1 Static model
    - 3.1.1.2 Dynamic model
  - 3.1.2 Variations in treatment intensity
    - Leve effects or slope effects
- 3.2 Some new estimators

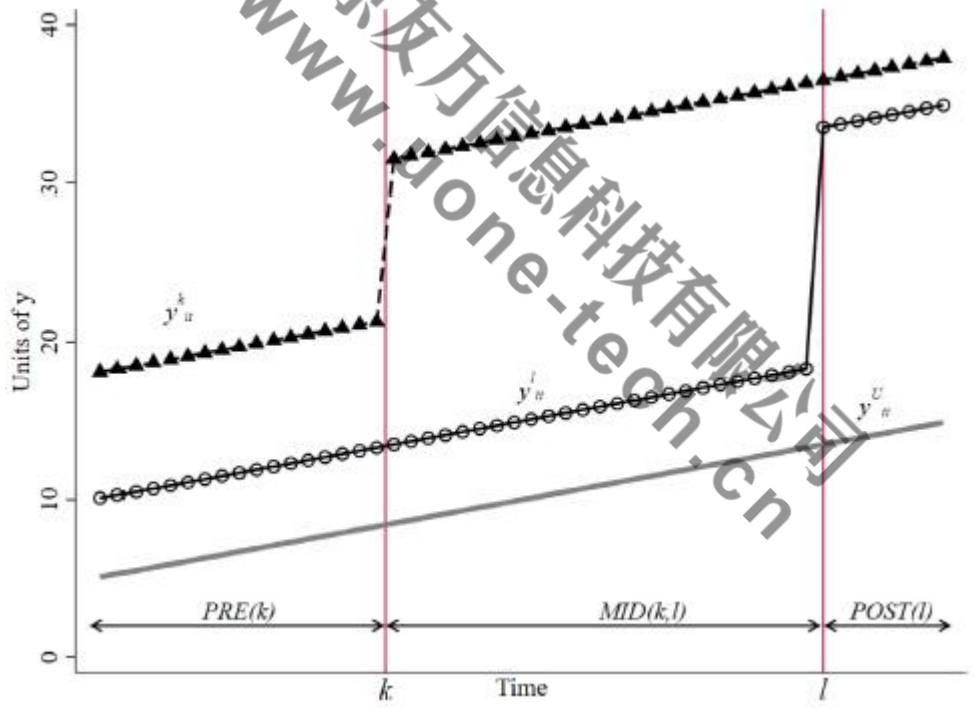
3.1.1 Variations in treatment timing

- **3.1.1.1 TWFE in static models**

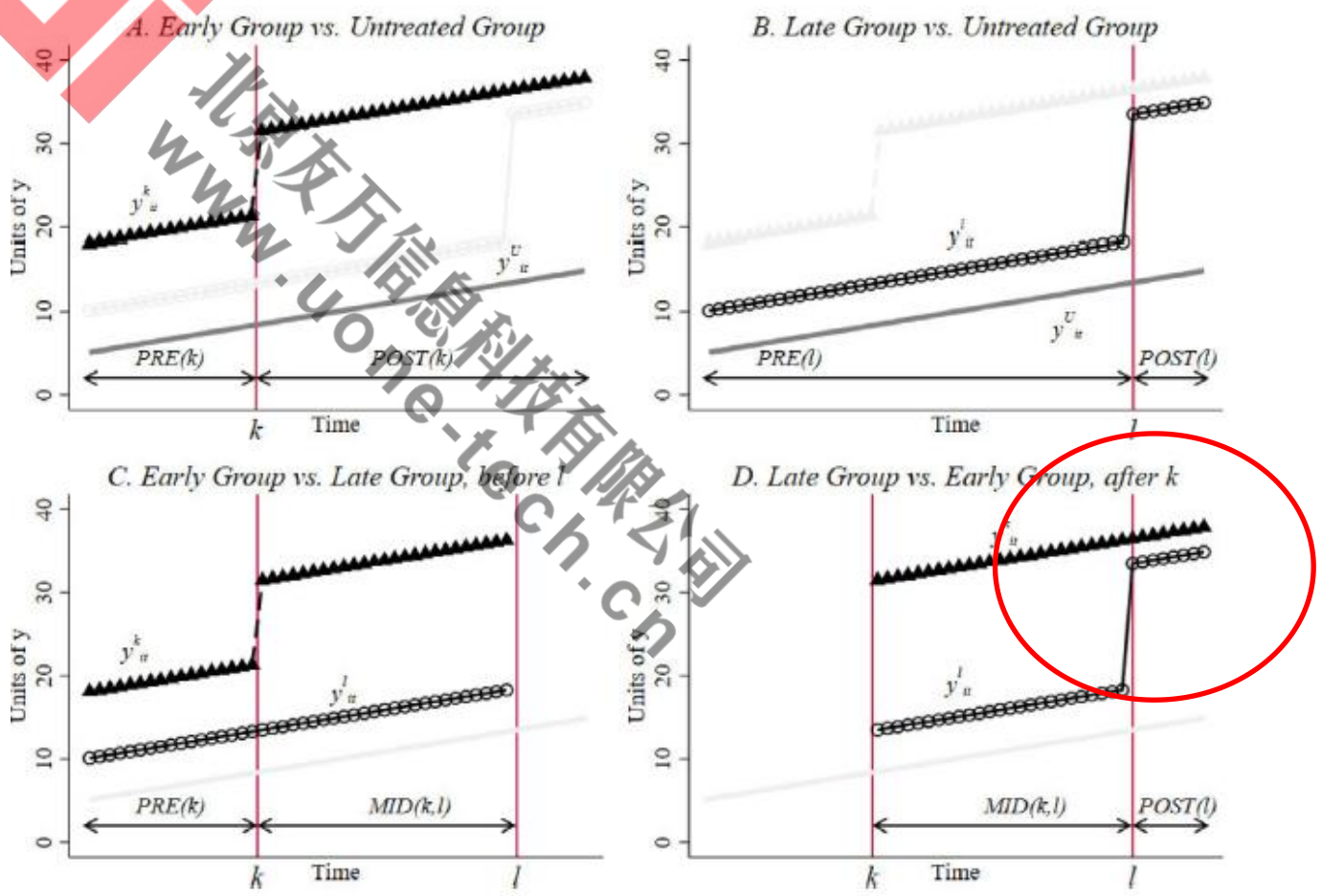
- Goodman-Bacon(2021)
  - The **TWFEDD** is a **weighted average of all possible 2x2 DD estimators** that compare **timing groups** to each other.
    - Some use units treated at a particular time as the treatment group and untreated units as the control group.
    - Some compare units treated at two different times, **using the later-treated group as a control** before its treatment begins and then **the earlier-treated group as a control** after its treatment begins.

