5: Introduktion til kausal inferens

Videregående kvantitative metoder i studiet af politisk adfærd

Frederik Hjorth
fh@ifs.ku.dk
fghjorth.github.io
@fghjorth

Institut for Statskundskab
Københavns Universitet

4. oktober 2017

1 Opsamling

2 Motivation

3 Hariri (2012)

4 ‘Bad controls’

5 Samii (2016)

6 Eckles & Bakshy (2017)

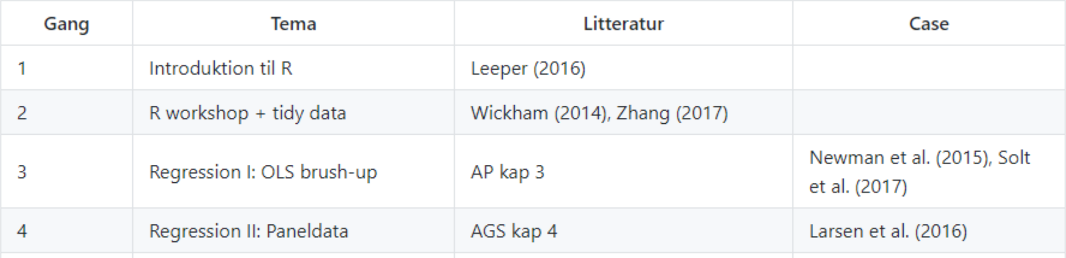
7 Kig fremad

Sidste gang:

- paneldata
- bredt vs. langt dataformat
- clustering
- fixed effects-modeller
- → kontrol for tidsinvariant uobserveret heterogenitet

Fagets opbygning

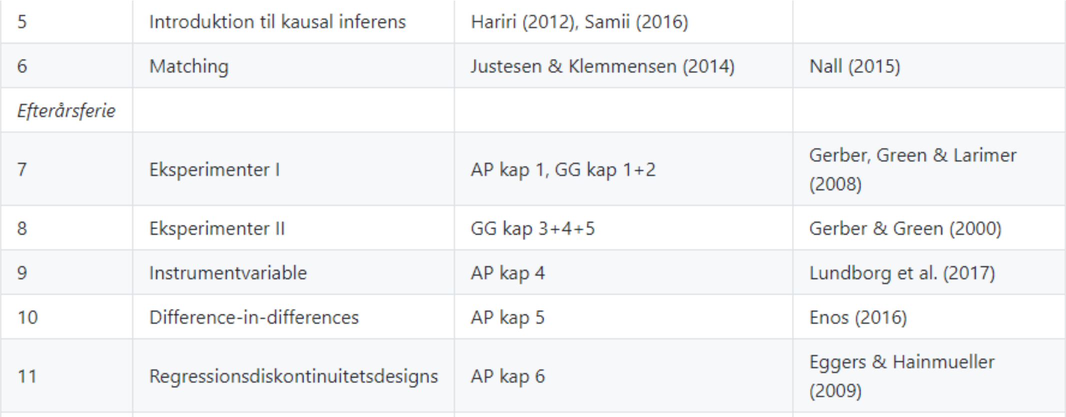
Blok 1



Gang	Tema	Litteratur	Case
1	Introduktion til R	Leeper (2016)	
2	R workshop + tidy data	Wickham (2014), Zhang (2017)	
3	Regression I: OLS brush-up	AP kap 3	Newman et al. (2015), Solt et al. (2017)
4	Regression II: Paneldata	AGS kap 4	Larsen et al. (2016)

