

Kausal inferens II: Paneldata

Metode 3

Link til slides: kortlink.dk/2acd

Søren Damsbo-Svendsen
sdas@ifs.ku.dk

Institut for Statskundskab
Københavns Universitet

Uge 13

GENÅBNING!

- Men ikke for os helt endnu...
- Jeg venter ligesom jer i spænding på næste ud melding
- Sandsynligvis afholder jeg en lille **spørgetime** inden eksamen, så forhåbentlig kan vi nå at mødes IRL en enkelt gang



CSS, 2021

Oversigt

Uge	Holdtime	Emne	Øvelsesopgave
6		Kvantitativ indholdsanalyse	1
7		Diskursanalyse I	2
8	25-26 februar 2021	Diskursanalyse II	3
9	04-05 marts 2021	Interaktioner og modelspecifikation	4
10	11-12 marts 2021	Logistisk regression	5
11	18-19 marts 2021	Multilevel analyse	6
12		Kausal inferens I: Kausalitet og instrumentvariable (IV)	7
13		Påskeferie	
14	08-09 april 2021	Kausal inferens II: Panedata	8
15	15-16 april 2021	Kausal inferens III: Eksperimentelle designs	9
16		Kausal inferens IIII: Regression Discontinuity (RD)	10
17		Social Data Science: Big Data	11
18		Process tracing	12
19		Kriterier for god videnskab	
20		Opsamling/spørgetime	
23		Aflevering af skriftlig hjemmeopgave (07 juni 2021)	

Dagens program

1. Info om vejledning
2. Kausalitet og lidt overblik og afgrænsning
3. Panel*data*
4. Fixed effects-modeller vs. almindelig OLS
 - Hvad er fidusen?
 - Øvelsesopgaver
 - Stata

Første vejledning

- **15 min. per gruppe** (10 min. hvis man skriver alene)
- Datoer: **16/4** (formiddag), **19/4** (eftermiddag) og **23/4** (eftermiddag)
- Foregår på **Zoom**: Ét link til alle (begge hold) → log på et par min. før jeres vejledning
- **Tilmelding i Google Sheets** - først til mølle
- Jeg sender **links** til Google Sheet og Zoom

Forberedelse

- Send en **superkort beskrivelse** af jeres idé indeholdende:
 - Navn(e), hold, vejledningsnummer (jf. Google Sheet)
 - Foreløbig problemformulering
 - Primær metode
 - Data
 - Evt. andre væsentlige overvejelser
 - Spørgsmål!
- **Format:** Max. en halv normalside → PDF-fil (filnavn der begynder med vejledningsnummer)
- **Deadline:** Send til mig på **sdas@ifs.ku.dk** senest 24 timer før jeres vejledning

Dagens formål

1. At forstå hvorfor paneldata ofte er smartere end tværsnitsdata
2. At få en fornemmelse for logikken bag fixed effects-modeller
3. At vide hvordan man klargør og analyserer paneldata i Stata

Pensum

Stock, J. H., & Watson, M. W. (2003). Introduction to Econometrics (Vol. 104). Boston: Addison Wesley. 396-419. **Kapitel 10: Regression with Panel Data.**

Recap på kausal inferens

- I har flere gange hørt mig sige, at vi bare har antaget, at sammenhængen mellem et X og Y er en "effekt", og at vi "skyder kausalitetsdiskussionen til hjørne" → det er slut nu!
- Hariris forelæsning om kausalitet og IV før påske var en ekstremt vigtig (og god!) forelæsning
 - Han præsenterede **potential outcomes**-frameworket, der er den dominerende forståelse af kausalitet
 - Kort sagt handler kausalitet om sammenligninger - **kontrafaktiske sammenligninger**
 - Effekten af $X \rightarrow Y$ er **forskellen mellem det faktiske Y for en person, der har fået "treatment" (X), og det Y personen ville have haft uden treatment**
 - Sidstnævnte ($Y_1 | D=0$) er **kontrafaktisk** og kan ikke observeres → derfor bruger vi diverse smarte påfund og krumspring (IV-regression og andre **naturlige eksperimenter**)
- Med **OLS og paneldataanalyse** forsøger vi at lave fornuftige sammenligninger mellem det, vi antager er essentielt set ens enheder *efter vi har kontrolleret for en række faktorer* ("all else being equal")
 - vi kan dog aldrig vide os helt sikre på ikke at mangle væsentlige kontrolvariable, og derfor er det en forholdsvis **usikker og følsom vej til kausal inferens** ift. naturlige eksperimenter
 - **paneldata** har dog nogle særlige fordele, som vi skal se på i dag