- **Forbidden comparisons** and the "**negative weights**" problem.
  - **Negative weights** arise when average treatment effects vary over time (i.e. heterogeneous treatment effects across time).
  - When **already-treated units act as controls**, changes in their outcomes are subtracted and these changes may include time-varying treatment effects.

- Three time periods, three groups.



- Four simple 2x2 DiDs



- ***ΔATT***
  - A weighted sum of **the change in treatment effects** within each timing group's before and after a later treatment time

$$\Delta ATT \equiv \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell}^{\ell} \left[ATT_k \left(POST\left(\ell\right) \right) - ATT_k \left(MID\left(k, \ell\right) \right) \right]$$

- When **use already-treated groups as controls**, the 2x2 DD subtract average changes in their untreated outcomes **and their treatment effects**.

- ***ΔATT*** is the source of the negative weights discussed in de Chaisemartin and D'Haultfoeuille(2020).

- ***ΔATT*** equals zero if average treatment effects are constant.

- Units that are **treated throughout the sample** can only ever act as controls (they enter into the decomposition theorem exactly like never-treated units), so if their treatment effects are changing during the sample periods they will also contribute to ***ΔATT***.
  - 2x2 DDs in which always-treated units are the control group use all time periods, so they get higher weight. If their treatment effects are changing they can substantially bias **TWFEDD** away from **VWATT**.

- **Diagnosis**
  - Conduct the decomposition suggested by Goodman-Bacon(2021), using a STATA package, **bacondecom**

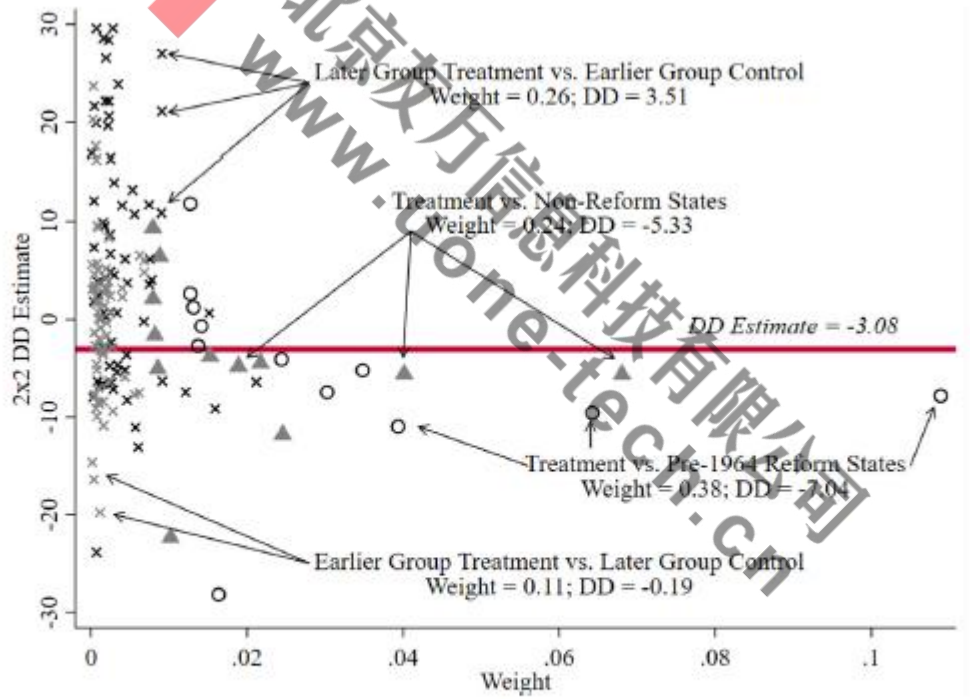


Fig. 6. Difference-in-differences decomposition for unilateral divorce and female suicide. Notes: The figure plots each 2x2 DD components from the decomposition theorem against their weight for the unilateral divorce analysis. The open circles are terms in which one timing group acts as the treatment group and the pre-1964 reform states act as the control group. The closed triangles are terms in which one timing group acts as the treatment group and the non-reform states act as the control group. The *x*'s are the timing-only terms. The figure notes the average DD estimate and total weight on each type of comparison. The two-way fixed effects estimate, −3.08, equals the average of the *y*-axis values weighted by their *x*-axis value.

Stata package

Syntax

bacondecomp outcome treatment

bacondecomp asmrs post , ddetail

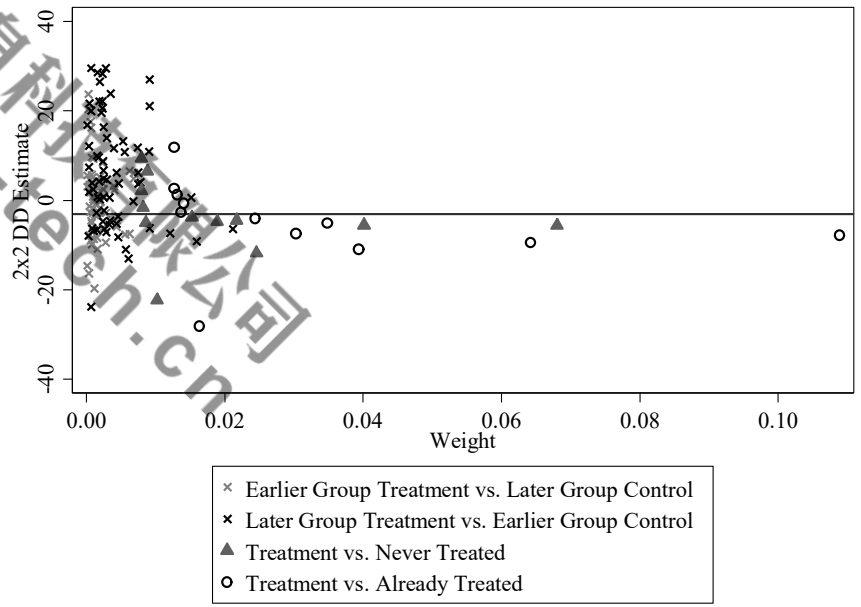
Calculating treatment times...
Calculating weights...
Estimating 2x2 diff-in-diff regressions...

