Gentagnede tværsnit og panel data

Økonometri A

Jakob Egholt Søgaard

Blok 1, 2023

Program

- Gentagende tværsnit (W13.1)
- Difference-in-Difference (W13.2)
  - MLR.4 i DiD analyser
  - Udvidelser af DiD setuppet
- Panel data
  - First differences (W13.3-W13.5, W13.A)
  - Fixed effects (W14.1)
  - Random effects (W14.2)
  - Andre typer panel data (W14.5)

Gentagende tværsnit

Gentagende tværsnit

Vi betragter nu gentagende tværsnit fra flere tidsperioder.

Eksempel ($T = 2$):

- Periode 1, $t = 1$ og en stikprøve på n_1 observationer:

$$(y_{i1}, x_{i11}, x_{i12}, x_{i13}, ..., x_{i1k}) \; i = 1, 2, 3, ..., n_1.$$

- Periode 2, $t = 2$ og en stikprøve på n_2 observationer:

$$(y_{i2}, x_{i21}, x_{i22}, x_{i23}, ..., x_{i2k}) \; i = n_1 + 1, n_1 + 2, ..., n_1 + n_2.$$

Vi følger ikke de samme individer over tid. Derfor tæller i videre fra n_1 i period 2.

Vi antager, at observationerne er en tilfældig stikprøve fra populationen i periode t.

Gentagende tværsnit

Eksempler på gentagende tværsnit:

- **Forbrugerundersøgelsen:** Hvert år udtager DST en nye tilfældig stikprøve fra populationen (ca. 2000 husstande), som registerer deres forbrug.
- **Hussalg:** Vi har kun huspriser på handlede huse. Kun de få huse sælges flere år i træk. Vi kan se på det som en stikprøve af huspriser på handlede huse.
- **Meningsmålinger:** Firmaer som fx Gallup og Epinion spørger løbende en stikprøve om deres opbakning til politiske partier. Typisk ny stikprøve hver gang.

Generelt er det meget billigere at indsample data som gentagnende tværsnit sammenlignet med at følge folk over tid.

Gentagende tværsnit

Vi kan grundlæggende tilgå data på 3 måder:

- **Metode 1: Pool alle observationer.**
  - Estimer $y = X\beta + u$ med alle observationerne: $\hat{\beta}^{pool}$.
  - Modellen indeholder i alt $k + 1$ parametre.
  - Fordel: flere observationer, mere variation.
- **Metode 2: Separate estimationer**
  - Estimer $y = X\beta + u$ kun med observationer fra $t = 1$: $\hat{\beta}^1$
  - Estimer $y = X\beta + u$ kun med observationer fra $t = 2$: $\hat{\beta}^2$
  - Modellen indeholder i alt $2(k + 1)$ parametre.
  - Fordel: kan besvare nye spørgsmål (har β ændret sig over tid?)
- **Metode 3: En mellemting mellem metode 1 og 2.**
  - Pooled estimation, men lad nogle parametrene variere på tværs af tidsperioderne.
  - Antallet af parametre er s med $k + 1 < s < 2(k + 1)$.

Gentagende tværsnit

Metode 3 tillader, at vi formelt kan teste om parameterne har ændret sig over tid. Betragt estimationsligningen

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + u_{it} \tag{1}$$

Hvis vi estimerer denne ligning på det poolede data, antager vi at β'erne er konstante over tid.

Vi kan tillade, at parameterne ændrer sig ved at interagere variablene med dummier (som vi så i W7).

Antag at $T = 2$ og definer en dummy variabel $d2$ som

$$d2_{it} = \begin{cases} 1 \text{ hvis individ } i \text{ er obs. i periode 2} \\ 0 \text{ hvis individ } i \text{ er obs. i periode 1} \end{cases}$$

Gentagende tværsnit

Simpel model, som tillader forskelligt konstantled:

$$y_{it} = \beta_0 + \delta_0 d2_{it} + \beta_1 x_{it1} + \beta_2 x_{it2} + u_{it}$$

"Fuld" model alle interaktionsled tillader ændringer i alle parametrene:

$$\begin{aligned} y_{it} = \beta_0 & & +\delta_0 d2_{it} \\ + \beta_1 x_{it1} & & +\delta_1 x_{it1} d2_{it} \\ + \beta_2 x_{it2} & & +\delta_2 x_{it2} d2_{it} + u_{it} \end{aligned}$$

I denne model svarer $\hat{\beta}$ til estimaterne fra en separat estimation på period 1 og $\hat{\beta} + \hat{\delta}$ svarer til en separat estimation på period 2.

Test for stabile parameter: $H_0 : \delta_j = 0$ for alle j. Teststørrelse: (Robust) F-test.

Gentagende tværsnit: Eksempel 13.2 i Wooldridge

Spørgsmål:

- Har afkastet til uddannelse ændret sig i perioden?
- Er løngabet mellem mænd og kvinder blevet mindre over tid?

Data:

- To uafhængige tværsnit: 1978-CPS og 1985-CPS.
- Indhold: Løn, uddannelse, erfaring for 1.084 ansatte.

Model (13.1):

$$\log(\textit{timeløn}) = \beta_0 + \delta_0 d85 + \beta_1 \textit{educ} + \delta_1 d85 \cdot \textit{educ} + \beta_2 \textit{exper} + \beta_3 \textit{exper}^2 + \beta_4 \textit{union} + \beta_5 \textit{female} + \delta_5 d85 \cdot \textit{female} + u$$

Fokus?

CPS data fra USA + meget andet kan findes på `www.ipums.org`.

Gentagende tværsnit: Eksempel 13.2 i Wooldridge

Result:

$$\begin{aligned}\log(wage) = {} & .459 + .118\, y85 + .0747\, educ + .0185\, y85{\cdot}educ \\ & \quad (.093) \quad (.124) \qquad (.0067) \qquad (.0094) \\ & + .0296\, exper - .00040\, exper^2 + .202\, union \\ & \quad (.0036) \qquad (.00008) \qquad (.030) \\ & - .317\, female + .085\, y85{\cdot}female \\ & \quad (.037) \qquad (.051) \\ n = 1{,}084,\ & R^2 = .426,\ \bar{R}^2 = .422.\end{aligned} \qquad \textbf{[13.2]}$$

Hvad er fortolkningen af parametrene for $y85 \cdot educ$ og $y85 \cdot female$?

Difference-in-Difference

Difference-in-Difference (DiD)

Gentagnede tværsnit gør det muligt at lave såkaldte **Difference-in-Difference (DiD)** analyser.

- En af de mest brugte estimationsmetoder i Økonomi.
- Særligt brugbar til analyser af "politikindgreb" / naturlige eksperimenter.

I sin rene form kræver en DiD analyse to ting af data:

- Data for to grupper, hvoraf den ene er påvirket af et politikindgreb (treatment gruppen) og en gruppe, som ikke er påvirket (control gruppen).
- Data for de to grupper både før og efter politikindgrebet.

Ved at have data for før indgrebet kan vi kontrollere for forskelle mellem grupperne, som ikke skyldes indgrebet.

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

Effekten af forbrændingsanlæg på huspriser

Hypotese: Et forbrændingsanlæg påvirker lokale huspriser negativt.

Naiv analyse:

$$rprice_i = \beta_0 + \beta_1 near_i + u_i, \tag{2}$$

hvor *rprice* er husprisen og *near* er en dummy, som er 1, hvis huset ligger tæt på et forbrændingsanlæg.

Er $\hat{\beta}_1$ et konsistent estimat for effekten af et forbrændingsanlæg? Næppe.

- Det er sjældent tilfældigt, hvor forbrændingsanlæg placeres.
- Ofte placeres de i mindre eftertragtede områder, pga. politisk pres, billigere jordpriser, eksisterende industriområder osv.
- Derfor vil $E(u|near = 1) < 0$ og $\hat{\beta}_1$ være negativt biased.

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

Analyse af effekten af et kommende forbrændingsanlæg i Boston.

- To tværsnit: 1978 og 1981. Data på huspriser og karakteristika for huse med forskellig afstand til et kommende forbrændingsanlæg.
- 1978 er før beslutningen om at opføre anlægget. 1981 er efter beslutningen. Selve anlægget blev først opført i 1985.

En naive estimation af ligning (2) på 1981 data giver

$$\widehat{rprice} = 101,3 - 30,7 \cdot near$$

Hvordan kan vi korrigere for $E(u|near = 1) < 0$?

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

Samme regression på 1978 data:

$$\widehat{rprice} = 82,5 - 18,8 \cdot near$$

Dvs. allerede før forbrændingsanlæg blev vedtaget, kostede huse i området omkring det kommende anlæg mindre end andre huse.

Hvis vi antager at bias i 1981 svarer til bias i 1978, kan vi fjerne bias ved at fratrække $\hat{\beta}_1^{1978}$ fra $\hat{\beta}_1^{1981}$

I så fald ender vi et estimat på

$$\hat{\delta}_1 = -30,7 - (-18,5) = -11,9.$$

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

$\hat{\delta}_1$ er et DiD estimate:

D1 Forskelle mellem huse tæt på og længere væk

D2 Forskel før og efter politikindgreb

Gennemsnitlige huspriser:

	Tæt på	Langt fra	Forskel
1978	63.7	82.5	-18.8
1981	70.6	101.3	-30.7
Forskel	6,9	18.8	-11.9

Differenserne svarer til OLS estimaterne ovenfor.

Ihukom at OLS med dummy variable måler forskelle i gennemsnit.

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

DiD i en regression:

$$rprice_{it} = \beta_0 + \delta_0 d81_{it} + \beta_1 near_{it} + \delta_1 d81_{it} near_{it} + u_{it}, \qquad (3)$$

- δ_0: prisudviklingen fra 1978 til 1981 for kontrol huse ($near_{it} = 0$).
- β_1: prisforskellen før interventionen.
- δ_1: DiD estimat af effekten forbrændingsanlægget.

$E(rprice_{it}|d81_{it}, near_{it})$:

	Tæt på	Langt fra	Forskel
1978	$\beta_0 + \beta_1$	β_0	β_1
1981	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\beta_0 + \delta_0$	$\beta_1 + \delta_1$
Forskel	$\delta_0 + \delta_1$	δ_0	δ_1

Difference-in-Difference (DiD): Eksempel 13.3 i Wooldridge

DiD i en regression med yderligere kontrolvariable:

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 near + \delta_1 y81 \cdot near + \mathbf{X}\boldsymbol{\theta} + u,$$

TABLE 13.2 Effects of Incinerator Location on Housing Prices

Dependent Variable: *rprice*			
Independent Variable	**(1)**	**(2)**	**(3)**
constant	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Other controls	No	*age*, age^2	Full Set
Observations	321	321	321
R-squared	.174	.414	.660

Difference-in-Difference (DiD)

Regression eller gennemsnit (DiD tabel)?