Lidt vigtig intro og afgrænsning

- Panedata er slet og ret en type **data** - ligesom tværsnitsdata (fx alm. surveys) → ikke så ophidsende
- Men denne type data giver nogle **fede analysemuligheder**:
 - panedata kan bruges til at undersøge dynamik (udvikling/påvirkning over tid)
 - kan besvare spørgsmål a la 'Hvad er effekten af at *b/ive* et demokrati' vs. 'Hvad er effekten af at *være* et demokrati' (ændringer i stedet for niveau)
 - i praksis bruges panedata først og fremmest som en smart **kontrolstrategi** → forbedret grundlag for **kausal inferens** ift. almindelig OLS
- Panedata kan analyseres på mange måder (bl.a. first-difference, random effects, tidsserieanalyse)
 - i praksis anvendes mest **fixed effects-modeller** og det er kun disse, vi ser på
- Hariri var meget grundig, teknisk og pædagogisk til **forelæsningen**
 - (gen)se hans forelæsning for at få fingerspitzengefühl med fixed effects og forskellige estimatorer
 - denne time bliver mere **anvendelsesorienteret** → hvornår/hvorfor/hvordan bruger vi panedata

Paneldata - Den Ultrakorte Version

1. Paneldata er kort sagt gentagne observationer af mange enheder. Når vi har mulighed for at arbejde med paneldata - *og alternativet er analyse af tværsnitsdata* - er paneldata næsten altid en fordel
2. Paneldata giver os redskaber til at håndtere nogle problemer - at eliminere nogle former for **omitted variable bias**. Dato. giver det et stærkere (omend upefekt) fundament for **kausal inferens**
3. De vigtigste redskaber/modeller at kende og forstå er **fixed effects-modeller**, som essentielt set kan koges ned til **OLS med kontrol for enheden samt evt. tidsperioden**
4. I praksis fungerer det ved, at man **tilføjer dummyer for hver enhed** (samt evt. for hver tidsperiode), hvormed effekter **estimeres inden for hver enhed**

Paneldata

Datatyper

Tværsnitsdata/cross-sectional data

- Et tværsnit af en given population målt på ét tidspunkt - eller et sample (udsnit)
- **Kun rumlig/spatial variation**, dvs. variation mellem enheder
- Eksempler: Valgundersøgelsen, European Social Survey, etc. → alt det I har lært i metode indtil nu

Tidsseriedata

- (Typisk) en enkelt enhed/observation målt over tid
- **Kun tidslig/temporal variation**, dvs. udvikling over tid
- Eksempler: Udviklingen i opbakning til statsministerpartiet fra 2001-2020

Paneldata

- Kombinerer tværsnits- og tidsseriedata: Data for flere enheder og flere perioder
- **Rumlig og tidslig variation**, dvs. tidslig variation *indenfor* enheder og mellem enheder
- Eksempler: Landepanel eller panel af surveyrespondenter

Paneldata - struktur

Paneldata kan overordnet set have to strukturer. Data i tabellerne nedenfor er ens. **Hvad er forskellen?**

Bredt format

enhed	x1	x2	y1	y2
a	1	1	4	5
b	0	1	5	8
c	1	0	3	3

Langt format

enhed	tid	x	y
a	1	1	4
a	2	1	5
b	1	0	5
b	2	1	8
c	1	1	3
c	2	0	3

Paneldata - struktur

Paneldata kan overordnet set have to strukturer. Data i tabellerne nedenfor er ens. **Hvad er forskellen?**

Bredt format

enhed	x1	x2	y1	y2
a	1	1	4	5
b	0	1	5	8
c	1	0	3	3

En række per enhed - en kolonne per variabel og periode

Langt format

enhed	tid	x	y
a	1	1	4
a	2	1	5
b	1	0	5
b	2	1	8
c	1	1	3
c	2	0	3

En række per enhed-periode - kun en kolonne per variabel

→ Vi skal altid bruge **langt format** til regressionsanalyse!

Hvordan ændrer man format fra bredt til langt?

Stata

reshape long ["brede" variable der skal transformeres], i([ny enheds-variabel]) j([ny tidsvariabel])

Det er vigtigt, at variabelnavnene, der angives efter "long", er forsynet med en angivelse af tidsperiode/bølge på en konsistent måde (f.eks. bnp10, bnp12 og bnp14 for BNP i år 2010, 2012 og 2014)

Eksempel

reshape long x y, i(enhed) j(tid)

Problemer? Skriv "help reshape" (sådan kan I altid få hjælp til en kommando) eller søg på Google

Analyse af paneldata

1. Notation
2. Alm. OLS
3. Fixed effects

Notation

For at holde styr på variable, der både varierer tidsligt og rumligt, tilføjes et ekstra fodtegn, **t**, udover det velkendte **i** → **i** refererer til de observerede *enheder*, mens **t** refererer til *tidsperioden*, enhederne observeres i

Når vi skriver en **statistisk model** på formel, angiver fodtegnet altså, **hvor(dan) hver variabel varierer**:

- $X_i \rightarrow$ variablen X varierer kun mellem enheder
- $X_t \rightarrow$ variablen X varierer kun over tid
- $X_{it} \rightarrow$ variablen X varierer *både* mellem enheder og over tid. Der er altså en måling for hver periode t for hver enhed i

Eksempel

$$Y_{it} = \alpha + \beta_1 X_{1,it} + \beta_2 X_{2,i} + \beta_3 X_{3,t} + \epsilon_{it}$$

Hvad kan vi sige om variationen i de tre uafhængige variable?