Diff-in-diff estimate: -3.080

DD Comparison	Weight	Avg DD Est
Earlier T vs. Later C	0.111	-0.187
Later T vs. Earlier C	0.265	3.512
T vs. Never treated	0.240	-5.331
T vs. Already treated	0.384	-7.044

T = Treatment; C = Control

Example: Stevenson and Wolfers' (2006)



- **3.1.1.2 TWFE in dynamic models**

- Sun and Abraham (2021)

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell} \mu_\ell \mathbf{1}\{t - E_i = \ell\} + \upsilon_{i,t}$$

  - Units can be categorized into different cohorts based on their initial treatment timing.

  - **Conclusions**
    - The coefficients on a given lead or lag (μ_l) can be expressed as **a linear combination of cohort-specific effects from both its own relative period and other relative periods**.
    - The terms that include treatment effects from other relative periods will not cancel out with **heterogeneous treatment effects** and will contaminate the estimate of μ_l.

- The coefficients on a given lead or lag (μ_l) can be expressed as a linear combination of cohort-specific effects from both its own relative period and other relative periods.
  - Using estimates of treatment leads in a dynamic model as a way of testing for parallel pretrends is problematic.
  - The estimate of μ_ℓ is affected by both pretrends and treatment effects heterogeneity.

- Sun and Araham (2021) define **the cohort-specific average treatment effect on the treated (CATT)** *l* periods from initial treatment,

$$CATT_{e,\ell} = E[Y_{i,e+\ell} - Y^{\infty}_{i,e+\ell} \mid E_i = e]$$

- The cohort is defined by the time at which the cohort was initially treated, *e*.
- The authors consider a TWFE model,

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{g \in \mathcal{G}} \mu_g \mathbf{1}\{t - E_i \in g\} + \nu_{i,t}$$

$$\mathbf{1}\{t - E_i \in g\} = \sum_{\ell \in g} \mathbf{1}\{t - E_i = \ell\} = \sum_{\ell \in g} D^{\ell}_{i,t} \qquad D^{\ell}_{i,t} := \mathbf{1}\{t - E_i = \ell\}$$

- Static specification,

$$Y_{i,t} = \alpha_i + \lambda_t + \mu_g \sum_{\ell \geq 0} D_{i,t}^{\ell} + \upsilon_{i,t}$$

- The more conventional model specification (dynamic specification) of event study,

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell=-K}^{-2} \mu_\ell D_{i,t}^{\ell} + \sum_{\ell=0}^{L} \mu_\ell D_{i,t}^{\ell} + \upsilon_{i,t}$$

  - Exclude relative period: {-T, …, -K-1, -1, L+1, …, T}
- Sometimes researchers bin or trim distant relative periods, instead of excluding them,

$$Y_{i,t} = \alpha_i + \lambda_t + \beta \cdot \sum_{\ell<-K} D_{i,t}^{\ell} + \sum_{\ell=-K}^{-2} \mu_\ell D_{i,t}^{\ell} + \sum_{\ell=0}^{L} \mu_\ell D_{i,t}^{\ell} + \gamma \cdot \sum_{\ell>L} D_{i,t}^{\ell} + \upsilon_{i,t}$$

- With the **parallel trends assumption**,

$$\mu_g = \sum_{\ell \in g} \sum_{e} \omega_{e,\ell}^{g} CATT_{e,\ell} + \sum_{g' \neq g, g' \in \mathcal{G}} \sum_{\ell' \in g'} \sum_{e} \omega_{e,\ell'}^{g} CATT_{e,\ell'} + \sum_{\ell' \in g^{excl}} \sum_{e} \omega_{e,\ell'}^{g} CATT_{e,\ell'}$$

- Since

$$E[Y_{i,e+\ell} - Y_{i,0}^{\infty} | E_i] - E[Y_{i,e+\ell}^{\infty} - Y_{i,0}^{\infty}] = CATT_{e,\ell} + \underbrace{E[Y_{i,t}^{\infty} - Y_{i,0}^{\infty} | E_i] - E[Y_{i,t}^{\infty} - Y_{i,0}^{\infty}]}_{=0}$$

- **Without anticipatory behavior**, pre-treatment $CATT_{e,\ell<0}$ is zero, and thus,

$$\mu_g = \sum_{\ell' \in g, \ell' \geq 0} \sum_{e} \omega_{e,\ell'}^{g} CATT_{e,\ell'} + \sum_{g' \neq g, g' \in \mathcal{G}} \sum_{\ell' \in g', \ell' \geq 0} \sum_{e} \omega_{e,\ell'}^{g} CATT_{e,\ell'} + \sum_{\ell' \in g^{excl}, \ell' \geq 0} \sum_{e} \omega_{e,\ell'}^{g} CATT_{e,\ell'}$$

- With **homogeneous treatment effect assumption**,

$$\mu_\ell = ATT_\ell + \sum_{\ell' \in g^{excl}} \omega_{\ell'}^{\ell} ATT_{\ell'}$$

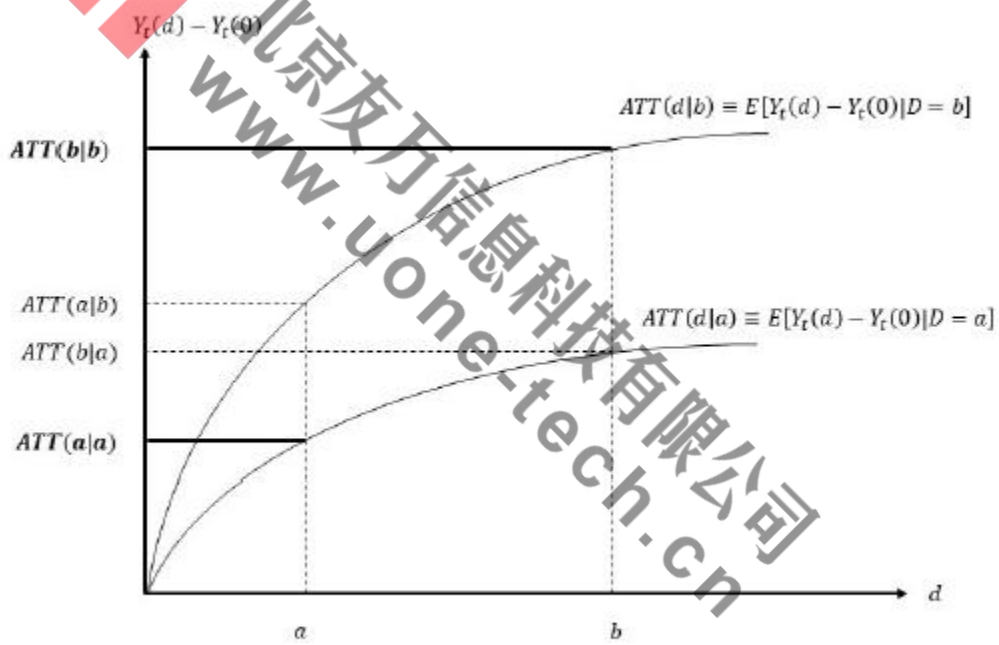
- The **constant ATT**$_{\ell'}$ s from other relative periods cancel out because of their weights are summed to zero.
- Even under the homogeneous treatment effect assumption, the coefficient μ_l can still be contaminated by treatment effects from the excluded periods.
  - This contamination can be avoided by adjusting the specification to only exclude periods with zero treatment effect.