Gennemsnit (DiD tabel):

$$\begin{aligned}\hat{\delta}_1 = &\quad (\bar{y}_{post,treatment} - \bar{y}_{post,control}) - (\bar{y}_{pre,treatment} - \bar{y}_{pre,control}) \\ = &\quad (\bar{y}_{post,treatment} - \bar{y}_{pre,treatment}) - (\bar{y}_{post,control} - \bar{y}_{pre,control})\end{aligned}$$

- Rækkefølgen i differenserne er ligegyldig.
- Meget transparent estimationssetup.

Regression:

$$y = \beta_0 + \delta_0 post + \beta_1 treatment + \delta_1 post \cdot treatment + \mathbf{X}\boldsymbol{\theta} + u,$$

- Kan tilføje ekstra kontrolvariable (mere præcision/fjerne bias).
- Kan let udregne variansen på estimatet.

MLR.4 i DiD analyser

Grundlæggende adskiller DiD analyser ikke fra andre regressionsanalyser.

Men vi er ikke interesseret i alle parameterestimaterne. For konsistent estimat af "behandlingseffekten" (δ_1) kræver vi kun:

$$\underbrace{E(u|post = 1, treatment = 1) - E(u|post = 1, treatment = 0)}_{\text{Bias efter indgreb}} =$$

$$\underbrace{E(u|post = 0, treatment = 1) - E(u|post = 0, treatment = 0)}_{\text{Bias før indgreb}}$$

Dette er en svagere antagelse end $E(u|post, treatment) = 0$. Vi kræver ikke at bias er 0 generelt. Blot at den er konstant over tid.

MLR.4 i DiD analyser

Antagelsen om konstant bias svarer til antagelse om at udviklingen i treatment og kontrolgruppen ville være parallel *i fravær af* treatment.

- I Wooldridge eksemplet svarer det til en antagelse om vi havde observeret samme ændring i priserne for huse tæt på og længere fra forbrændingsanlægget, hvis beslutningen om anlægget aldrig var blevet taget.

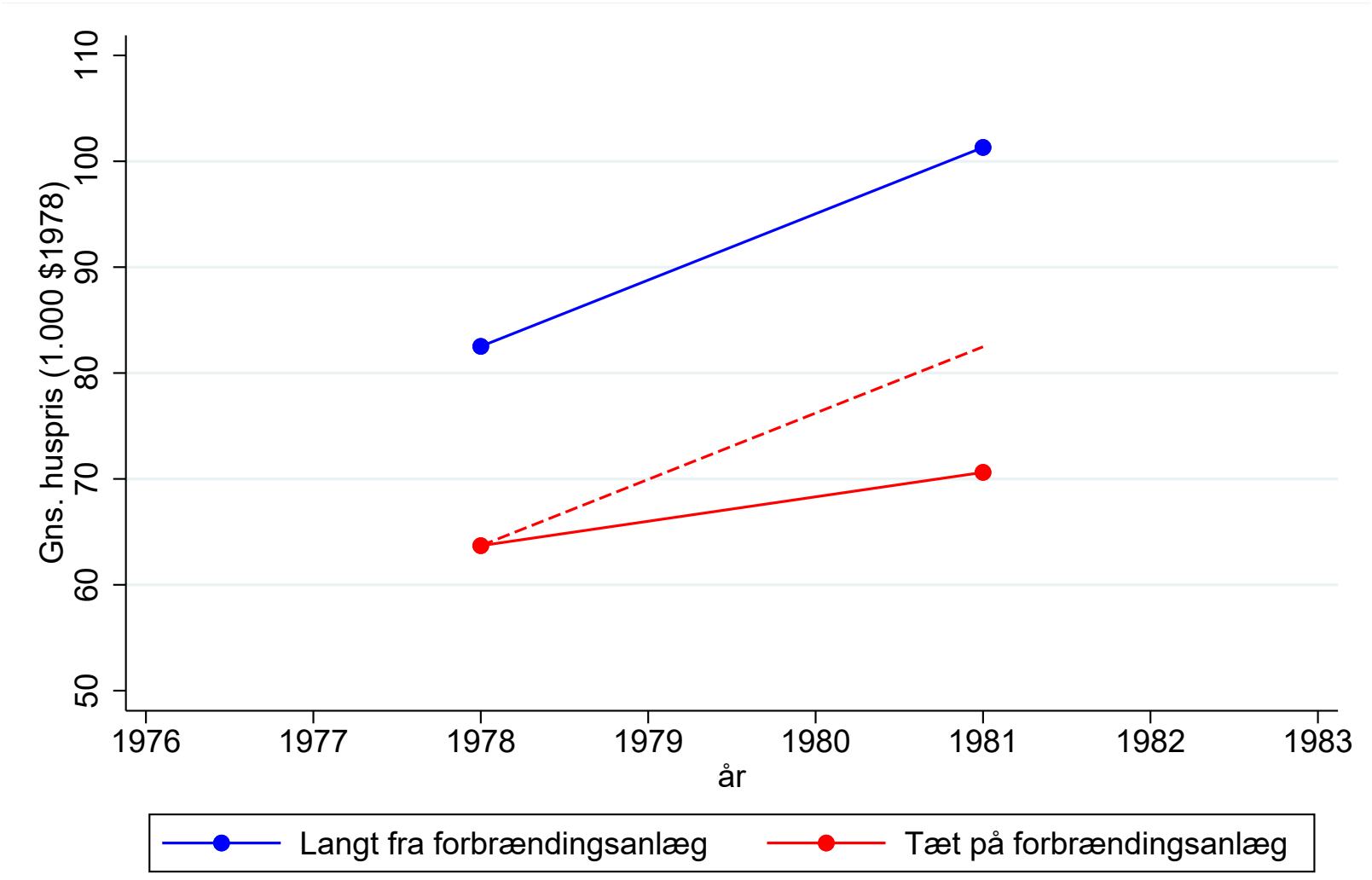
Vi kan ikke teste antagelsen om **parallel trends**

- Det kræver en tidsmaskine at observere udviklingen både med og uden treatment.

Men der findes måder at validere den på.

MLR.4 i DiD analyser: Eksempel 13.3 i Wooldridge

Antagelsen om parallel trends.



DiD med flere tidsperioder

Ovenfor betragtede vi data med to tidsperioder (før og efter et politikindgreb).

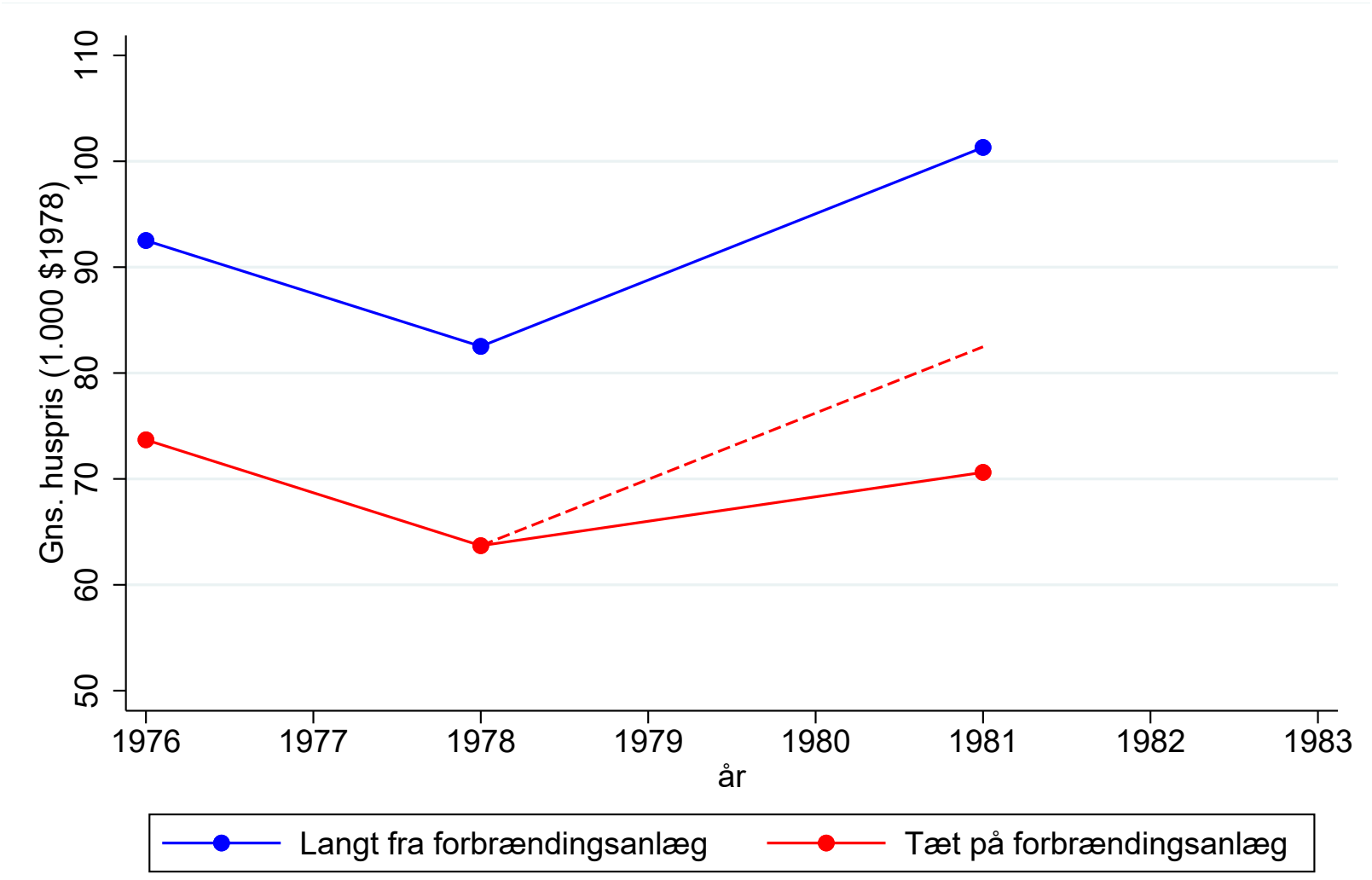
Vi kan nemt udvide DiD analyser med flere perioder.

- Flere pre-perioder: mulighed for at validere antagelsen om parallel trends.
- Flere post-perioder: mulighed for at undersøge om "treatment" effekten ændres over tid.

I dag vil de fleste DiD analyser være utroværdige uden flere pre-perioder.