Fagets opbygning

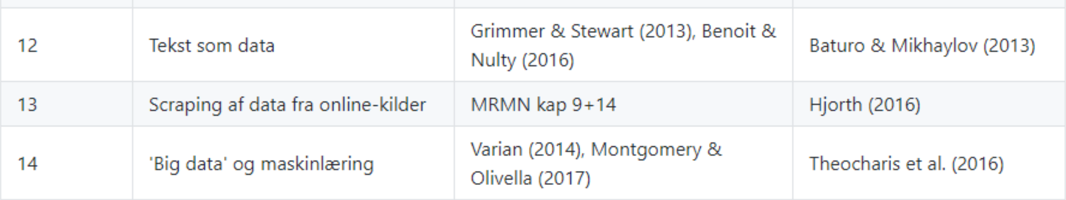
Blok 2



5	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	
6	Matching	Justesen & Klemmensen (2014)	Nall (2015)
Efterårsferie			
7	Eksperimenter I	AP kap 1, GG kap 1+2	Gerber, Green & Larimer (2008)
8	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	Instrumentvariable	AP kap 4	Lundborg et al. (2017)
10	Difference-in-differences	AP kap 5	Enos (2016)
11	Regressionsdiskontinuitetsdesigns	AP kap 6	Eggers & Hainmueller (2009)

Fagets opbygning

Blok 3



12	Tekst som data	Grimmer & Stewart (2013), Benoit & Nulty (2016)	Baturo & Mikhaylov (2013)
13	Scraping af data fra online-kilder	MRMN kap 9+14	Hjorth (2016)
14	'Big data' og maskinlæring	Varian (2014), Montgomery & Olivella (2017)	Theocharis et al. (2016)

Om midterm og eksamen

- midterm
  - frist: fredag d. 3. november kl. 23.59
  - omfang: 1-2 ns.
  - hvad planlægger jeg at gøre i min seminaropgave?
  - hvilke data bruger jeg?
  - hvilke metoder fra faget anvendes?
  - ej bindende
- eksamen
  - frist: fredag d. 22. december kl. 23.59
  - omfang: 10-20 ns.
  - skal demonstrere opfyldelse af fagets læringsmål
  - skal skrives individuelt
  - evt. genindlevering aftales (frist primo januar)
- begge afleveres på Absalon

Peer-effekter i digitale sociale netværk

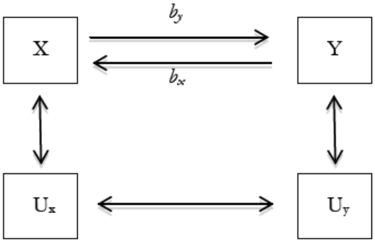
Traditionel tilgang: TEST-kriterierne

- Tidsrækkefølge
- Empirisk sammenhæng
- Fravær af Spuriøsitet
- Teoretisk forklaring

Hariri: disse kriterier hjælper os ikke med at opnå troværdige estimater af kausale effekter

Udgangspunkt: observeret korrelation ml. x og y

Figur 1: Korrelation mellem *x* og *y*



→ vi har confounding selv om hverken u_x eller u_y *i sig selv* er korreleret med X og Y!

Situationen i Fig. 1 som ligningssystem:

$$y = b_y x + a_y u_y \tag{1}$$

$$x = b_x y + a_x u_x \tag{2}$$

Nødvendige restriktioner a og b:

$$b_x = 0 \tag{a}$$

$$E[u_x u_y] = 0 \tag{b}$$

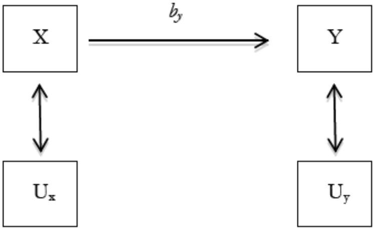
Hermed kan (1) og (2) omskrives til:

$$y = b_y x + a_y u_y \tag{1'}$$

$$x = a_x u_x \tag{2'}$$

Hermed:

Figur 2: Kausal identifikation af størrelsen af påvirkningen af *x* på *y*



> »Imidlertid minder samfundet mere om figur 1 end om et eksperimentelt laboratorium: Næsten alle faktorer påvirker hinanden.(...) Dermed er det også vanskeligt at identificere kausale sammenhænge i empirisk samfundsforskning.Den konstruktive lære er imidlertid, at selvom det er vanskeligt, er det ikke umuligt.Det kræver dog et velgennemtænkt forskningsdesign.«

→ fra *modelbaseret* til *designbaseret inferens*

Vi antager det gammelkendte setup m. skills s_i, indkomst Y_i og confounder ability a_i:

$$Y_i = \alpha + \rho s_i + \gamma a_i + e_i \tag{3}$$

men vi kan kun observere 'late ability' a_{li}, som delvist er en konsekvens af skills s_i:

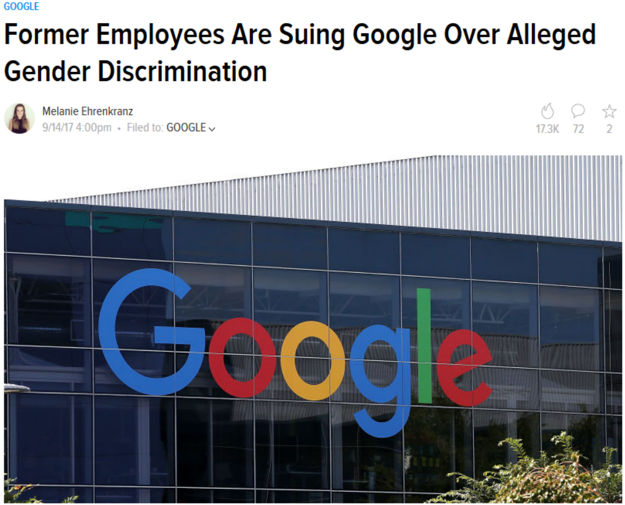
$$a_{li} = \pi_0 + \pi_1 s_i + \pi_2 a_i \tag{4}$$

Konsekvens: vores estimat af *rho* afhænger af γ, π_1 og π_2:

$$Y_i = \left(\alpha - \gamma \frac{\pi_0}{\pi_2}\right) + \left(\rho - \gamma \frac{\pi_1}{\pi_2}\right) s_i + \frac{\gamma}{\pi_2} a_{li} + e_i \tag{5}$$

→ a_{li} er en 'bad control'

> »[W]hen thinking about controls, timing matters. Variables measured before the variable of interest was determined are generally good controls. (...) Because these variables were determined before the variable of interest, they cannot themselves be outcomes in the causal nexus.«



GOOGLE

Former Employees Are Suing Google Over Alleged Gender Discrimination

Melanie Ehrenkranz
9/14/17 4:00pm • Filed to: GOOGLE

17.3K 72 2

`https://gizmodo.com/`
`former-employees-are-suing-google-over-alleged-gender-d-1810184079`

> »After the *New York Times* detailed the employee spreadsheets on Friday, Google spokesperson Gina Scigliano told Gizmodo that its own data shows, when you take "location, tenure, job role, level and performance" into account, that "women are paid 99.7% of what men are paid at Google." Scigliano described the *Times* story as "extremely flawed."«

→ hvad fortæller tallet 99.7 pct. os?

Eks. på 'kitchen sink' OLS: Putnam (2007)