- **Intuition behind the contamination**
  - In an event study where individuals receive the treatment at different times, **the panel can never be balanced in both calendar time and time relative to the initial treatment**.
  - The relative time indicators are correlated even after controlling for unit and time fixed effects in a TWFE regression.
  - Consider a true model with some **missing regressors, $D_{l'}$**
    - The omitted variable bias formula implies that there will be a bias equals the interaction between the coefficients on the missing regressors (**ATT**$_{l'}$) and the estimate (the weight) from an auxiliary regression .

3.1.2 Variations in treatment intensity

- Callaway, Goodman-Bacon and Sant'Anna (NBER WP32117, 2024)
  - Many DiD applications study treatments that do not simply turn "on", they have a "dose" or operate with varying intensity.
  - Two types of causal effects arise in a non-binary DiD setting:
    - **The level effect**: the treatment effect of "dose" **d**, which equals the difference between a unit's potential outcome under treatment **d** and its untreated potential outcome.
    - **The slope effect**: the causal response to an incremental change in the "dose" at **d**.

Figure 1: Average Treatment Effects on the Treated, Two Doses



Notes: The figure plots $ATT(d|a)$ (the average effect of experiencing dose d among units that actually experienced dose a) and $ATT(d|b)$ (the average effect of experiencing dose d among units that actually experienced dose b).

3.2 Some new estimators

- **3.2.1 Estimating only instantaneous treatment effects**
- **3.2.2 Estimating weighted treatment effect based on some "building blocks"**
  - Group-time ATT
  - Cohort-specific ATT
- **3.2.3 Two-stage estimators/imputation estimators**

3.2.1 Estimating only instantaneous treatment effects

- de Chaisemartin and D'Haultfoeuille (2020)
  - Focus on the ATE of all switching cells, the **leavers** or **joiners**.
  - Defining the averaged treatment effect (**ATE**) for switching cells,

$$\delta^S = E\left[\frac{1}{N_S}\sum_{(i,g,t):t\geq 2, D_{g,t}\neq D_{g,t-1}}\left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]\right]$$

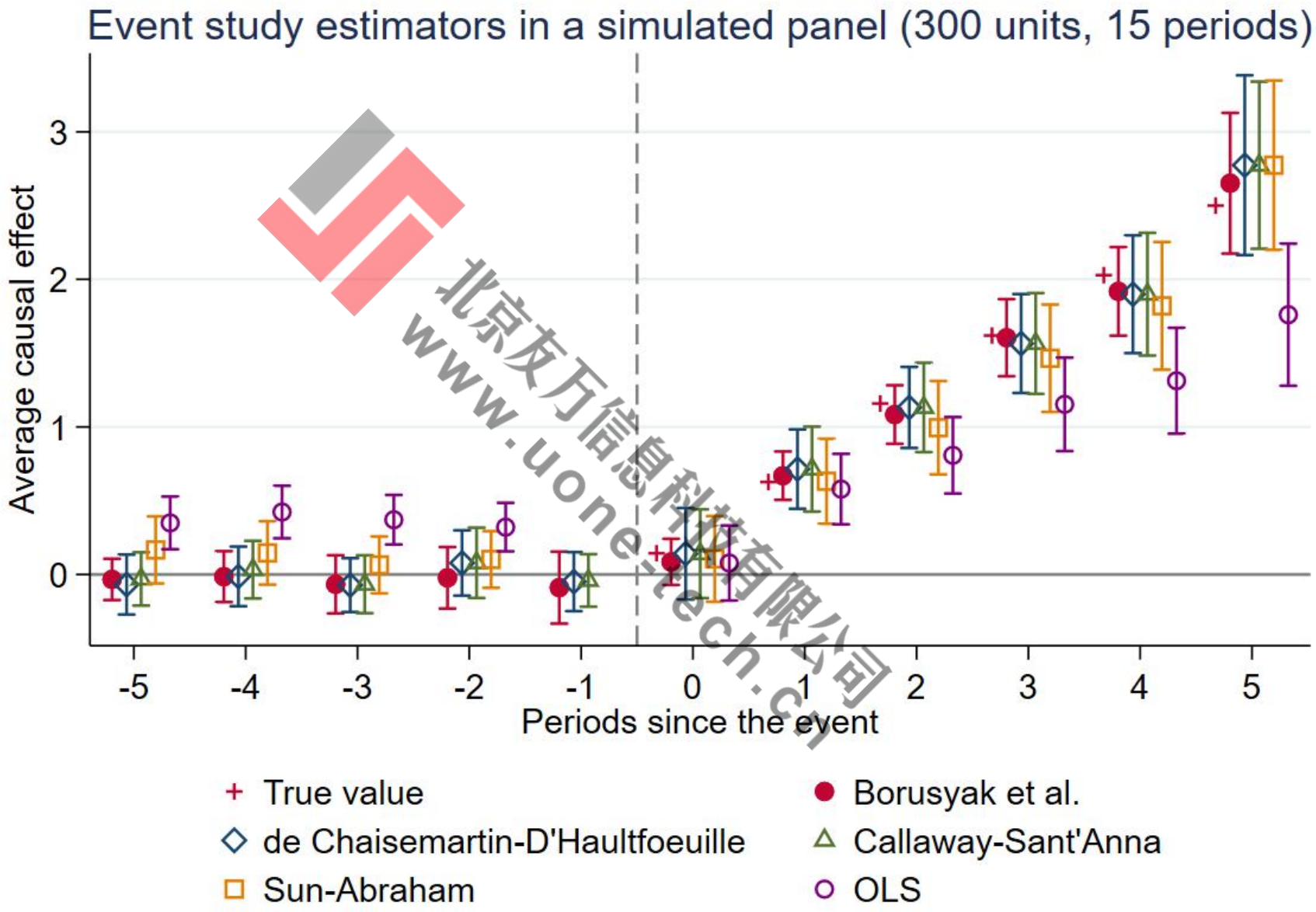
3.2.2 Estimating weighted treatment effect based on some "building blocks"

- Callaway and Sant'Anna (2021)
  - **Group-time specific treatment effect**
- Sun and Abraham (2021)
  - **Cohort-specific treatment effect**

- Basically, these estimators separate the DiD estimation into two steps:
  - Identification of disaggregated causal effects, i.e., the building blocks.
  - Aggregating (some of) these disaggregated causal effects to form summary measures of the causal effects.