Øvelsesopgave 8

Hvorfor vandt Trump det amerikanske præsidentvalg i 2016?



- Mutz hævder, at Trump-vælgere er motiveret af en **følelse af tabt status** snarere end økonomiske vanskeligheder
- Vi har paneldata med samme respondenter (**MNO**) målt to gange - i 2012 og 2016 (**wave**)
 - Den *afhængige* variabel er **cutdifftherm**, som udtrykker respondentens **positive følelser over for Trump** fratrukket positive følelser over for Hillary (0-19; højere værdi = mere positiv holdning til Trump)
 - Den primære *uafhængige* variabel er **chinasef** ("China threat"), som udtrykker **i hvilken grad respondenten opfatter Kina som en trussel mod USA** fra 1 (trussel) til 7 (mulighed)
 - En **positiv sammenhæng** mellem X og Y ville altså indikere, at mindre trussel → mere støtte til Trump

Øvelsesopgave 8.1

Cross-sectional analyse af 2016-bølgen

- Vi ignorerer datas panelstruktur og anvender kun observationer fra 2016 (= tværsnit)
- Fit en OLS-model med holdning til Trump (*cutdifftherm*) som afhængig variabel og *chinaself* som uafhængig
 - **Kontrolvariable:** indkomst (income), ledighed (lookingforwork), personlig økonomi (personeco) og parti (xparty3)
- Hvordan ville sådan en model se ud?

$$\begin{aligned} \text{cutdifftherm}_i = & \alpha + \beta_1 \text{ChinaThreat}_i + \beta_3 \text{Indkomst}_i + \beta_4 \text{Ledig}_i \\ & + \beta_5 \text{Økonomi}_i + \beta_6 \text{Parti_Independent}_i + \beta_7 \text{Parti_Democrat}_i + \epsilon_i \end{aligned}$$

Øvelsesopgave 8.1

Cross-sectional analyse af 2016-bølgen

Hvad er sammenhængen
mellem "China threat"
(chinaside) og støtte til
Trump?

... efter kontrol for de fire
kontrolvariable

```
. reg cutdifftherm c.chinaside i.xparty3 c.income i.lookingforwork c.personeco if wave ==  
> 1, vsquish
```

Source	SS	df	MS	Number of obs	=	1,124
Model	24287.3756	6	4047.89594	F(6, 1117)	=	310.12
Residual	14579.6208	1,117	13.0524806	Prob > F	=	0.0000
Total	38866.9964	1,123	34.6099701	R-squared	=	0.6249

	cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaside	-.4631329	.073134	-6.33	0.000	-.6066285	-.3196373
xparty3						
2	-2.698191	.5687546	-4.74	0.000	-3.814139	-1.582243
3	-8.047411	.2362199	-34.07	0.000	-8.510896	-7.583927
income	.0224104	.0268811	0.83	0.405	-.0303327	.0751535
i.lookingforwork						
1	-.8710731	.602032	-1.45	0.148	-2.052314	.3101678
personeco						
_cons	-1.1316	.1210906	-9.35	0.000	-1.369191	-.8940099
_cons	18.14209	.4374188	41.48	0.000	17.28384	19.00035

Øvelsesopgave 8.1

Cross-sectional analyse af 2016-bølgen

Overordnet set, en **stærk negativ sammenhæng**

Når en respondent går 1 point højere op på China threat-skalaen (1-7), altså opfatter Kina *mindre* som trussel, reduceres støtten til Trump med 0.46 skalapoint ($p < 0.000$)

Det peger altså isoleret set på, at "China threat" havde en stærk effekt på støtten til Trump

```
. reg cutdifftherm c.chinaself i.xparty3 c.income i.lookingforwork c.personeco if wave ==  
> 1, vsquish
```

Source	SS	df	MS	Number of obs	=	1,124
Model	24287.3756	6	4047.89594	F(6, 1117)	=	310.12
Residual	14579.6208	1,117	13.0524806	Prob > F	=	0.0000
Total	38866.9964	1,123	34.6099701	R-squared	=	0.6249

cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaself	-.4631329	.073134	-6.33	0.000	-.6066285 -.3196373
xparty3					
2	-2.698191	.5687546	-4.74	0.000	-3.814139 -1.582243
3	-8.047411	.2362199	-34.07	0.000	-8.510896 -7.583927
income	.0224104	.0268811	0.83	0.405	-.0303327 .0751535
i.lookingforwork	-.8710731	.602032	-1.45	0.148	-2.052314 .3101678
personeco	-1.1316	.1210906	-9.35	0.000	-1.369191 -.8940099
_cons	18.14209	.4374188	41.48	0.000	17.28384 19.00035