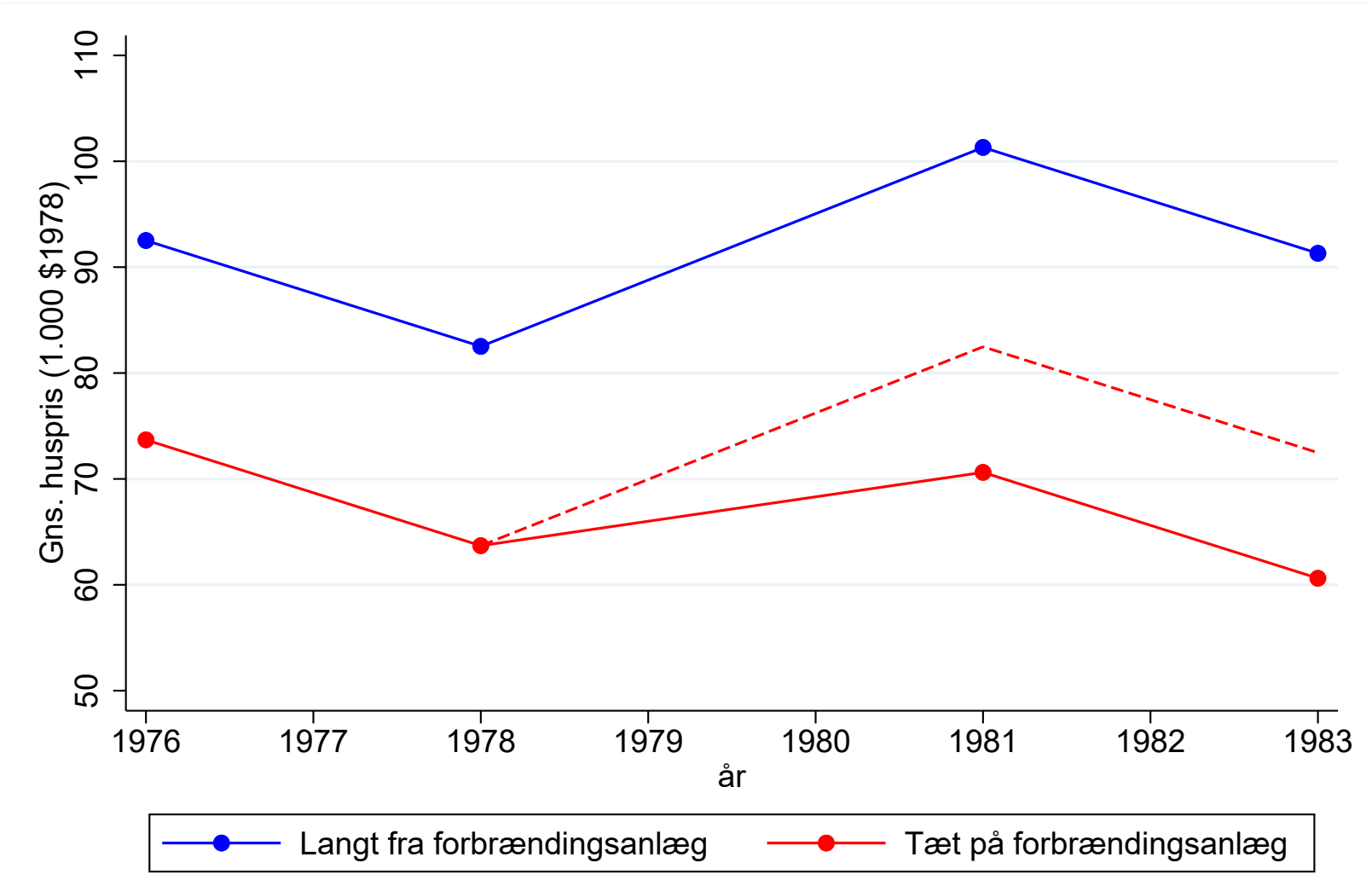
DiD med flere tidsperioder: Eksempel 13.3 i Wooldridge

Flere pre-perioder



DiD med flere tidsperioder: Eksempel 13.3 i Wooldridge

Flere post-perioder



DiD med flere tidsperioder

DiD regression med flere tidsperioder.

For hver ekstra tidsperiod, skal vi bruge en ekstra tidsdummy.

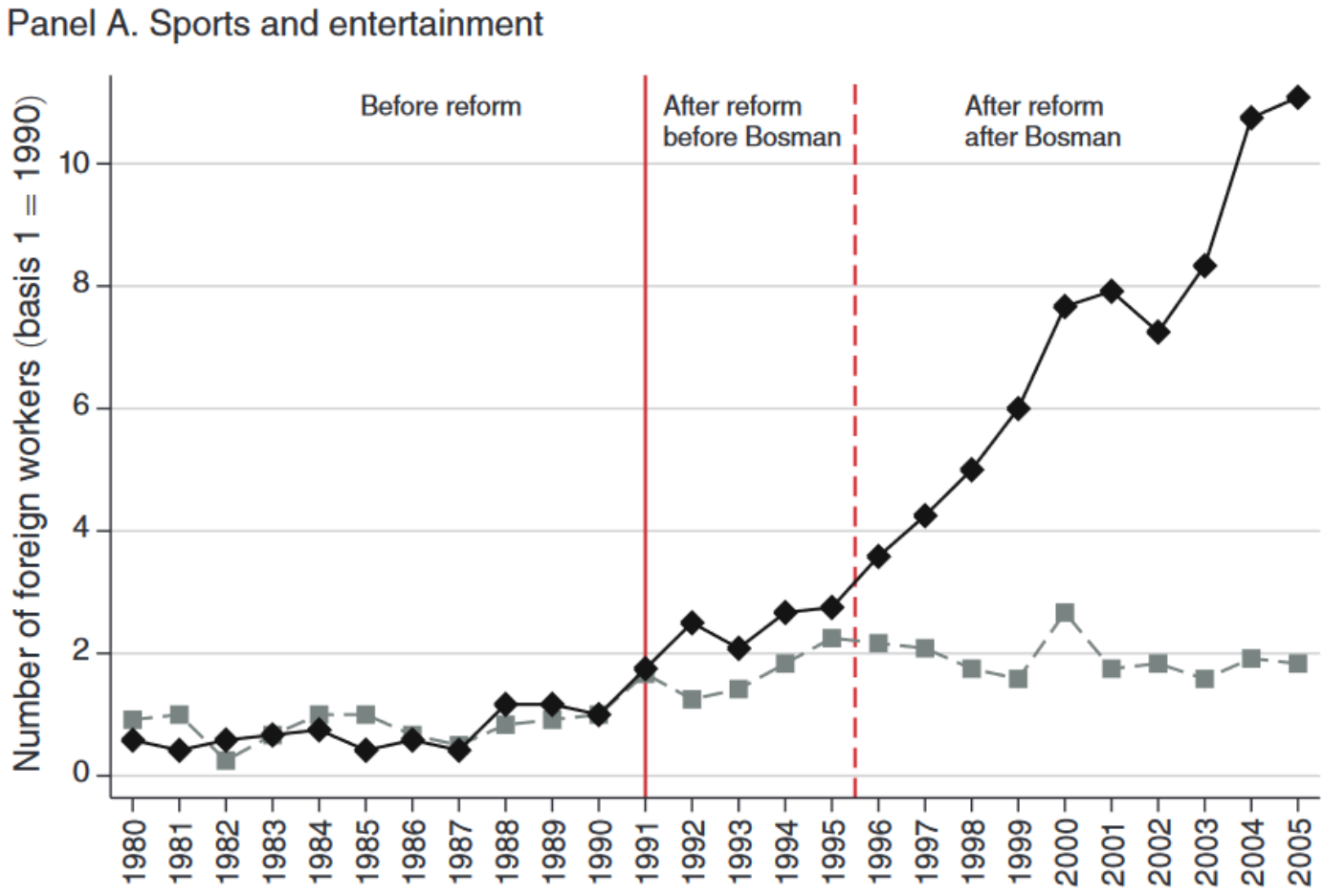
- Typisk årsdummier D_t^{Year} eller tid relevant til treatment D_t^{Event}
- Vi er dog nød til at undlade én dummy (dummy fælden).

Regressionsmodel

$$y_{it} = \beta_0 + \beta_1 treat_i + \sum_t \delta_t D_t + \sum_t \gamma_t D_t \cdot treat_i + \mathbf{X}\boldsymbol{\theta} + u,$$

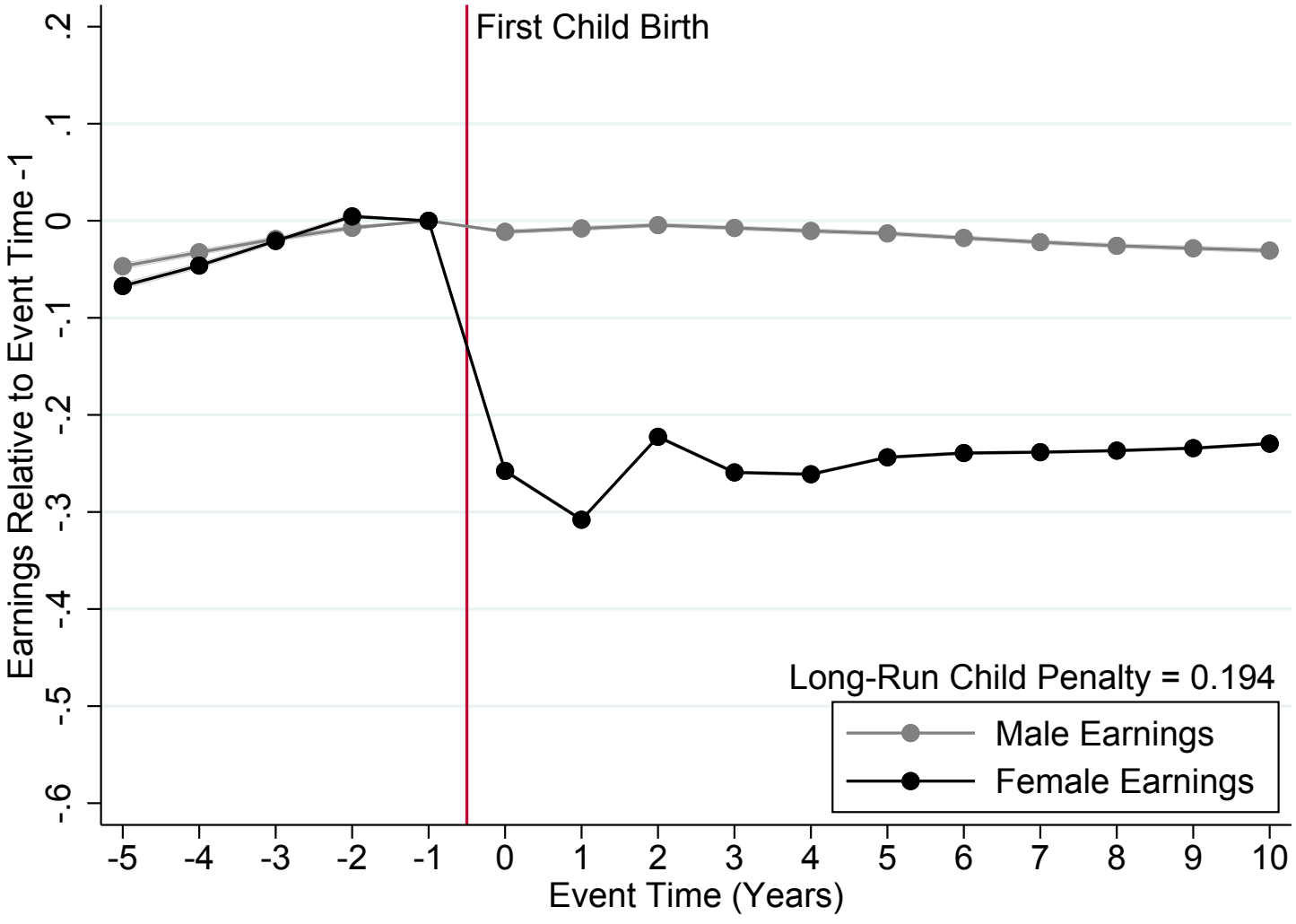
- Referencegruppen er her året før treatment.
- γ'erne er de dynamiske DiD estimater.
- H_0 for parallelle pre-trends?