3.2.3 Two-stage estimators/imputation estimators

- Borusyak, Jaravel and Spiess (2022, CEPR Discussion Paper No. DP17247)
  - Estimate unit-specific treatment effect.
    - Use untreated observations to parametrically identify the unit and period fixed effects, then impute the untreated potential outcomes of each treated observation.
  - Aggregate unit-specific treatment effects with some reasonable weights to obtain the estimation target.



RDD

1. Designs and Parameters

- Canonical RD settings
- Multidimensional RD designs
- Related designs

2. Estimation and inference

- Local polynomial regression methods
- Experiments methods

3. Validation and falsification

1. Designs and Parameters

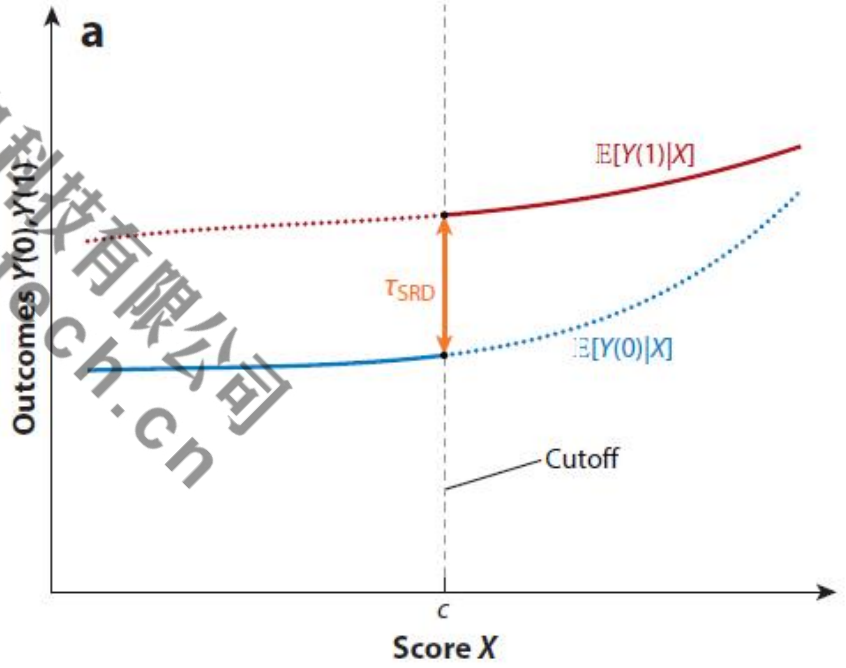
- Canonical RD settings
  - Sharp RD
  - Fuzzy RD
- Multidimensional RD designs
  - Multi-cutoff, multi-score, geographic, multiple-treatment, time-varying designs
- Related designs
  - Kink, bunching, before-after, threshold regression designs

1.1 Canonical RD Settings

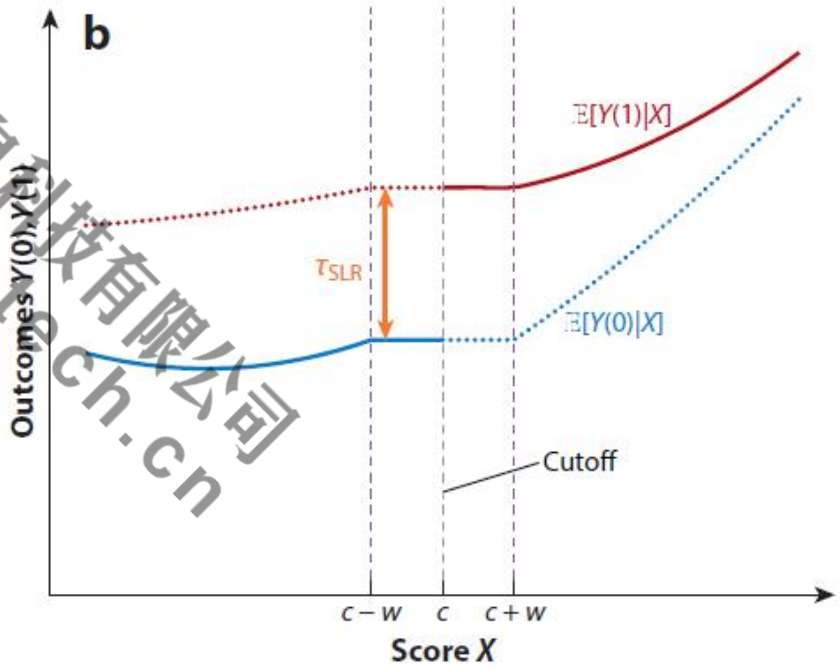
- RD design
  - All the units in the study are assigned a value of the score (also called a running variable or index), and the treatment is assigned only to units whose score value exceeds a known cutoff (also called threshold)
  - The probability of treatment assignment changes from zero to one at the cutoff
  - The most important threat
    - The possibility that units might be able to strategically and precisely change their score to be assigned to their preferred treatment condition (Lee 2008,McCrary 2008)

- RD design
  - To study causal RD treatment effects, the score, cutoff, and treatment assignment rule must exist ex-ante and be well defined
  - The treatment assignment rule is known and verifiable

- **Sharp Designs**
  - The treatment assigned and the treatment received coincide for all units (perfect compliance or focus on ITT)
  - **Continuity-based framework** (Hahn et al., 2001)



- **Sharp Designs**
  - The treatment assigned and the treatment received coincide for all units (perfect compliance or focus on ITT)
  - **Local randomization framework** (Lee, 2008; Lee & Lemieux, 2010)



- **Fuzzy Designs**
  - The treatment assigned and the treatment received do not coincide for at least some units (imperfect compliance)
  - **Fuzzy RD is IV** (Angrist and Pischke, 2009)

1.2 Multidimensional Designs

- **Multi-score RD design**
  - Two or more scores assigning units to a range of different treatment conditions
  - The score is multidimensional but the treatment is still binary
    - **Geographic RD design**: the RD score is two dimensional to reflect each unit's position in space, usually latitude and longitude (Dell, 2010)