Øvelsesopgave 8.2

Cross-sectional analyse af 2016-bølgen - med yderligere kontrolvariable

Tilføj kontrolvariable, der angiver om respondenten selv har haft gavn af international handel (**tradeper**) og er *for* indvandring (**proimmself**)

Er der stadigvæk en sammenhæng mellem chinaself og cutdifftherm?
Har den ændret sig?

```
. reg cutdifftherm c.chinaself i.xparty3 c.income i.lookingforwork c.personeco c.tradeper  
> c.proimmself if wave == 1, vsquish
```

Source	SS	df	MS	Number of obs	=	1,101
Model	25560.6134	8	3195.07667	F(8, 1092)	=	278.68
Residual	12519.9389	1,092	11.4651455	Prob > F	=	0.0000
Total	38080.5522	1,100	34.6186838	R-squared	=	0.6712
				Adj R-squared	=	0.6688
				Root MSE	=	3.386

cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaself	-.1956404	.0744996	-2.63	0.009	-.341819 -.0494618
xparty3					
2	-2.173155	.5485854	-3.96	0.000	-3.249556 -1.096754
3	-6.796203	.2446972	-27.77	0.000	-7.276333 -6.316073
income	.0440298	.0255099	1.73	0.085	-.0060242 .0940837
i.lookingforwork	-.5092596	.5654403	-0.90	0.368	-1.618732 .6002127
personeco	-.8485995	.1221547	-6.95	0.000	-1.088284 -.6089151
tradeper	-.3400787	.1304048	-2.61	0.009	-.5959509 -.0842064
proimmself	-.673535	.0572567	-11.76	0.000	-.7858806 -.5611894
_cons	19.18884	.4384577	43.76	0.000	18.32853 20.04916

Øvelsesopgave 8.2

Cross-sectional analyse af 2016-bølgen - med yderligere kontrolvariable

Sammenhængen reduceres markant fra 0.46 til 0.20, men er stadig signifikant ($p < 0.01$)

Side note: Man skal være yderst varsom med at kontrollere for (holdnings)variable, der i sig selv er påvirket af X → risiko for **post-treatment bias**

```
. reg cutdifftherm c.chinaself i.xparty3 c.income i.lookingforwork c.personeco c.tradeper  
> c.proimmsself if wave == 1, vsquish
```

Source	SS	df	MS	Number of obs	=	1,101
Model	25560.6134	8	3195.07667	F(8, 1092)	=	278.68
Residual	12519.9389	1,092	11.4651455	Prob > F	=	0.0000
Total	38080.5522	1,100	34.6186838	R-squared	=	0.6712
				Adj R-squared	=	0.6688
				Root MSE	=	3.386

cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaself	-.1956404	.0744996	-2.63	0.009	-.341819 -.0494618
xparty3					
2	-2.173155	.5485854	-3.96	0.000	-3.249556 -1.096754
3	-6.796203	.2446972	-27.77	0.000	-7.276333 -6.316073
income	.0440298	.0255099	1.73	0.085	-.0060242 .0940837
i.lookingforwork	-.5092596	.5654403	-0.90	0.368	-1.618732 .6002127
personeco	-.8485995	.1221547	-6.95	0.000	-1.088284 -.6089151
tradeper	-.3400787	.1304048	-2.61	0.009	-.5959509 -.0842064
proimmsself	-.673535	.0572567	-11.76	0.000	-.7858806 -.5611894
_cons	19.18884	.4384577	43.76	0.000	18.32853 20.04916

Men er det en *kausal* sammenhæng?

- Kan vi med ro i maven konkludere, at X (opfattet trussel fra Kina) har en kausal effekt på Y (støtte til Trump), dvs. *forårsager* Y?
 - Hvorfor/hvorfor ikke?
 - Hvad skal der til?
 - Snak om det i 3-5 min.
-
- Den helt **centrale forudsætning** er, at der ikke er endogenitetsbias, herunder **omitted variable bias** (OVB)
 - at vi har kontrolleret for alle **faktorer, som kan påvirke både X og Y** (a.k.a. fravær af spuriøsitet)
 - Vi kunne ret hurtigt finde nogle faktorer, der potentielt påvirker trusselsopfattelse *og* støtte til Trump og dermed *confounder* sammenhængen

Omitted variable bias

Typologi over faktorer, der kan medføre OVB

	Tidsvariant	Tidsinvariant (konstant)
Enhedsvariant	A - farlig	B - håndterbar
Enhedsinvariant (ens)	C - håndterbar	D - OK

- D. Fuldstændigt **ens og konstante faktorer** (Planet = jorden) → behøver vi ikke at bekymre os om
- C. Faktorer der **varierer over tid**, men er ens for alle
- B. Faktorer der **varierer mellem enheder**, men er konstante over tid
- A. Faktorer der **både varierer over tid og mellem enheder**
- Med **tværsnitsdata/OLS** skal vi kunne kontrollere for det hele (bortset fra D)