Effekten af beskatning på international mobilitet



Kleven, Landais & Saez (2013). American Economic Review, vol. 103(5)

DiD med flere tidsperioder: Effekten af børn forældres løn



Kleven, Landais & Søgaard (2019), American Economic Journal: Applied Economics, vol. 11(4)

DiD med flere kontrolgrupper

Vi kan også tilføje en ekstra kontrolgruppe til DiD setuppet. Ideen:

- Find tilsvarende data og foretage en tilsvarende opdeling af population, fx i et område, som ikke er påvirket af politikindgrebet.
- Foretag en DiD på dette data. Hvis antagelsen om parallel trends holder, burde dette DiD estimat være insignifikant.
- Hvis vi er bekymret for om antagelsen om parallel trends holder, kan vi trække det nye DiD estimat fra det oprindelige.
- Derved får vi et DiDiD estimat.
- Ny antagelse: Konstante forskelle i trends.

Som med DiD kan vi opstille DiDiD, som en regressionsanalyse.

Dog bliver regressionsligningen ret kluntet (Se ligning 13.14).

Et generelt DiD setup

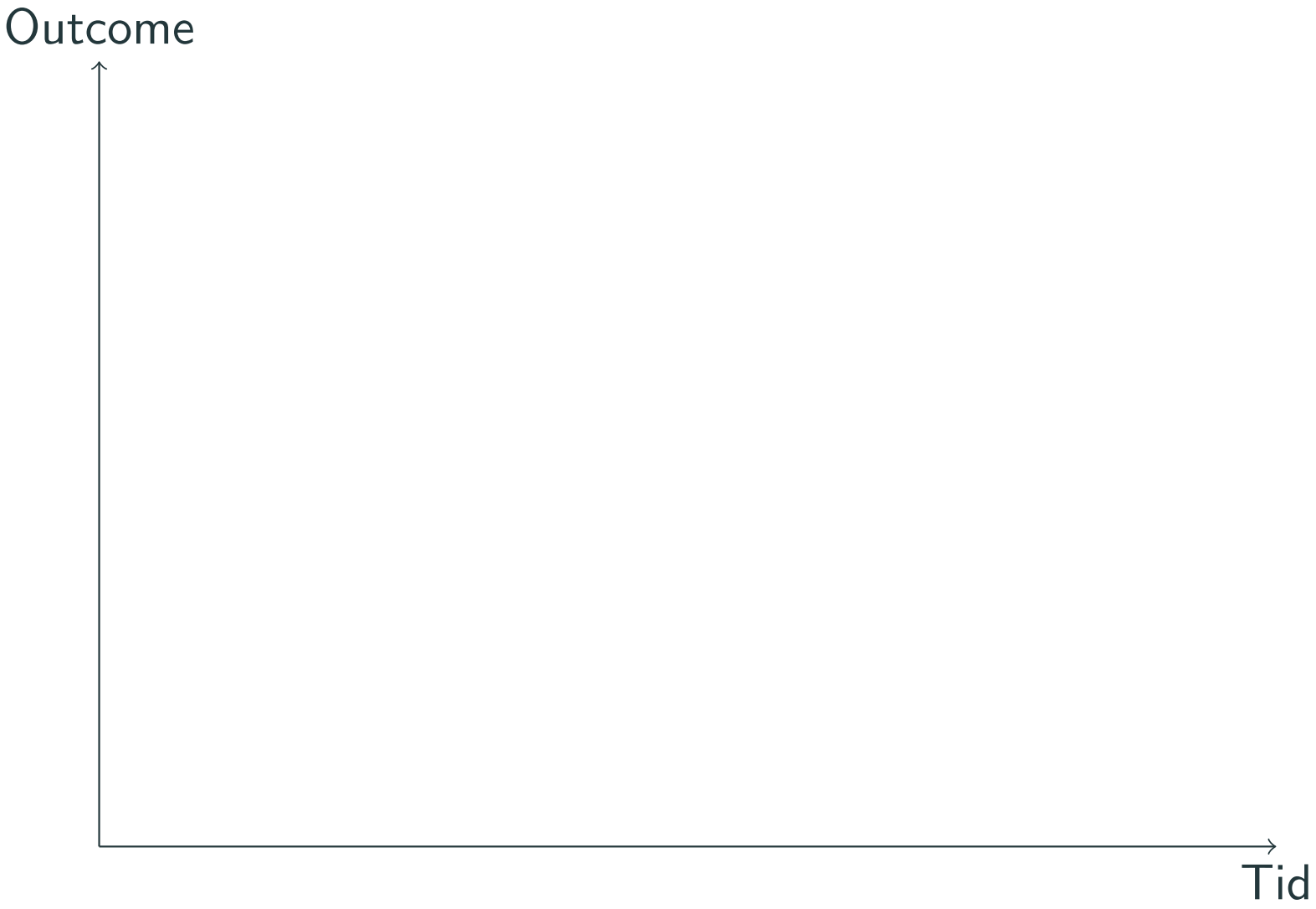
Wooldridge opstiller et generelt setup til DiD analyser (W13.2b), som

$$y_{igt} = \boldsymbol{\lambda}_t + \boldsymbol{\alpha}_g + \beta x_{igt} + \mathbf{Z}\boldsymbol{\gamma} + u_{igt},$$

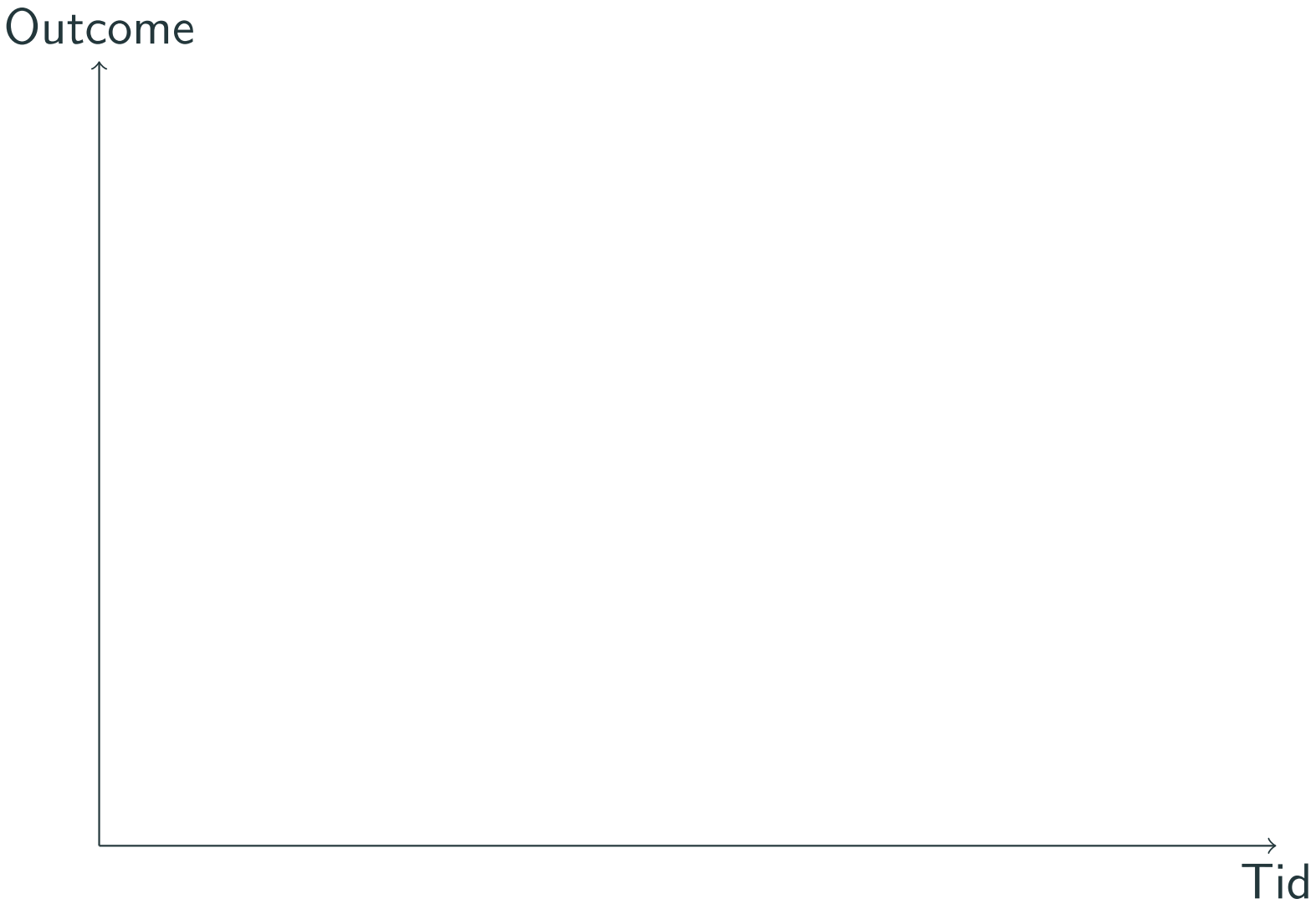
hvor $\boldsymbol{\lambda}_t$ er en vektor af tidsdummier og $\boldsymbol{\alpha}_g$ er gruppedummier.

- x_{igt} er et mål for det politikindgreb vi er interessede i, fx mindstelønnen, en skattesats eller lignende.
- Fordelen er at vi i et setup kan bruge alle politikændringer (stigninger, fald mv.)
- I Wooldridge eksempel 13.3 er $x_{igt} = d81_{it} near_{it}$, dvs. en dummy for om der er planlagt et forbrændingsanlæg.