- Dell M. 2010. The persistent effects of Peru's mining mita. Econometrica 78(6):1863–903

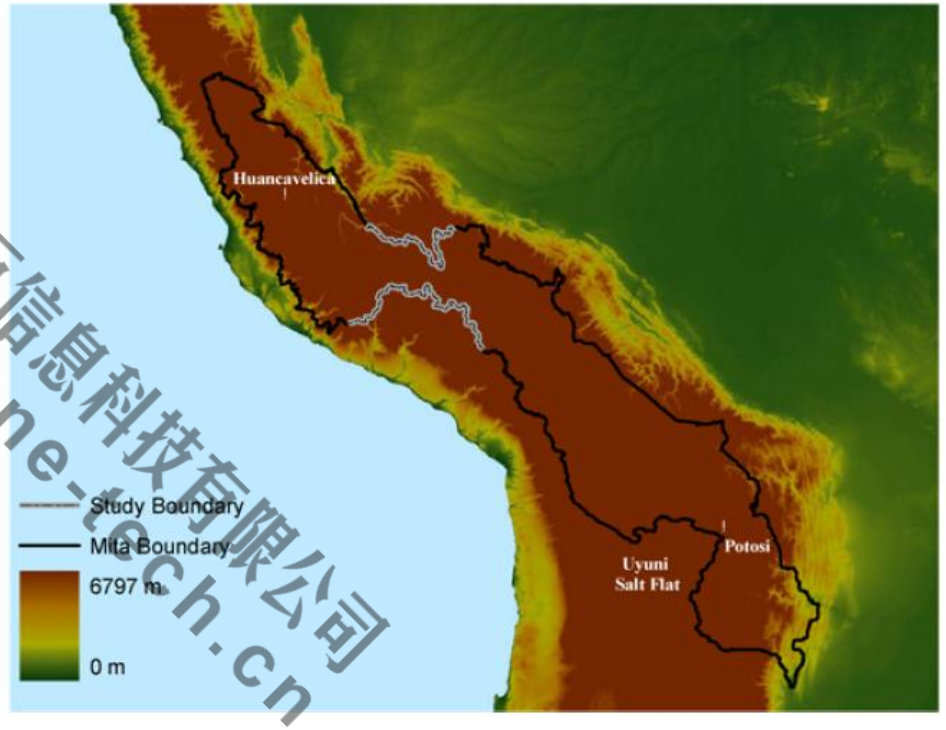


FIGURE 1.—The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.

- **Multi-cutoff RD design**
  - Different units in the study receive the treatment according to different cutoff values along a univariate score
    - We can normalize and pool the data along the treatment assignment boundary curve or the multiple cutoff values to consider a single, pooled RD treatment effect (Cattaneo MD, Idrobo N, Titiunik R. 2022a)

- **Multi-valued/continuous treatment**
  - RD causal effects can be identified based on changes in the probability distribution of the continuous treatment at the cutoff (Dong et al. 2021)

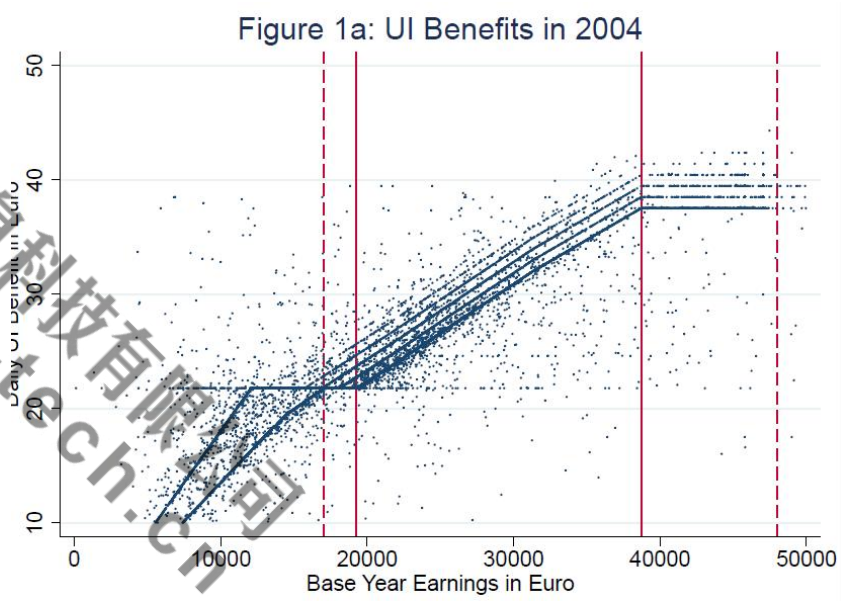
- **Time-varying designs**
  - Difference-in-discontinuities design (Grembi et al., 2016)