Omitted variable bias

Typologi over faktorer, der kan medføre OVB

	Tidsvariant	Tidsinvariant (konstant)
Enhedsvariant	A - farlig	B - håndterbar
Enhedsinvariant (ens)	C - håndterbar	D - OK

- Med **panedata** kan vi anvende **fixed effects** til at eliminere nogle former for OVB
 - **enheds-fixed effects** eliminerer B
 - **tids-fixed effects** eliminerer C
 - D er ufarlig
- Det magiske er, at **vi ikke engang behøver kunne observere disse faktorer**
- Kun **A er tilbage** (*ideosynkratisk OVB*, jf. Hariri) og må håndteres med kontrolvariable

"[Two-way fixed effects] is **immune to omitted variable bias from variables that are constant either over time or across [units]**"

(Stock & Watson)

Så hvordan estimeres (two-way) fixed effects?

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \gamma_t + \epsilon_{it}$$

Sådan! Det ser avanceret ud, men det er i virkeligheden tæt på, hvad I allerede kan. Læg mærke til følgegnene

- **Enheds-fixed effects**, α_i , opnår vi ved tilføje en **dummy for hver enhed** som kontrolvariabel
 - vi "kontrollerer for enhedsvariation" og giver hver enhed sit eget α_i (konstantled), som tager højde for enheds-konstante faktorer
- **Tids-fixed effects**, γ_t , opnår vi ved at tilføje en **dummy for hver tidsperiode** som kontrolvariabel
 - vi "kontrollerer for tidsvariation" og fjerner den overordnede tidstrend
- Den tilbageværende variation er kun **indenfor hver enhed** → kaldes derfor også en *within-estimator*

I Stata er der to muligheder

- (1) "**xtset enhedsid tidsvariabel**" efterfulgt af "**xtreg Y X i.tidsvariabel, fe**"
- (2) "**reg Y X i.tidsvariabel, absorb(enhedsvariabel)**"
 - absorb(enhedsvariabel) svarer til at tilføje "i.enhedsvariabel" bare uden at estimatorne printes
- samme kommandoer uden "**i.tidsvariabel**", hvis vi kun vil have enheds-fixed effects

Øvelsesopgave 8.3

Tilbage til Trump og Mutz!



Brug **xtset**-kommandoen til at specifiere panelstrukturen for Stata:

```
. xtset MNO wave // angiver panelstruktur "xtset unit time"  
    panel variable: MNO (unbalanced)  
    time variable: wave, 0 to 1  
        delta: 1 unit
```



Øvelsesopgave 8.4

Enheds-fixed effects

- I opgave 8.4 bliver vi igen bedt om at estimere β_1
- Men denne gang udnytter vi panelstrukturen og den tidslige variation (*t*-fodtegn) og tilføjer enheds-fixed effects, α_i

$$\begin{aligned} cutdifftherm_{it} = & \alpha_i + \beta_1 ChinaThreat_{it} + \beta_3 Indkomst_{it} + \beta_4 Ledig_{it} + \beta_5 Økonomi_{it} \\ & + \beta_6 Parti_Independent_{it} + \beta_7 Parti_Democrat_{it} + \epsilon_{it} \end{aligned}$$

- Konsekvens: Nu udtrykker β_1 , **hvor meget Y ændrer sig, når X vokser med +1 indenfor personer**

Øvelsesopgave 8.4

Enheds-fixed effects

Hvordan adskiller resultaterne sig fra
tværsnitsanalysen?

Hænger *chinaself* stadig sammen med støtte til
Trump?

```
. xtreg cutdifftherm c.chinaself i.xparty3 c.income i.lookingforwork c.personeco, fe vsqu  
> ish
```

```
Fixed-effects (within) regression                               Number of obs     =      2,268  
Group variable: MNO                                     Number of groups  =      1,201  
  
R-sq:                                                 Obs per group:  
    within     =  0.0665                                         min =          1  
    between    =  0.6287                                         avg =        1.9  
    overall    =  0.5626                                         max =          2  
  
                                                F(6,1061)     =     12.60  
corr(u_i, Xb)  =  0.6270                                         Prob > F      =  0.0000
```

	cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaself	-.0974394	.0733319	-1.33	0.184	-.2413314	.0464527
xparty3						
2	1.432148	.7353776	1.95	0.052	-.0108113	2.875108
3	-2.725938	.4089121	-6.67	0.000	-3.528306	-1.92357
income	-.0333072	.0434521	-0.77	0.444	-.1185689	.0519546
1.lookingforwork	-.375675	.4692109	-0.80	0.424	-1.296362	.5450118
personeco	-.3818377	.1162381	-3.28	0.001	-.6099204	-.1537549
_cons	12.81947	.6248674	20.52	0.000	11.59336	14.04559
sigma_u	4.3405224					
sigma_e	2.6778744					
rho	.72431007	(fraction of variance due to u_i)				