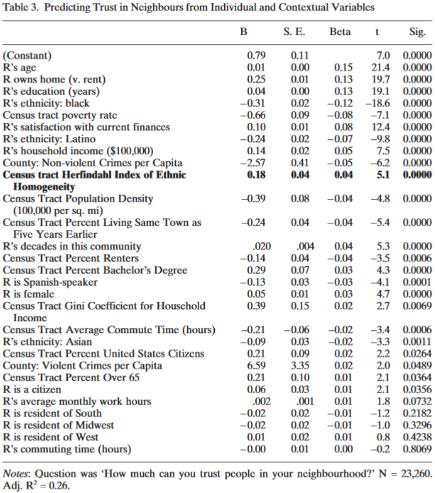


Table 3. Predicting Trust in Neighbours from Individual and Contextual Variables

	B	S. E.	Beta	t	Sig.
(Constant)	0.79	0.11		7.0	0.0000
R's age	0.01	0.00	0.15	21.4	0.0000
R owns home (v. rent)	0.25	0.01	0.13	19.7	0.0000
R's education (years)	0.04	0.00	0.13	19.1	0.0000
R's ethnicity: black	–0.31	0.02	–0.12	–18.6	0.0000
Census tract poverty rate	–0.66	0.09	–0.08	–7.1	0.0000
R's satisfaction with current finances	0.10	0.01	0.08	12.4	0.0000
R's ethnicity: Latino	–0.24	0.02	–0.07	–9.8	0.0000
R's household income ($100,000)	0.14	0.02	0.05	7.5	0.0000
County: Non-violent Crimes per Capita	–2.57	0.41	–0.05	–6.2	0.0000
Census tract Herfindahl Index of Ethnic Homogeneity	**0.18**	**0.04**	**0.04**	**5.1**	**0.0000**
Census Tract Population Density (100,000 per sq. mi)	–0.39	0.08	–0.04	–4.8	0.0000
Census Tract Percent Living Same Town as Five Years Earlier	–0.24	0.04	–0.04	–5.4	0.0000
R's decades in this community	.020	.004	0.04	5.3	0.0000
Census Tract Percent Renters	–0.14	0.04	–0.04	–3.5	0.0006
Census Tract Percent Bachelor's Degree	0.29	0.07	0.03	4.3	0.0000
R is Spanish-speaker	–0.13	0.03	–0.03	–4.1	0.0001
R is female	0.05	0.01	0.03	4.7	0.0000
Census Tract Gini Coefficient for Household Income	0.39	0.15	0.02	2.7	0.0069
Census Tract Average Commute Time (hours)	–0.21	–0.06	–0.02	–3.4	0.0006
R's ethnicity: Asian	–0.09	0.03	–0.02	–3.3	0.0011
Census Tract Percent United States Citizens	0.21	0.09	0.02	2.2	0.0264
County: Violent Crimes per Capita	6.59	3.35	0.02	2.0	0.0489
Census Tract Percent Over 65	0.21	0.10	0.01	2.1	0.0364
R is a citizen	0.06	0.03	0.01	2.1	0.0356
R's average monthly work hours	.002	.001	0.01	1.8	0.0732
R is resident of South	–0.02	0.02	–0.01	–1.2	0.2182
R is resident of Midwest	–0.02	0.02	–0.01	–1.0	0.3296
R is resident of West	0.01	0.02	0.01	0.8	0.4238
R's commuting time (hours)	–0.00	0.01	0.00	–0.2	0.8069

Notes: Question was 'How much can you trust people in your neighbourhood?' N = 23,260. Adj. $R^2 = 0.26$.

Den klassiske tilgang: masseproduktion af 'pseudo-general pseudo-facts'

»At the turn of the millennium, the modal quantitative research design was one in which researchers assembled data on theoretically interesting dependent and independent variables (...) Researchers then assessed the presumably causal relationships in these data using regressions with informally motivated sets of control variables to reduce the potential for confounding.«

Sidenhen: en 'credibility revolution' i samfundsvidenskaben

»This convention in quantitative causal research appears to be breaking down, and more quantitative causal research is moving toward causal empiricism. This (...) represents a major change in what researchers believe are credible ways of doing causal inference.«

Problemer i klassiske regressionstilgange:

1. mgl. ekstern validitet: nominel ctr. effektiv stikprøve
2. mgl. intern validitet: misspecifikation

Ad (1):

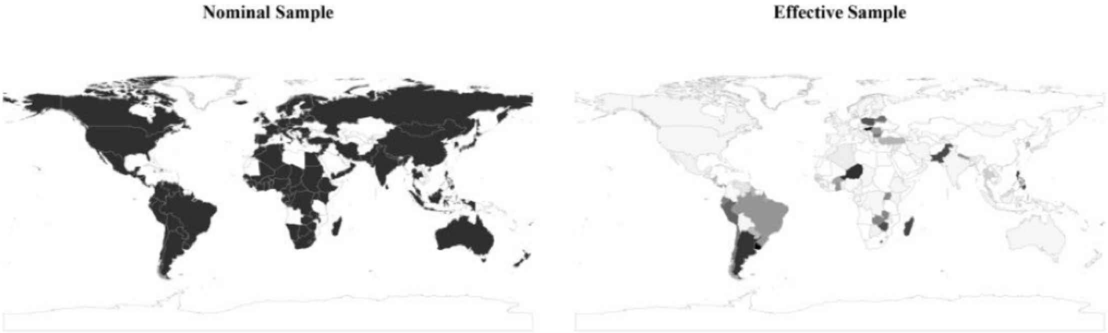


Figure 1. Nominal and effective samples from Jensen (2003), reproduced from Aronow and Samii (2016)

Ad (2):