$$\hat{\tau}_{DD} \equiv (Y^- - Y^+) - (\tilde{Y}^- - \tilde{Y}^+)$$

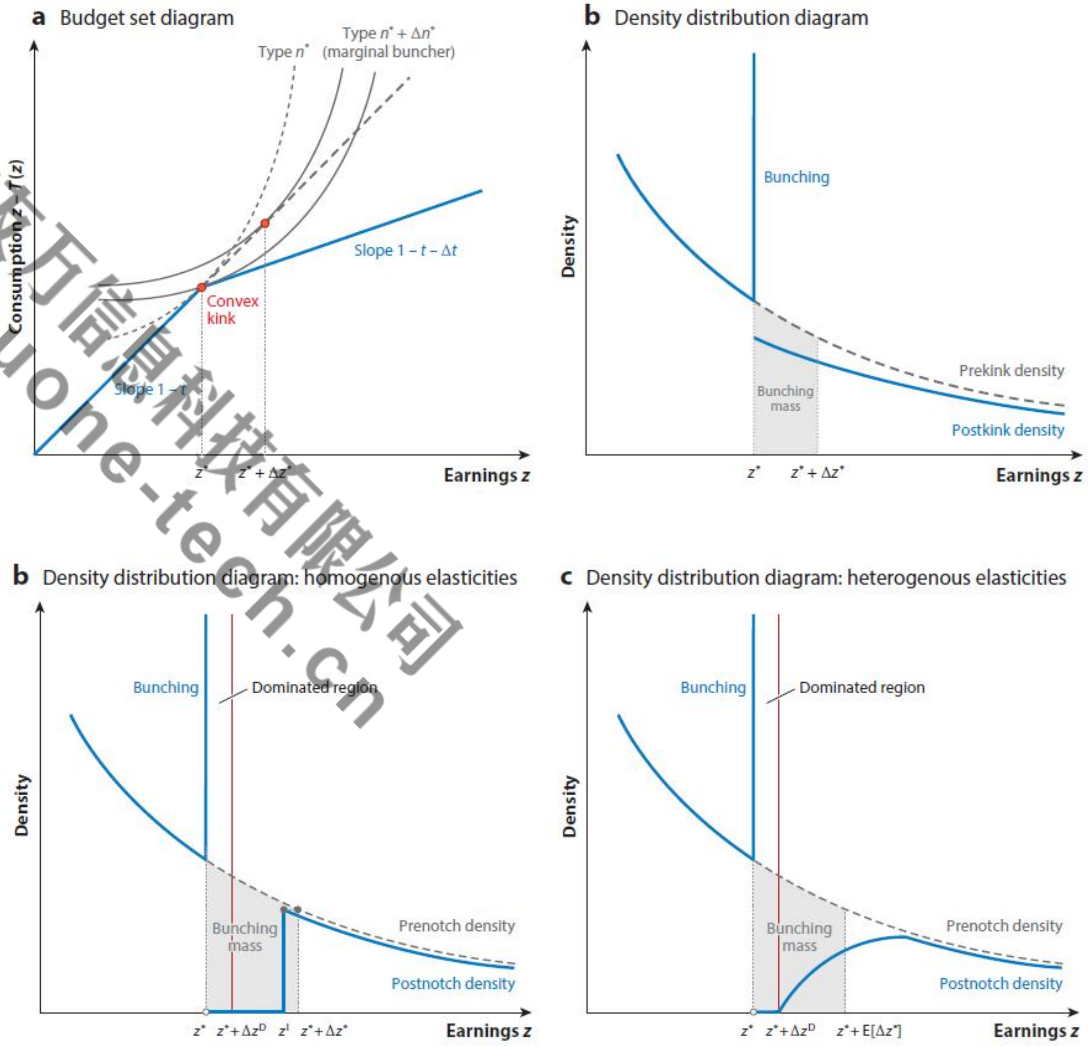
- RD designs have high **internal validity** but low **external validity**
  - In the absence of additional assumptions, it is not possible to learn about treatment effects away from the cutoff
    - Dong & Lewbel (2015) and Cerulli et al. (2017) study local extrapolation methods via derivatives of the RD average treatment effect
    - Angrist & Rokkanen (2015) employ **pre-intervention covariates** under a conditional ignorability condition
    - Rokkanen (2015) relies on **multiple measures of the score**, which are assumed to capture a common latent factor
    - Bertanha & Imbens (2020) exploit **variation in treatment assignment** generated by imperfect compliance

1.3 Related Designs

- **Regression Kink Design** (Card et al., 2015, 2017)
  - The assignment rule that links the treatment and the score is assumed to change slope at a known cutoff point
  - The regression function of the observed outcome will be continuous at all values of the score, but its slope will be discontinuous at the cutoff point
  - Differences of first derivatives of regression functions at the cutoff, or ratios thereof—are referred to as kink RD designs

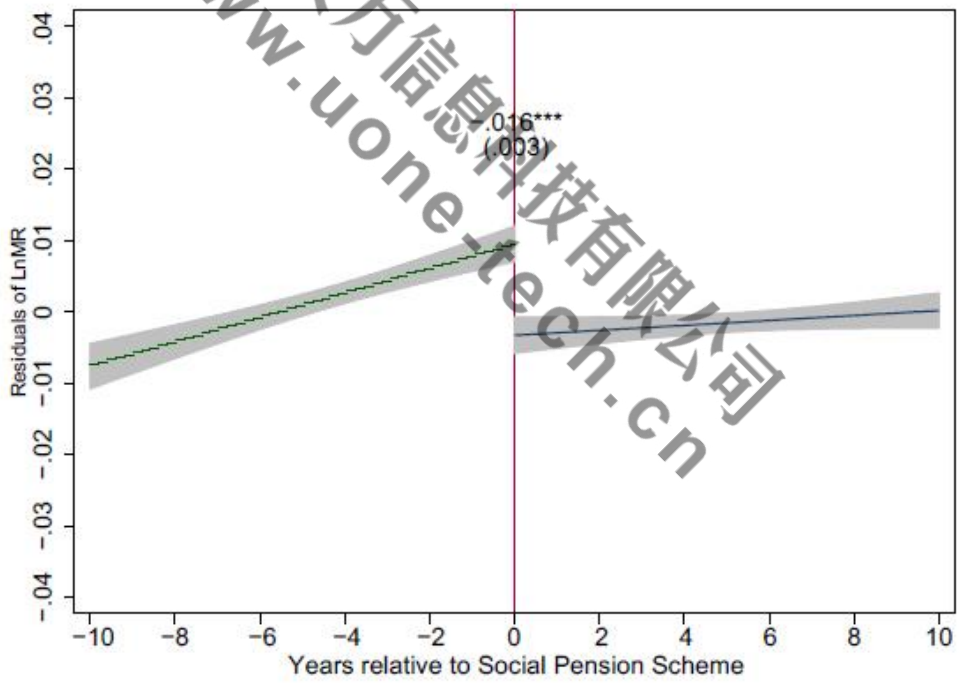


- **Bunching and density discontinuities** (Kleven 2016, Jales & Yu 2017)
  - The objects of interest are related to discontinuities and to other sharp changes in probability density functions
  - Identification (as well as estimation and inference) requires additional parametric modeling assumptions that are invoked for extrapolation purposes



- RD designs in time: Before-and-after analysis/event studies

Figure C1: Regression Discontinuity Estimation for the Effects of Social Pensions on Mortality



(a) Age-eligible group

2. Estimation and Inference

- Visualization and its limitations
  - Global polynomial fit for the outcome on the score (Gelman & Imbens, 2019)
  - Local sample means of the outcome computed in small bins of the score variable
  - **Changing the specification of RD plots while keeping the underlying model constant leads participants to draw different conclusions** (Korting et al., 2021)