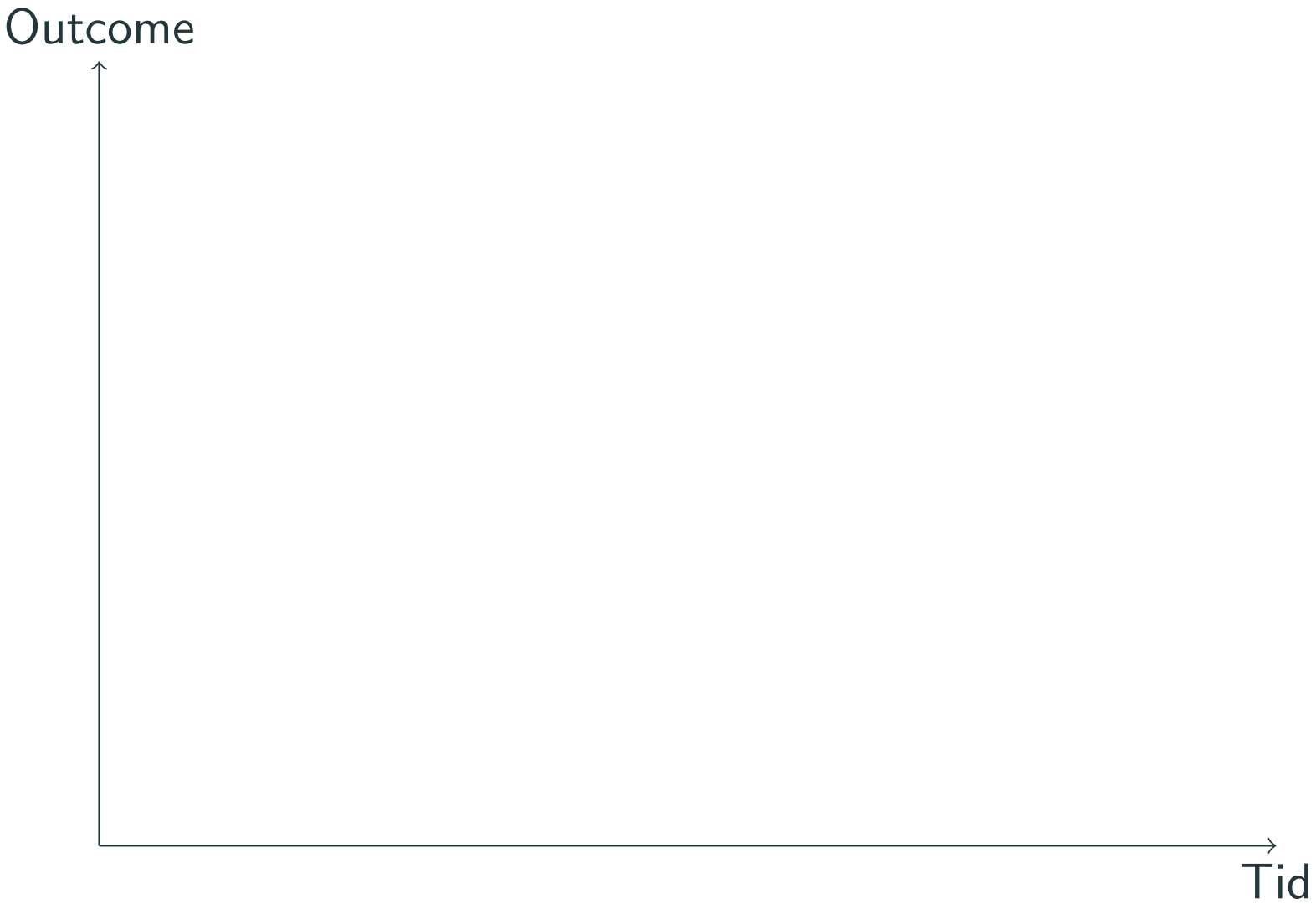
Et generelt DiD setup: Identification 1



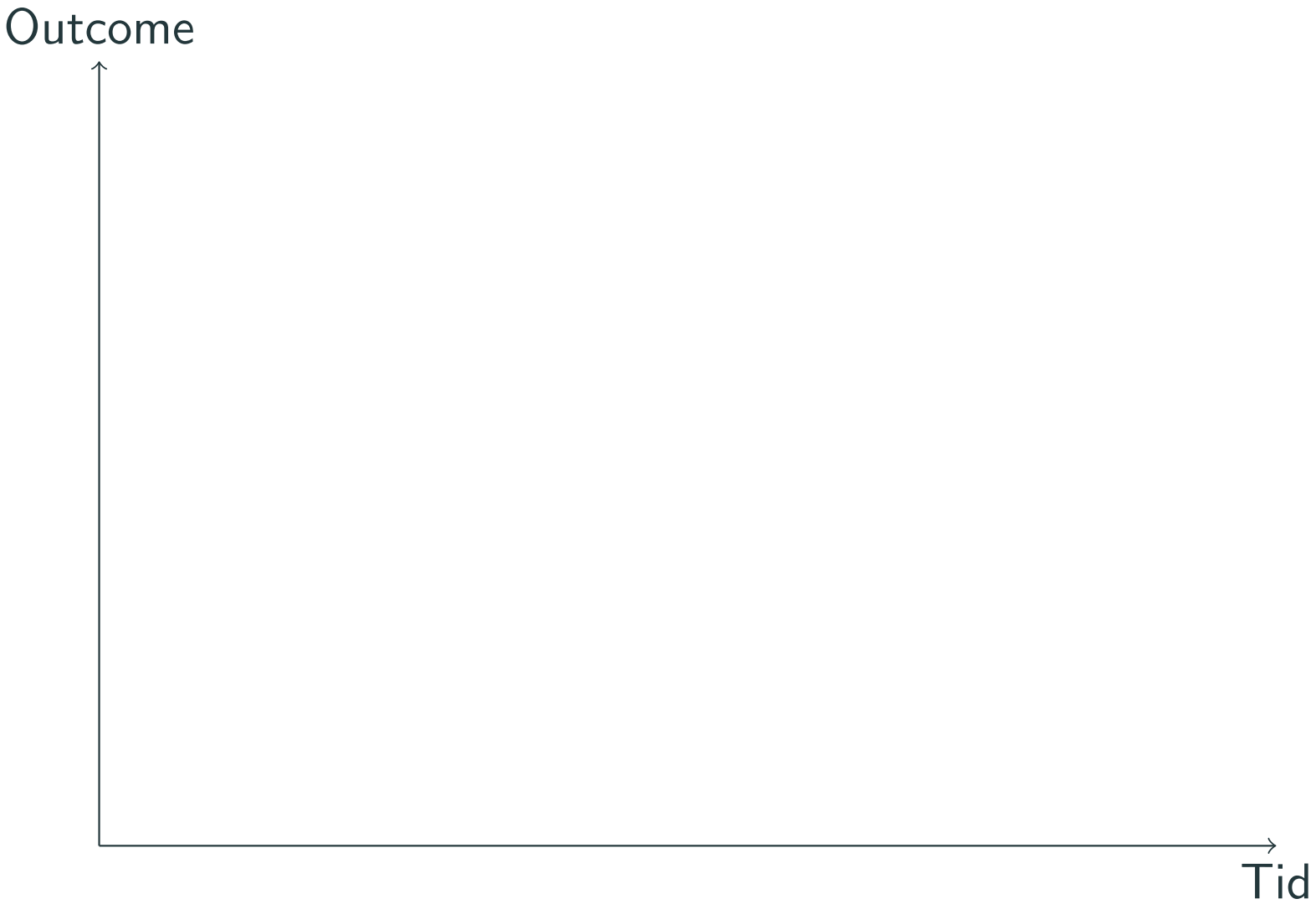
Et generelt DiD setup: Identification 2



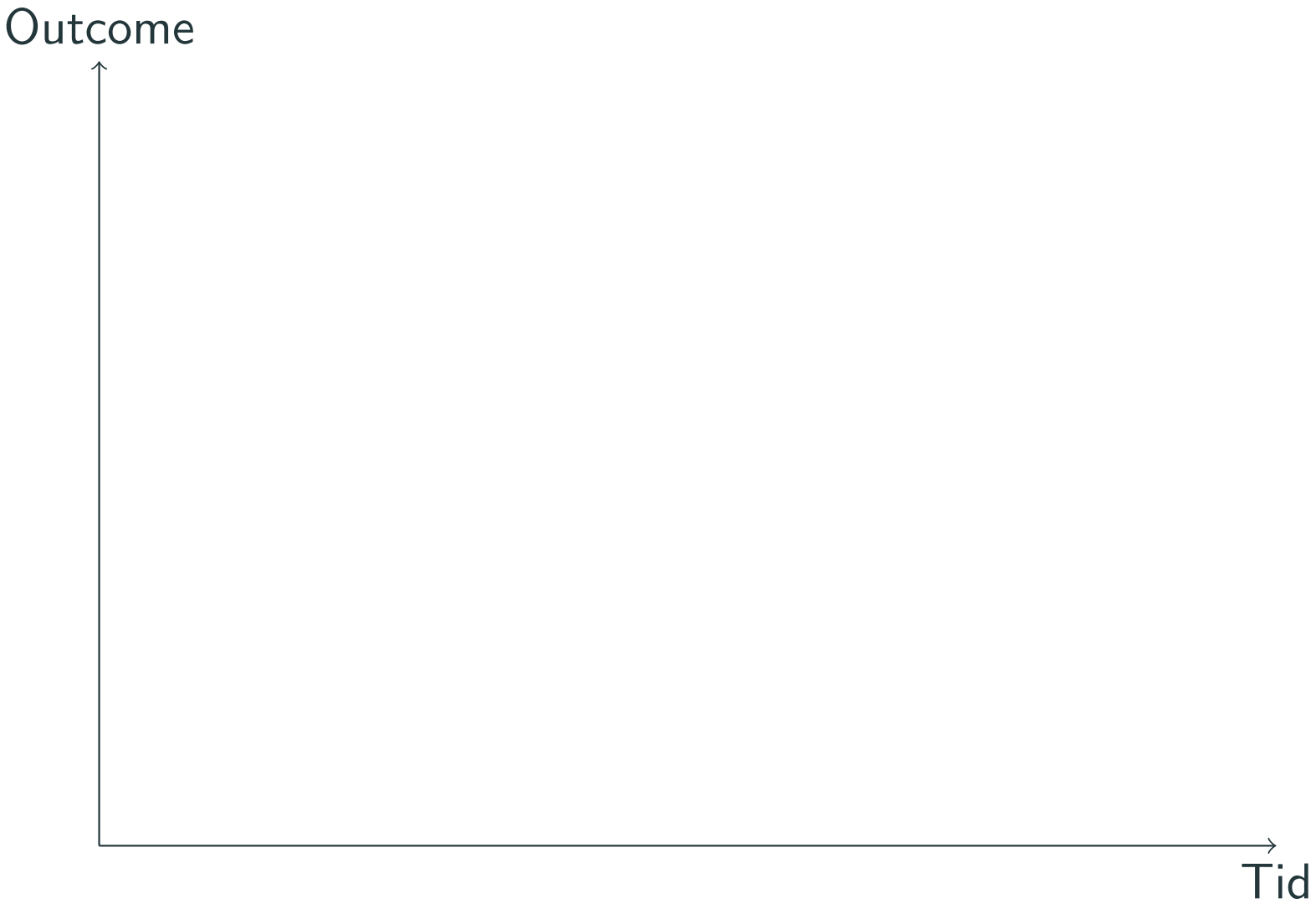
Et generelt DiD setup: Identification 3



Et generelt DiD setup: Identification 4



Et generelt DiD setup: Identification 5



Et generelt DiD setup: Opsamling

Generaliseret DiD estimationsligning

$$y_{igt} = \boldsymbol{\lambda}_t + \boldsymbol{\alpha}_g + \beta x_{igt} + \mathbf{Z}\boldsymbol{\gamma} + u_{igt},$$