F test that all u_i=0: F(1200, 1061) = 2.57

Prob > F = 0.0000

Øvelsesopgave 8.4

Enheds-fixed effects

Koefficienten for *chinself* er -0.10 i den forventede retning, men sammenhængen er **ikke signifikant** ($p=0.18$)

Selvom folk, der opfatter Kina som en trussel, bedre kan lide Trump, **jf. tværsnitsanalysen**, betyder det ikke, at *ændringer i opfattelsen af Kina* hænger sammen med (eller forårsager) *ændringer i støtten til Trump*, **jf. paneldata-analysen**

Derfor: Tværsnitsanalysens sammenhæng mellem opfattet trussel og støtte til Trump er interessant, men det vil være misvisende at fortolke sammenhængen kausalt

```
. xtreg cutdifftherm c.chinself i.xparty3 c.income i.lookingforwork c.personeco, fe vsqu  
> ish
```

```
Fixed-effects (within) regression                               Number of obs     =      2,268  
Group variable: MNO                                Number of groups  =      1,201  
  
R-sq:                                                 Obs per group:  
    within  =  0.0665                                         min =         1  
    between =  0.6287                                       avg =       1.9  
    overall =  0.5626                                      max =         2  
  
                                                F(6,1061)        =     12.60  
corr(u_i, Xb)  =  0.6270                                 Prob > F        =  0.0000
```

cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinself	-0.0974394	.0733319	-1.33	0.184	-.2413314 .0464527
xparty3					
2	1.432148	.7353776	1.95	0.052	-.0108113 2.875108
3	-2.725938	.4089121	-6.67	0.000	-3.528306 -1.92357
income	-0.0333072	.0434521	-0.77	0.444	-.1185689 .0519546
1.lookingforwork	-0.375675	.4692109	-0.80	0.424	-1.296362 .5450118
personeco	-0.3818377	.1162381	-3.28	0.001	-.6099204 -.1537549
_cons	12.81947	.6248674	20.52	0.000	11.59336 14.04559
sigma_u	4.3405224				
sigma_e	2.6778744				
rho	.72431007	(fraction of variance due to u_i)			

F test that all $u_i=0$: F(1200, 1061) = 2.57

Prob > F = 0.0000

Øvelsesopgave 8.4

Enheds-fixed effects med yderligere kontrolvariable

Vi tilføjer **to kontrolvariable** mere (*tradeper* og *proimmself*)

Se på koefficienten for **chinasef** igen

Nu er den blevet endnu mindre signifikant (p=0.776)

```
. xtreg cutdifftherm c.chinasef i.xparty3 c.income i.lookingforwork c.personeco c.tradeper  
> er c.proimmself, fe vsquish
```

```
Fixed-effects (within) regression                               Number of obs     =      2,210  
Group variable: MNO                                     Number of groups  =      1,188  
  
R-sq:                                                 Obs per group:  
    within  =  0.0969                                         min =          1  
    between =  0.6356                                        avg =        1.9  
    overall =  0.5807                                       max =          2  
  
                                                F(8,1014)       =     13.60  
corr(u_i, Xb)  =  0.5979                                 Prob > F        =  0.0000
```

cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinasef	-.0210522	.0739628	-0.28	0.776	-.1661899 .1240855
xparty3					
2	2.083756	.7857964	2.65	0.008	.5417827 3.625729
3	-2.283514	.4086308	-5.59	0.000	-3.085372 -1.481655
income	.0041297	.0431087	0.10	0.924	-.0804628 .0887221
i.lookingforwork	-.5377623	.4627657	-1.16	0.245	-1.44585 .3703257
personeco	-.3621272	.116696	-3.10	0.002	-.5911205 -.1331339
tradeper	-.3296306	.1030299	-3.20	0.001	-.5318067 -.1274544
proimmself	-.3270944	.067024	-4.88	0.000	-.4586161 -.1955727
_cons	14.00227	.6837934	20.48	0.000	12.66046 15.34409
sigma_u	4.1634888				
sigma_e	2.5968456				
rho	.71992946	(fraction of variance due to u_i)			

F test that all u_i=0: F(1187, 1014) = 2.47

Prob > F = 0.0000

Øvelsesopgave 8.4

Two-way fixed effects

Vi tilføjer tids-fixed effects ("i.wave"), dvs. en dummy for hver periode

Dermed estimeres en model med **two-way fixed-effects** (kaldes også *difference-in-differences-estimatoren*)

Koefficienten udtrykker, hvor meget *mere Y* ændrer sig over tid for personer med X+1 (mindre trussel) sammenlignet med personer med X (mere trussel)?
Eller **hvor meget ændres tidstrenden i Y, når X øges med +1 (mindre trussel)?**