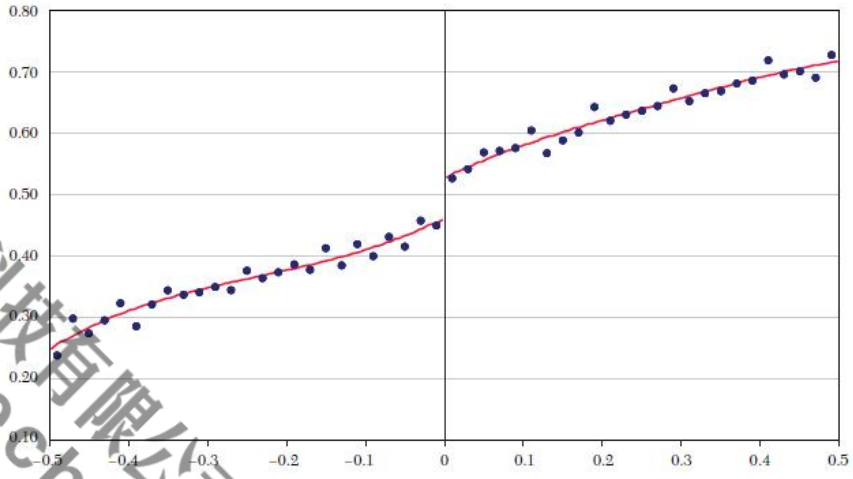


Figure 6. Share of Vote in Next Election, Bandwidth of 0.02 (50 bins)

- Local polynomial methods
  - The standard approach for estimation and inference in the RD design under continuity conditions (Calonico et al., 2014)
  - Local polynomial analysis for RD designs is implemented by fitting Yi on a low-order (*p*) polynomial expansion of Xi, separately for treated and control observations, and in each case using only observations near the cutoff rather than all available observations, as determined by the choice of a kernel function(*k*) or weighting scheme and a bandwidth parameter (*h*)
  - **Choice of bandwidth (*h*) is critical for RD estimation** (Imbens & Kalyanaraman, 2012; Cattaneo & Vazquez-Bare, 2016; Calonico et al., 2014, 2020)

- Local polynomial methods: Estimation

$$\widehat{\beta}_{-} = \underset{b_0,\ldots,b_p}{\arg\min} \sum_{i=1}^{n} \mathbb{1}(X_i < c) \left(Y_i - b_0 - b_1(X_i - c) - b_2(X_i - c)^2 - \cdots - b_p(X_i - c)^p\right)^2 K\left(\frac{X_i - c}{h}\right)$$

and

$$\widehat{\beta}_{+} = \underset{b_0,\ldots,b_p}{\arg\min} \sum_{i=1}^{n} \mathbb{1}(X_i \geq c) \left(Y_i - b_0 - b_1(X_i - c) - b_2(X_i - c)^2 - \cdots - b_p(X_i - c)^p\right)^2 K\left(\frac{X_i - c}{h}\right)$$

$$\widehat{\tau}_{\mathrm{SRD}}(h) = \widehat{\beta}_{+,0} - \widehat{\beta}_{-,0}$$

- Local polynomial methods: Inference
  - The usual confidence interval:

$$I_{\text{LS}} = \left[\hat{\tau}_{\text{SRD}}(h_{\text{MSE}}) \pm 1.96 \cdot \sqrt{\hat{\text{V}}} \right]$$

  - Robust bias-corrected confidence interval:

$$I_{\text{RBC}} = \left[\left(\hat{\tau}_{\text{SRD}}(h_{\text{MSE}}) - \hat{\text{B}} \right) \pm 1.96 \cdot \sqrt{\hat{\text{V}} + \hat{\text{W}}} \right]$$

  - where B denotes the estimated bias correction and W denotes the adjustment in the standard errors.

- Local randomization methods: Estimation
  - Chose a window W where the local randomization is assumed to hold
    - **It is analogous to the bandwidth selection step in the continuity framework**
    - Cattaneo et al. (2015) recommend to select the window based on pre-treatment covariates or placebo outcomes known to be unaffected by the treatment
  - Estimate the average effects as the simple difference-in-means for observations inside W

$$\hat{\tau}_{\text{SLR}} = \bar{Y}_W^+ - \bar{Y}_W^- \quad \text{and} \quad \hat{\tau}_{\text{FLR}} = \frac{\bar{Y}_W^+ - \bar{Y}_W^-}{\bar{D}_W^+ - \bar{D}_W^-}$$

- Local randomization methods: Inference
  - Assuming the observations in the study are the population of interest, not as a random sample from a larger population. The only randomness stems from the random assignment of the treatment
    - **Use the permutation method to obtain p-value and confidence interval (Abadie et al. 2020)**
  - Assuming the superpopulation exist, the observations in the study are seen as a random sample taken from a larger population
    - Use methods based on normal distribution assumption and large-sample approximation

- **Discrete score variable**
  - A score with discrete support implies that multiple units will share the same value of Xi, leading to repeated values, or mass points, in the data
  - With discrete scores, identification and estimation of continuity-based RD treatment effects would necessarily require extrapolation outside the support of the score
  - A key consideration for RD analysis with discrete scores is the number of distinct values M in the support of the running variable
  - Dong (2015) and Barreca et al. (2016) investigate the phenomenon of heaping, which occurs when the score variable is rounded so that units that initially had different score values appear in the data set as having the same value

3. Validation and Falsification

- Analysis of pre-intervention covariates and placebo outcomes
- Density continuity test (McCrary, 2008) to detect endogenous sorting around the cutoff
- Cattaneo et al. (2017) propose a binomial test for counts near the cutoff as an additional manipulation test.
  - Unlike the continuity-based density test, the binomial test can be used when the score is continuous or discrete, and it does not rely on asymptotic approximations

- Placebo Cutoffs and Continuity of Regression Functions
  - Researchers choose a grid of artificial cutoff values and repeat estimation and inference of the RD effect on the outcome of interest at each artificial cutoff value

- Donut Hole
  - Reimplementing estimation and inference for the RD treatment effect with different subsets of observations, as determined either by excluding the observations closest to the cutoff or by varying the bandwidth used for estimation and inference
    - The intuition is that if there is endogenous sorting of units across the cutoff, such sorting might occur only among units whose scores are very close to the cutoff, and thus when those observations are excluded the RD treatment effect may change

- Bandwidth Sensitivity
  - Reestimate the RD treatment effect for bandwidths (or neighborhood lengths) that are smaller or larger than the one originally chosen
  - In the continuity-based framework, if the original bandwidth is MSE optimal, considering much larger bandwidths is not advisable due to the implied misspecification bias
  - In the local randomization framework, considering larger neighborhoods may not be justifiable if important covariates become imbalanced; in this case, the approach will be uninformative

- Future development
  - External validity: To extrapolate RD treatment effects
  - Experimental design and data collection in RD settings
  - Incorporate modern high-dimensional and machine learning methods in RD settings

THANKS!