Antagelser:

- α_g er konstant over tid
- δ_t er fælles for treatment og kontrolgrupper
- β er konstant over tid

Vi kan validere antagelerne ved at inkludere leads og lags af x:

- Lags ($x_{igt-1}, x_{igt-2}, ...$) fanger ændring i β over tid.
- Leads ($x_{igt+1}, x_{igt+2}, ...$) fanger forskelle i trends op til treatment

Et generelt DiD setup: Quiz

Betragt ligningen estimationsmodellen:

$$y_{igt} = \boldsymbol{\lambda}_t + \boldsymbol{\alpha}_g + \beta x_{igt} + \mathbf{Z}\boldsymbol{\gamma} + u_{igt},$$

Kan vi estimere β i følgende tilfælde?

1. x_{igt} varierer på tværs af grupper, men for hver gruppe er x_{igt} fast over tid.

2. x_{igt} varierer på tværs over tid, men i hver period er x_{igt} den samme for hver gruppe.

3. x_{igt} varierer både over over tid og på tværs af gruppe.

Et generelt DiD setup: Forskningseksempel

Universal Investment in Infants and Long-Run Health: Evidence from Denmark's 1937 Home Visiting Program†

By Jonas Hjort, Mikkel Sølvsten, and Miriam Wüst*

This paper examines the long-run health effects of a universal infant health intervention, the 1937 Danish home visiting program, which targeted all infants. Using administrative population data and exploiting variation in the timing of implementation across municipalities, we find that treated individuals enjoy higher age-specific survival rates during middle age (45–64), experience fewer hospital nights, and are less likely to be diagnosed with cardiovascular disease. These results suggest that an improved nutrition and disease environment in infancy "programmed" individuals for lower predisposition to serious adult diseases. (*JEL* H51, I12, I18, J13, N34)

Hjort, Sølvsten, & Wüst. (2017). American Economic Journal: Applied Economics, 9(4)

Et generelt DiD setup: Forskningseksempel

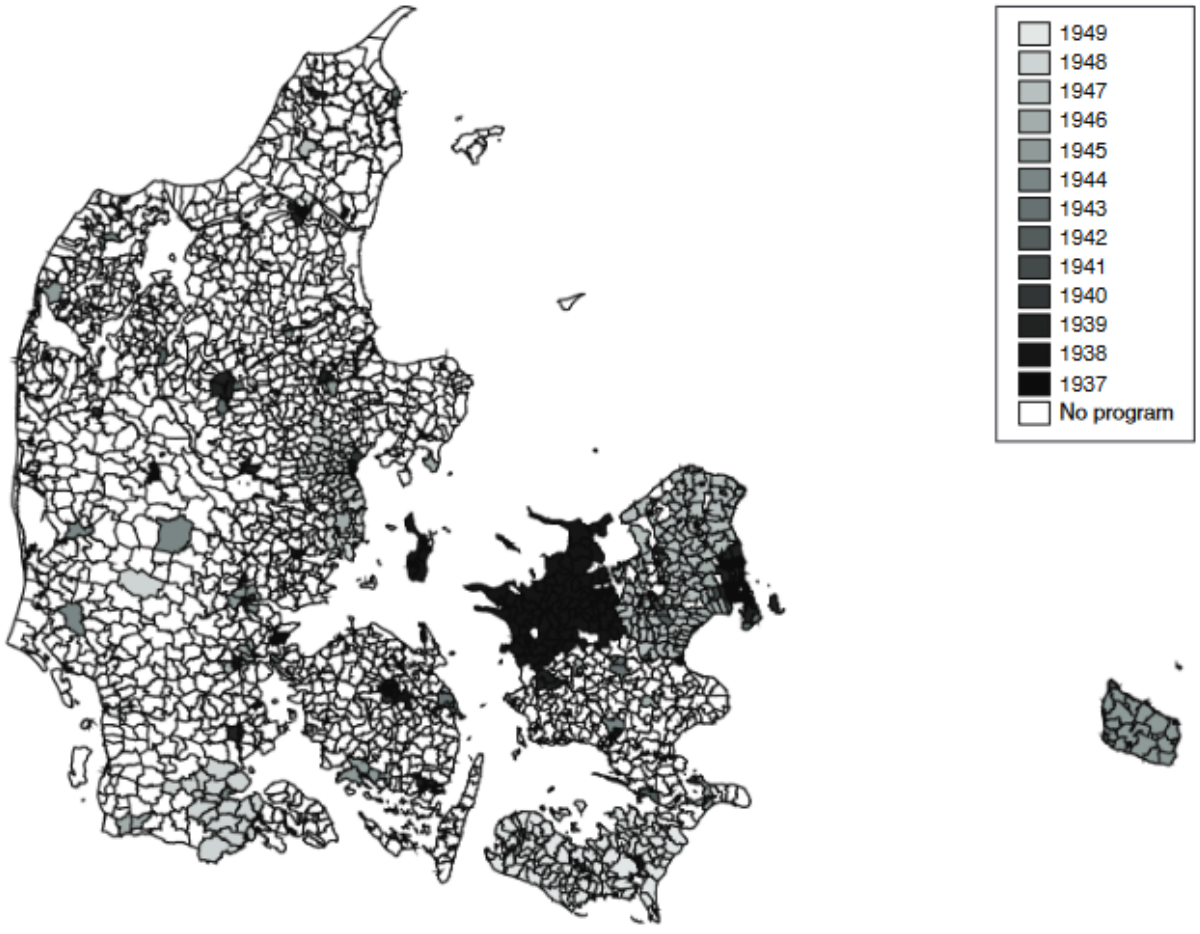


FIGURE 1. MUNICIPALITIES AND THEIR DATE OF ENTRY INTO TREATMENT, 1937–1949

Hjort, Sølvsten, & Wüst. (2017). American Economic Journal: Applied Economics, 9(4)

Et generelt DiD setup: Forskningseksempel

II. Empirical Strategy

To estimate the effect of the home visiting program on long-run health, we follow a difference-in-difference approach (DiD), beginning with the following baseline specification:

$$y_{jt} = \alpha + \beta\, homevisit_{jt} + \gamma_j + \delta_t + \epsilon_{jt}, \tag{1}$$

Hjort, Sølvsten, & Wüst. (2017). American Economic Journal: Applied Economics, 9(4)

Et generelt DiD setup: Forskningseksempel

TABLE 2—EFFECT OF THE HOME VISITING PROGRAM ON SURVIVAL BEYOND VARIOUS AGES, COHORTS 1935–1949

Outcome	All (1)	All (2)	All (3)	Matched (4)	Ever impl. (5)
Survival until age 50	0.064	0.087	0.165	0.063	0.198
	(0.063)	(0.078)	(0.081)	(0.171)	(0.096)
Mean of dependent variable × 100	98.342	98.342	98.342	98.377	98.218
Observations	20,078	20,078	20,078	6,204	5,769
Survival until age 55	0.228	0.264	0.312	0.317	0.335
	(0.091)	(0.119)	(0.120)	(0.273)	(0.151)
Mean of dependent variable × 100	95.880	95.880	95.880	95.925	95.610
Observations	20,078	20,078	20,078	6,204	5,769
Survival until age 60	0.337	0.260	0.469	0.587	0.469
	(0.123)	(0.140)	(0.167)	(0.372)	(0.207)
Mean of dependent variable × 100	92.432	92.432	92.432	92.535	91.967
Observations	20,078	20,078	20,078	6,204	5,769
Survival until age 64	0.382	0.257	0.454	0.592	0.387
	(0.155)	(0.178)	(0.195)	(0.435)	(0.235)
Mean of dependent variable × 100	88.756	88.756	88.756	88.791	88.121
Observations	20,078	20,078	20,078	6,204	5,769
Cohort fixed effects	Yes	Yes	Yes	Yes	Yes
Municipal					
Fixed effects	Yes	Yes	Yes	Yes	Yes
X (level) × year interactions	No	No	Yes	No	Yes
X (trend) × year interactions	No	No	Yes	No	Yes
Linear time trends	No	Yes	No	No	No

Panel data

Panel data

Med **Panel data** observerer vi - modsat gentagende tværsnit - hver enhed flere gange.

- Fx lande, kommuner, virksomheder og individer.

Antag at vi observer hver enhed i til hvert t (**balanceret panel**):

- Periode 1, $t = 1$ og en stikprøve på n observationer:

$$(y_{i1}, x_{i11}, x_{i12}, x_{i13}, ..., x_{i1k}) \; i = 1, 2, 3, ..., n.$$

- Periode 2, $t = 2$ og en stikprøve på n observationer:

$$(y_{i2}, x_{i21}, x_{i22}, x_{i23}, ..., x_{i2k}) \; i = 1, 2, 3, ..., n.$$

Vi har i alt $2n$ observationer.

OBS: Der er ofte større problemer med manglende data ved panel data pga. sample attrition.

Panel data

Med panel data kan vi estimere modeller med uobserveret heterogenitet, som fx

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + a_i + u_{it}$$

hvor fejlleddet nu består af to led: a_i og u_{it}

- a_i er uobserveret heterogenitet/individuel effekt:
  - Tidinvariant og individspecifikt.
- u_{it} er et idiosynkratisk fejlled:
  - Varierer tilfældigt på tværs af individer og tidsperioder.

Panel data

Estimation af:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + a_i + u_{it}, \tag{4}$$

OLS estimatoren er konsistent og middelret, hvis MLR.4 er opfyldt for både a_i og u_{it}. Dvs.

$$E(v_{it}|x) = E(a_i + u_{it}|x) = E(a_i|x) + E(u_{it}|x) = 0$$

I dette tilfælde kan vi glemme panelstrukturen, og benytte alle observationerne i en **pooled OLS**.