Er der en effekt? (se på *chinasef* igen)

```
. xtreg cutdifftherm c.chinasef i.xparty3 c.income i.lookingforwork c.personeco c.tradeper  
> er c.proimmself i.wave, fe vsquish
```

```
Fixed-effects (within) regression                               Number of obs     =      2,210  
Group variable: MNO                                     Number of groups  =      1,188  
  
R-sq:                                                 Obs per group:  
    within  = 0.1040                                         min =          1  
    between = 0.6375                                         avg =        1.9  
    overall = 0.5824                                         max =          2  
  
                                                F(9,1013)       =     13.06  
corr(u_i, Xb)  = 0.6004                                 Prob > F        = 0.0000
```

	cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinasef	-.0241714	.0737169	-0.33	0.743	.1688267	.1204839
xparty3						
2	2.265217	.7857184	2.88	0.004	.7233951	3.807039
3	-2.417693	.4099789	-5.90	0.000	-3.222198	-1.613188
income	.0149198	.0431295	0.35	0.729	-.0697136	.0995532
i.lookingforwork	-.629207	.4623068	-1.36	0.174	-1.536396	.2779815
personeco	-.3302001	.1168412	-2.83	0.005	-.5594786	-.1009215
tradeper	-.3168097	.1027758	-3.08	0.002	-.5184875	-.1151319
proimmself	-.2893898	.0681102	-4.25	0.000	-.4230431	-.1557365
1.wave	-.3389965	.1198184	-2.83	0.005	-.5741171	-.1038759
_cons	13.85788	.6833522	20.28	0.000	12.51693	15.19883
sigma_u	4.1678496					
sigma_e	2.5879224					
rho	.72173611					(fraction of variance due to u_i)

F test that all u_i=0: F(1187, 1013) = 2.49

Prob > F = 0.0000

Øvelsesopgave 8.4

Two-way fixed effects

Nej! Nu er β_1 (-0.24) endnu mindre signifikant
(p=0.74)

Hvad fortæller estimatet for **wave** os?

At folk gennemsnitligt set bliver mere negative overfor Trump fra 2012 til 2016

```
. xtreg cutdifftherm c.chinaself i.xparty3 c.income i.lookingforwork c.personeco c.tradeper> er c.proimmself i.wave, fe vsquish
```

```
Fixed-effects (within) regression Number of obs = 2,210  
Group variable: MNO Number of groups = 1,188
```

```
R-sq:  
within = 0.1040 Obs per group:  
between = 0.6375 min = 1  
overall = 0.5824 avg = 1.9  
max = 2
```

```
F(9,1013) = 13.06  
corr(u_i, Xb) = 0.6004 Prob > F = 0.0000
```

	cutdifftherm	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
chinaself	-.0241714	.0737169	-0.33	0.743	.1688267	.1204839
xparty3						
2	2.265217	.7857184	2.88	0.004	.7233951	3.807039
3	-2.417693	.4099789	-5.90	0.000	-3.222198	-1.613188
income	.0149198	.0431295	0.35	0.729	-.0697136	.0995532
1.lookingforwork	-.629207	.4623068	-1.36	0.174	-1.536396	.2779815
personeco	-.3302001	.1168412	-2.83	0.005	-.5594786	-.1009215
tradeper	-.3168097	.1027758	-3.08	0.002	-.5184875	-.1151319
proimmself	-.2893898	.0681102	-4.25	0.000	-.4230431	-.1557365
1.wave	-.3389965	.1198184	-2.83	0.005	-.5741171	-.1038759
_cons	13.85788	.6833522	20.28	0.000	12.51693	15.19883
sigma_u	4.1678496					
sigma_e	2.5879224					
rho	.72173611					(fraction of variance due to u_i)

F test that all $u_i=0$: F(1187, 1013) = 2.49

Prob > F = 0.0000

Klyngerobuste standardfejl

- **Standardfejl** angiver *præcisionen* af et estimat (en effekt)
- Normale standardfejl antager at alle observationer er uafhængige af hinanden
- Når vi observerer den samme observation flere gange, vil der ofte være udfordringer med **autokorrelation**, dvs. at samme enheds fejllæd er korreleret over tid
 - **normale standardfejl undervurderer derfor usikkerheden**
- En mulig løsning: **Klyngerobuste standardfejl** på enhedsniveau
 - det får Stata til at tage højde for, at observationerne ikke er uafhængige, men at der er flere per person
 - at data indeholder sammenklumpede (clustered) observationer
 - ofte en god idé - i det mindste som supplement
- Tilføjes i **Stata** med optionen **cluster(enhedsvariabel): ", cluster(MNO)"**

Finale: Sammenligning af tre modeller

Tre modeller med alle kontrolvariable samt (1) **enheds-fixed effects**, (2) **two-way fixed effects** og (3) **two-way fixed effects med klyngerobuste standardfejl** på enhedsniveau