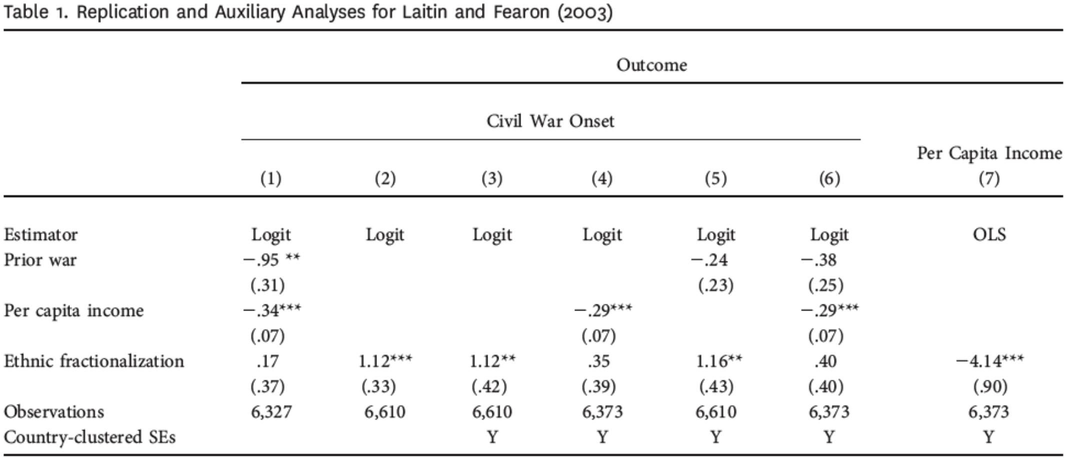


Table 1. Replication and Auxiliary Analyses for Laitin and Fearon (2003)

	Outcome						
	Civil War Onset						Per Capita Income
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Estimator	Logit	Logit	Logit	Logit	Logit	Logit	OLS
Prior war	−.95 **				−.24	−.38	
	(.31)				(.23)	(.25)	
Per capita income	−.34***			−.29***		−.29***	
	(.07)			(.07)		(.07)	
Ethnic fractionalization	.17	1.12***	1.12**	.35	1.16**	.40	−4.14***
	(.37)	(.33)	(.42)	(.39)	(.43)	(.40)	(.90)
Observations	6,327	6,610	6,610	6,373	6,610	6,373	6,373
Country-clustered SEs			Y	Y	Y	Y	Y

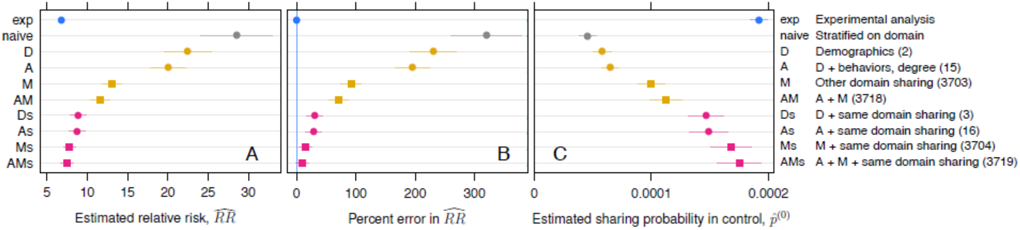


Figure 2: **Comparison of experimental and observational estimates of peer effects.** (A) The experiment estimates that users are 6.8 times as likely to share when exposed to a peer sharing, while the observational point estimates are larger. (B) Treating the experimental estimate as the truth, the naive observational estimate overestimates peer effects by 320%. This bias is substantially reduced by adjusting for prior same domain sharing (magenta) and prior sharing for 3,703 other domains (squares). (C) All discrepancies in the estimates of relative risk are due to underestimating $p^{(0)}$ when using observational data. Error bars are 95% confidence intervals. Brief descriptions of the estimators with number of covariates in parentheses are shown for reference.

Næste gang:

- matching
- Justesen & Klemmensen
- case: Nall
- specifik metode: *coarsened excat matching*
- ekstra lektie 1: hent `cem`-pakken
- ekstra lektie 2: læs afsnit 3 t.o.m. 3.1.2 i Iacus et al. (link på GH)

Tak for i dag!