Panel data

Ofte er formålet med panel data dog netop at tillade for uobserveret heterogenitet $E(a_i|x) \neq 0$.

Hvordan gør det vi?

Mulighed 1: Kan vi ikke bare estimere a_i?

- Bemærk at model (4) grundlæggende svarer til den generelle DiD model med individet som "gruppe" ($a_i = \alpha_g$).
- Kan vi så ikke bare estimere modellen på samme måde ved at inkludere en dummy for hvert individ?

Problem:

Panel data

Ofte er formålet med panel data dog netop at tillade for uobserveret heterogenitet $E(a_i|x) \neq 0$.

Hvordan gør det vi?

Mulighed 2: Vi kan omskrive modellen:

- FD: Første differenser (First differences).
- FE: Faste effekter (Fixed effects).

Første differenser (FD)

Vi kan omskrive ligning (4):

- Periode 1:

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1}$$

- Periode 2:

$$y_{i2} = \beta_0 + \delta_0 + \beta_1 x_{i2} + a_i + u_{i2}$$

- Første-differenser:

$$y_{i2} - y_{i1} = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + a_i - a_i + u_{i2} - u_{i1}$$
$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i.$$

- Den uobserverede heterogenitet a_i forsvinder.

Vi har nu et tværsnitsdatasæt med første-differenser.

FD-estimatoren

"Standard" OLS estimation af

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i.$$

Giver FD-estimatoren:

$$\hat{\beta}_1^{FD} = \frac{cov(\Delta x_i, \Delta y_i)}{var(\Delta x_i)} = \frac{\sum_{i=1}^{n} (\Delta x_i - \overline{\Delta x_2}) \Delta y_i}{\sum_{i=1}^{n} (\Delta x_i - \overline{\Delta x_2}) \Delta x_i}$$

FD-estimatoren

FD-estimatoren er konsistent hvis FD.1-FD4 er opfyldt

- FD.1: Modellen er lineær: $y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + a_i + u_{it}$
- FD.2: Tilfældige stikprøve af individer
- FD.3: Variation i variablene over tid.
- FD.4: $E(u_t|x_1, x_2, d2_1, d2_2, a) = 0$

FD.2 og FD.4 $\Rightarrow$ Streng eksogenitet: $cov(u_{it}, x_{is}) = 0$ for alle t og s

FD-estimatoren er konsistent: Bevis

FD-estimatoren: Uobserveret heterogenitet

Bemærk at vi ikke antager noget om $cov(a_i, x_{it})$.

- Vi kan tillade, at den individspecifikke effekt a_i kan være korreleret med de forklarende variable: $cov(a_i, x_{it}) \neq 0$.

Hvis $cov(a_i, x_{it}) \neq 0$ ville **Pooled OLS** vil være biased og inkonsistent.

Hvorfor det?

FD-estimatoren og DiD: Eksempel

Spørgsmål: Reducerer politiovervågning antallet af biltyverier

Setup: Terrorangreb mod en jødisk institution i Buenos Aires, Argentina den 18. juli 1994. Politiet opstiller derefter bejente ved jødiske institutioner.[1]

Data:

- Gennemsnitlige antal biltyverier per boligblok i Buenos Aires april-juni (før angrebet) og august-oktober (efter angrebet).
- Information om hvorvidt en boligblok indeholder beboelse, forretninger (butikker, banker, servicestationer mv.) og institutioner, herunder religiøse (kirker, synagoger mv.).

[1] Di Tella and Schargrodsky. 2004. "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack." American Economic Review, 94 (1): 115-133.

FD-estimatoren og DiD: Eksempel

Difference-in-Difference model:

$$biltyverier_{it} = \beta_0 + \beta_1 institution_i + \beta_2 Defter_t + \beta_3 (institution_i \cdot Defter_t) + \alpha_i + u_{it}$$

Hvordan ser First Difference modellen ud?

$$A : \Delta biltyverier_i = \beta_0 + \beta_1 Institution_i + \Delta u_i$$
$$B : \Delta biltyverier_i = \beta_0 + \beta_3 Institution_i \cdot Defter_t + \Delta u_i$$
$$C : \Delta biltyverier_i = \beta_2 + \beta_3 Institution_i + \Delta u_i$$

FD-estimatoren og DiD: Eksempel

Difference-in-Difference tabel:

Gennemsnitligt antal biltyverier per blok pr måned:[2]

	Institution	Ej institution	Diff
apr-jun	0.113	0.096	0.017
aug-okt	0.041	0.114	-0.073
Diff	-0.072	0.018	-0.090

DiD Regressionsresult: -0.090 (0.037)

FD Regressionsresult: -0.090 (0.033)

[2] Egne beregninger pga. Di Tella and Schargrodsky data (se Absalon)

FD-estimatoren: Quiz

"Sand" model:

$$\log(\mathit{timeløn}_{it}) = \beta_0 + \delta_0 d2_t + \beta_1 \mathit{udd}_{it} + \beta_2 \mathit{erfaring}_{it} + \beta_3 \mathit{lokalledighed}_{it} + \beta_4 \mathit{kvinde}_{it} + a_i + u_{it}.$$

FD estimationsmodel:

$$\Delta \log(\mathit{timeløn}_{i2}) = \delta_0 + \beta_1 \Delta \mathit{udd}_{i2} + \beta_2 \Delta \mathit{erfaring}_{i2} + \beta_3 \Delta \mathit{lokalledighed}_{i2} + \beta_4 \Delta \mathit{kvinde}_{i2} + \Delta u_{i2}.$$

Vil være muligt at estimere:

- β_4?
- β_1?

FD-estimatoren med flere end 2 perioder

Det er ligefrem at udvide FD-estimatoren til mere end 2 perioder:

$$y_{it} - y_{it-1} = \delta_0 + \beta_1(x_{it} - x_{it-1}) + a_i - a_i + u_{it} - u_{it-1}$$
$$\Delta y_{it} = \delta_0 + \beta_1 \Delta x_{it} + \Delta u_{it}.$$

Med $T > 2$ har vi et panel af første-differenser:

- Ændringer fra $t = 1$ til $t = 2$.
- Ændringer fra $t = 2$ til $t = 3$, osv.

FD-estimatoren vil stadig værre konsistent under FD.1-FD.4.

- Men de normale standardfejl vil kun være valide hvis Δu_{it} er ukorrelerede over tid.
- Mere om det nedenfor.

FE-estimatoren

Ved FD estimatoren fjernede vi a_i vha. første differenser.

Vi kan også fjerne a_i vha. en såkaldt **within transformation**.

Betragt igen modellen:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \delta_0 d2_t + a_i + u_{it}, \quad i = 1, 2, .., n, \quad t = 1, 2, .., T.$$

Beregn individgennemsnit

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \delta_0 \overline{d2} + a_i + \bar{u}_i, \quad i = 1, 2, .., n, \tag{5}$$

hvor $\bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it}$.

Ligning (5) indeholder kun variation mellem individer **(between variation)**.

FE-estimatoren

Within transformation:

$$y_{it} - \bar{y}_i = \beta_0 - \beta_0 + \beta_1(x_{it} - \bar{x}_i) + \delta_0(d2_t - \overline{d2}) + a_i - a_i + (u_{it} - \bar{u}_i)$$
$$\Leftrightarrow \ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \delta_0(\ddot{d2}_t) + \ddot{u}_{it} \quad i = 1, .., n, \quad t = 1, .., T. \tag{6}$$

Denne ligning indeholder kun variation indenfor individer **(within variation)**.

Vi har nu et paneldatasæt med afvigelser fra individgennemsnit.

FE-estimatoren

"Standard" OLS estimation af

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \delta_0 \ddot{d2}_t + \ddot{u}_{it}$$

Giver FE-estimatoren:

$$\hat{\beta}_1^{FE} = \frac{cov(\ddot{x}_{it}, \ddot{y}_{it})}{var(\ddot{x}_{it})} = \frac{\sum_{i=1}^{n} (\ddot{x}_{it} - \bar{\ddot{x}}_2)\ddot{y}_{it}}{\sum_{i=1}^{n} (\ddot{x}_{it} - \bar{\ddot{x}}_2)\ddot{x}_{it}} = \frac{\sum_{i=1}^{n} \ddot{x}_{it}\ddot{y}_{it}}{\sum_{i=1}^{n} \ddot{x}_{it}^2}$$

FE-estimatoren er middelret under antagelse FE.1-FE.4 (se W14.A), som grundlæggende svarer til FD.1-FD.4.

- Særligt antager vi at x er strengt eksogen
- Dvs. $E(u_{it}|x_{is}, a_i) = 0$ for alle t og s

FE eller FD?

For $T = 2$ er FE og FD estimatoren identiske. Det ses ved at

$$\ddot{x}_{i1} = x_{i1} - \bar{x}_i = x_{i1} - \tfrac{1}{2}(x_{i2} + x_{i1}) = \tfrac{1}{2}(x_{i1} - x_{i2}) = \quad -\tfrac{1}{2}\Delta x_{i2}$$
$$\ddot{x}_{i2} = x_{i2} - \bar{x}_i = x_{i2} - \tfrac{1}{2}(x_{i2} + x_{i1}) = \tfrac{1}{2}(x_{i2} - x_{i1}) = \quad \tfrac{1}{2}\Delta x_{i2}$$

Dvs. ved $T = 2$ svarer hver observation i FE til en halvdelen af en observation i FD.

Tilgengæld har vi dobbelt så mange observation med FE end med FD.

$$\hat{\beta}_1^{FE} = \frac{cov(\ddot{x}_{it}, \ddot{y}_{it})}{var(\ddot{x}_{it})} = \frac{cov(\frac{1}{2}\Delta x_{i2}, \frac{1}{2}\Delta y_{i2})}{var(\frac{1}{2}\Delta x_{i2})} = \hat{\beta}_1^{FD}$$

Det gælder ikke ved $T > 2$.

FE eller FD?

For $T > 2$ adskiller FE og FD sig særligt i forhold til, hvornår fejlleddene er ukorrelerede - dvs. hvornår FD.6/FE.6 er opfyldt.

- FD.6: $cov(\Delta u_{it}, \Delta u_{is}|X) = 0$
- FE.6: $cov(\ddot{u}_{it}, \ddot{u}_{is}|X) = 0 \Rightarrow cov(u_{it}, u_{is}|X) = 0$

for alle $t \neq s$.

- FE.6 er opfyldt, hvis u_{it} er ukorreleret over tid.
- FD.6 er opfyldt, hvis Δu_{it} er ukorreleret over tid (u_{it} er en random walk).

FE eller FD?