Beta-koefficienter øverst, p-værdier nederst

Hvor stor en forskel tids-fixed effects og klyngerobuste standardfejl gør, **afhænger af den konkrete undersøgelse**

Her gør det praktisk talt ingen forskel, hvilket i og for sig er en god ting → indikerer at sammenhængen er **robust**

Bemærk at **klyngerobuste standardfejl** ikke ændrer ved estimererne, men kun ved usikkerheden (standardfejl og p-værdier)

Variable	FE_unit	FE_unittime	FE_ut_clusterrob
chinself	-0.021 0.776	-0.024 0.743	-0.024 0.744
xparty3			
2	2.084 0.008	2.265 0.004	2.265 0.044
3	-2.284 0.000	-2.418 0.000	-2.418 0.000
income	0.004 0.924	0.015 0.729	0.015 0.756
lookingforwork			
1	-0.538 0.245	-0.629 0.174	-0.629 0.144
personeco	-0.362 0.002	-0.330 0.005	-0.330 0.007
tradeper	-0.330 0.001	-0.317 0.002	-0.317 0.006
proimmself	-0.327 0.000	-0.289 0.000	-0.289 0.000
wave			
1		-0.339 0.005	-0.339 0.004
_cons	14.002 0.000	13.858 0.000	13.858 0.000

Opsummering - en samlet fremgangsmåde

1. **Indsaml/find panedata** (enheder observeret i flere perioder) - skal være i **langt format**
2. Fortæl Stata, at data er **panedata**: "xtset enhedsvariabel tidsvariabel"
3. **Kør jeres regressionsanalyser**: "xtreg Y X Z1 Z2 Z3, fe"
4. Undersøg og demonstrér **robusthed** af resultater overfor forskellige specifikationer, herunder kontrol for *tidsvarierende faktorer* (vha. **tids-fixed effects**) og **klyngerobuste standardfejl**
 - xtreg Y X Z1 Z2 Z3 i.tidsvariabel, fe
 - xtreg Y X Z1 Z2 Z3 i.tidsvariabel, fe cluster(enhedsvariabel)
5. Test **forudsætninger** som normalt med OLS baseret på genkørsel af model(ler)
 - reg Y X Z1 Z2 Z3 i.tidsvariabel, absorb(i.enhedsvariabel)
 - samme antagelser som OLS plus **ingen autokorrelation**

Ulemper ved paneldata

1. Panel data med fixed effects-estimation løser mange af udfordringerne med **omitted variable bias (OVB)** og endogenitet, **men ikke alle**:
 - med **two-way-fixed effects** kan vi eliminere OVB fra tidsinvariante variable og enhedsinvariante variable, men **ideosynkratisk OVB** (fra tids- og enhedsvariante faktorer) **skal stadig håndteres med alm. kontrol**, før vi kan kalde et estimat for en kausaleffekt
 - vi kan ikke afhjælpe **omvendt kausalitet** med paneldata → (også) her må vi argumentere teoretisk
 - (begge elimineres, hvis den uafhængige variabel tildeles tilfældigt → hint til eksperimenter i næste uge)
2. Det kræver langt flere **ressourcer** og mere **tid** at indsamle paneldata end tværsnitsdata. Derfor er paneldata sværere at indsamle selv, mindre udbredt og **mindre tilgængeligt**

Higher-order fixed effects

En nice-to-know sidenote (evt. relevant i bachelorprojekt)

- Vi har kun berørt *enheds*-fixed effects
- Man kan også tilføje **rumlige fixed-effects på et højere niveau end enheds-niveauet** - også med tværsnitsdata (klar parallel til multilevel-analyse)
- Eksempel: Hvis man undersøger en sammenhæng for **individer i forskellige lande**, kan man fx tilføje fixed effects (og klyngerobuste standardfejl) på lande-niveau og dermed håndtere alt det, der gør folk forskellige på tværs af lande

Dagens pointer

- Observationelle tværsnitsanalyser (fx OLS) kan kun stræbe efter at identificere kausaleffekter ved at kontrollere for alle faktorer, der påvirker X og Y → aldrig helt sikkert
- **Paneldata** har den store fordel, at man kan udnytte kombinationen af rumlig og tidslig variation til at lave (two-way) fixed effects-modeller, der **eliminerer nogle typer omitted variable bias**
 - herefter kun problemer med faktorer, der varierer mellem enheder *og* over tid
 - klar forbedring fra OLS men ikke på niveau med naturlige eksperimenter
- Fixed effects tilføjer dummyvariable for enheder og tid, hvilket giver effekten af X på Y *inden for enhederne* over tid eller *difference-in-differences*

Næste gang

- Sidste holdtime 😞 
- Kausal inferens III: Eksperimentelle designs
 - "guldstandarden" inden for kausal inferens
 - vigtigt at få en fornemmelse for → idealet for de fleste tilgange til kausal inferens
- Første vejledning (husk at tilmelde jer)

Tak for i dag!