For $T > 2$ adskiller FE og FD sig særligt i forhold til, hvornår fejlleddene er ukorrelerede - dvs. hvornår FD.6/FE.6 er opfyldt.

- FD.6: $cov(\Delta u_{it}, \Delta u_{is}|X) = 0$
- FE.6: $cov(\ddot{u}_{it}, \ddot{u}_{is}|X) = 0 \Rightarrow cov(u_{it}, u_{is}|X) = 0$

for alle $t \neq s$.

- FE.6 er opfyldt, hvis u_{it} er ukorreleret over tid.
- FD.6 er opfyldt, hvis Δu_{it} er ukorreleret over tid (u_{it} er en random walk).

Afhængig af korrelationen af u_{it} over tid vil enten FD eller FE være mest efficient (mindste standardfejl).

Generelt vil vi gerne tillade for at u_{it} kan være korreleret over tid vha. clustered standardfejl (ud over pensum i dette fag).

RE-estimatoren

Hvis $E(v_{it}|x) = E(u_{it} + a_i|x) = 0$ vil pooled OLS, FD og FE være middelrette og konsistente. Ingen af dem er dog den mest efficiente estimator:

- FD og FE smider al between variation væk.
- Omvendt ignorer pooled OLS at v_{it} er korreleret over tid og lægger for meget vægt på between variationen.

Selvom u_{it} er ukorreleret over tid, gælder der for det sammensatte fejlled v_{it}:

$$corr(v_{it}, v_{is}) = \frac{cov(u_{it} + a_i, u_{is} + a_i)}{(var(u_{it} + a_i)var(u_{is} + a_i))^{1/2}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}$$

Korrelationen mellem v_{it} betyder også at OLS standardfejlene er forkerte.

RE-estimatoren

Man kan få en efficient estimator ved kun at smide en del af between variationen væk **Partial Demeaning)**:

$$y_{it} - \theta \bar{y}_i = \beta_0 + \beta_1 (x_{it} - \theta \bar{x}_i) + v_{it} - \theta \bar{v}_i$$

hvor

$$\theta = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2} \right)^{1/2}$$

Dette er RE estimatoren.

RE estimatoren er generelt ikke særlig brugt i økonomi.

- Bygger på samme (strenge) antagelser om pooled OLS.
- Vi vil typisk tillade for en mere generel korrelation mellem fejlleddene.

Sammenligning af estimatorer: Simulering

Den datagenererende proces

$$\beta_0 = 1,\ \beta_1 = 2$$
$$a_i \sim iiN(0,4),\ v_{it} \sim iiN(0,4)\ e_{it} \sim iiN(0,1),$$
$$x_{it} = 2 + v_{it} + a_i,$$
$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + e_{it}$$

Vi simulerer med $T = 2$ og $n = 500$ og beregner OLS, FD, FE og RE.

Quiz: Hvad er fortegnet på asymptotiske bias af OLS estimatoren?

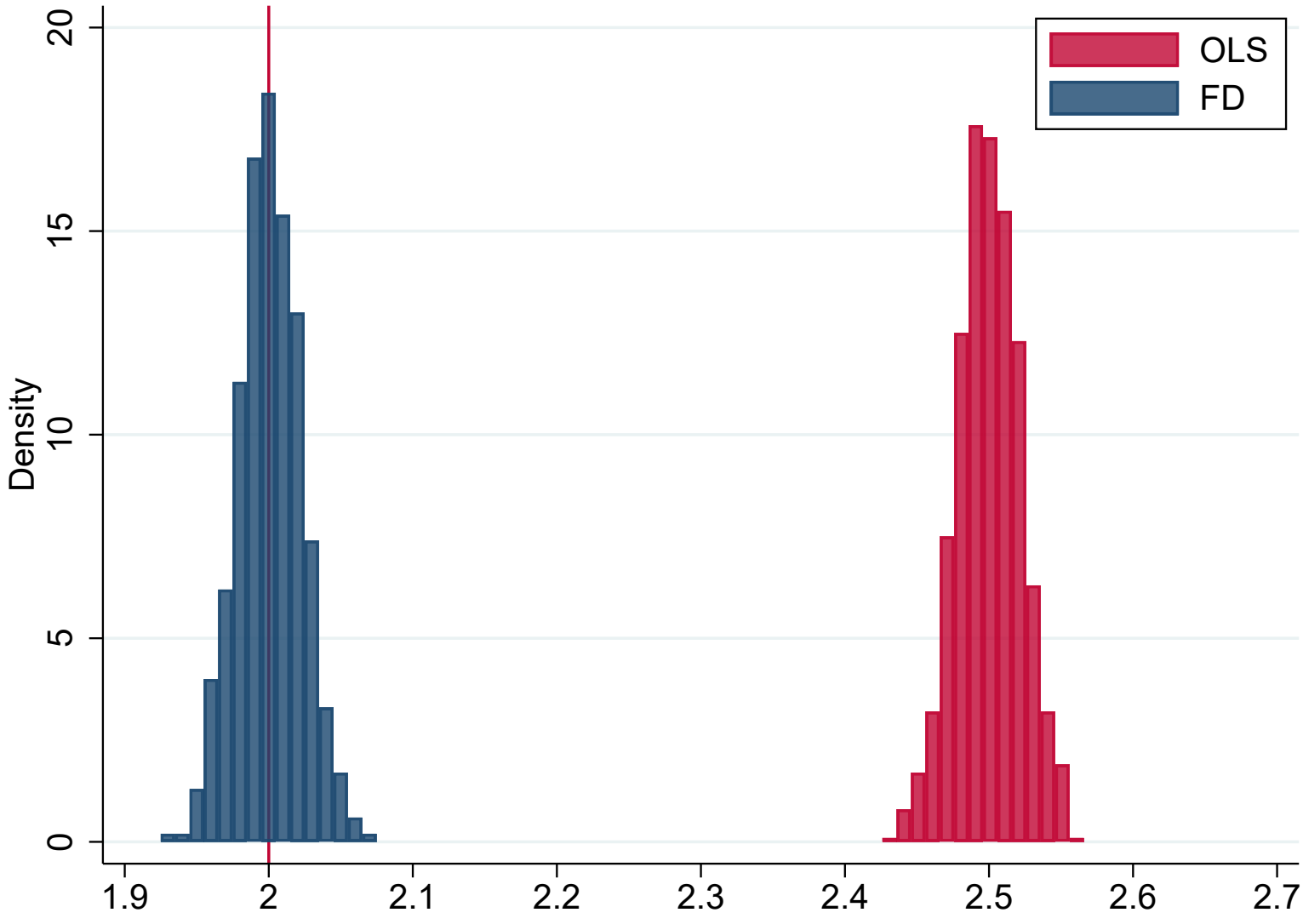
$$A : p\lim \hat{\beta}_1^{OLS} - \beta_1 > 0$$
$$B : p\lim \hat{\beta}_1^{OLS} - \beta_1 = 0$$
$$C : p\lim \hat{\beta}_1^{OLS} - \beta_1 < 0$$

Sammenligning af estimatorer: Simulering

```
. estimates table OLS FD FE RE, b(%9.4f) se(%9.4f)

------------------------------------------------------------
    Variable |    OLS         FD          FE          RE
-------------+----------------------------------------------
           x |    2.5289                  2.0167      2.3269
             |    0.0190                  0.0225      0.0195
          dx |                2.0167
             |                0.0225
       _cons |   -0.0026                  0.9620      0.3778
             |    0.0666                  0.0531      0.0828
------------------------------------------------------------
                                                 legend: b/se
```

Sammenligning af estimatorer: Simulering



Andre typer panel data

Indtil nu har vi anvendt panel data i tid til at fjerne konstante individeffekter.

Vi kan bruge andre typer af "panel data" til at fjerne andre typer af effekter. Fx kan vi sammenligne

- søskende for at fjerne familie fixed effects.
- kollegaer for at fjerne virksomheds fixed effects.
- klassekammerater for at fjerne klasse fixed effects.

FE/FD estimatoren på disse former for data udnytter at der er variation af i x inden for hvert "cluster" (familie, virksomhed, klasse).

Andre typer panel data: Eksempel

Estimates of the Economic Return to Schooling from a New Sample of Twins

By ORLEY ASHENFELTER AND ALAN KRUEGER*

This paper uses a new survey to contrast the wages of genetically identical twins with different schooling levels. Multiple measurements of schooling levels were also collected to assess the effect of reporting error on the estimated economic returns to schooling. The data indicate that omitted ability variables do not bias the estimated return to schooling upward, but that measurement error does bias it downward. Adjustment for measurement error indicates that an additional year of schooling increases wages by 12–16 percent, a higher estimate of the economic returns to schooling than has been previously found. (*JEL* J31)

Ashenfelter & Krueger (1994). The American Economic Review

Andre typer panel data: Eksempel

Et berømt studie af Ashenfelter and Krueger (1994) estimerer afkastet af uddannelse ved brug af data for tvillinger.

Model:

$$\log(wage_{fj}) = \beta_0 + \beta_1 educ_{fj} + \alpha_f + u_{fj},$$

hvor f referer til familie $f = 1, ..., n$ og $j = 1, 2$ til tvilling 1 og 2.

- Hvad er fortolkningen af modellen?
- Hvorfor er det problematisk at estimere modellen med OLS?

FD modellen ser således ud

$$\Delta \log(wage_{fj}) = \beta_1 \Delta educ_{fj} + \Delta u_{fj},$$

Er antagelse FD.4: $E(u_{fj}|educ_{f1}, educ_{f2}, a_i) = 0$ rimelig?

Opsummering

Opsummering

Gentagende tværsnitsdata:

- Tillader at vi undersøge for parameterstabilitet over tid.
- Vi laver politikevalueringer ved **Difference-in-difference estimation**, hvis vi kan finde **en kontrolgruppe**.
- Difference-in-difference estimation bygger på en antagelse om parallel trends **i fravær af politikindgrebet**.

Paneldata:

- Det er normalt mere besværligt og dyrere at indsamle panel data på de samme individer (risiko for sample attrition).
- Med paneldata kan vi tillade for **uobserverede individspecifikke faktorer** i fejlledet.
- Vi kan fjerne den individspecifik heterogenitet vha. FD eller FE.

Opsummering

Paneldata:

- FD og FE bruger kun variation inden for individerne (within variation).
  - FD estimator er OLS på første differencer.
  - FE estimator er OLS på afvigelser fra individ gennemsnit.
- FD og FE er konsistente selv hvis $cov(x_{it}, a_i) \neq 0$.
- RE og pooled OLS er kun konsistente hvis $cov(x_{it}, a_i) = cov(x_{it}, u_{ti}) = 0$.
- Ikke muligt at estimere parametre til tidsinvariante variable med FD og FE.
- Vi kan bruge FE og FD til at fjerne andre typer af uobserveret heterogenitet (fx familie, virksomhed og klasseeffekter).