6: Matching

Videregående kvantitative metoder i studiet af politisk adfærd

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

Institut for Statskundskab
Københavns Universitet

11. oktober 2017

- endogenitet
- 'bad controls' / post-treatment adjustment bias
- 'credibility-revolutionen'
- modelbaseret ctr. designbaseret inferens

Om kontrol for outcomevariable ctr. 'late' proxyvariable:

»There is an interesting ambiguity in the proxy-control story that is not present in the first bad-control story. Control for outcome variables is simply misguided; you do not want to control for occupation in a schooling regression if the regression is to have a causal interpretation. In the proxy-control scenario, however, your intentions are good. And while proxy control does not generate the regression coefficient of interest, it may be an improvement on no control at all.« (68)

Danske kilder til surveydata

- Valgprojektet
- Dansk Data Arkiv

Internationale kilder til surveydata

- Afrobarometer
- American National Election Studies
- Americas Barometer
- Arab Barometer
- Asian Barometer
- British Election Study
- British Social Attitudes
- Caucasus Barometer
- Comparative Study of Electoral Systems
- Election Studies Eastern Europe

Internationale kilder til surveydata II

- Eurobarometer
- European Election Studies: Voter Study
- European Social Survey
- European Values Study
- German General Social Survey
- International Social Survey Programme
- Latinobarómetro
- Pew Global Attitudes Survey
- Swedish National Election Studies
- World Values Survey

Internationale kilder til surveydata III

- American National Election Studies
- British Household Panel Survey
- German Socio-Economic Panel
- Longitudinal Internet Studies for the Social sciences
- Swiss Household Panel
- Understanding Society

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)

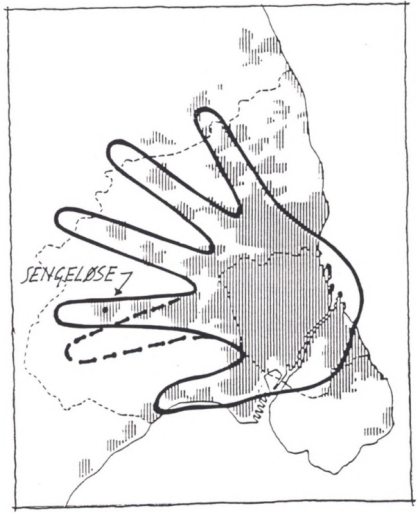
Fagets opbygning

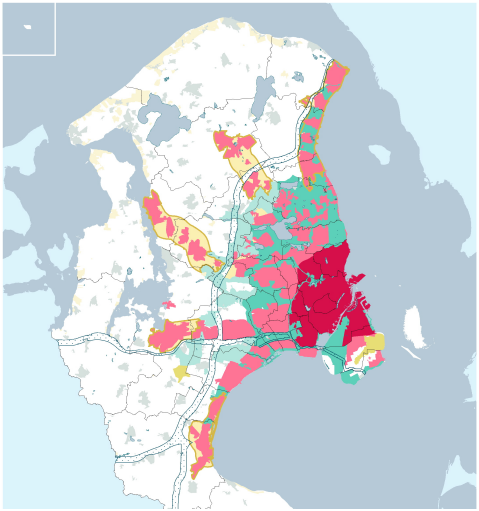
Blok 3

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)

Politiske spillovers af suburbanisering

Suburbanisering i København:





Lad os antage en påvirket gruppe ($D_i = 1$) og en upåvirket gruppe ($D_i = 0$). Vi definerer nu *for hvert individ i*:

$$\text{Potential ouctome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \tag{1}$$

$\Longrightarrow$

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i \tag{2}$$

$\rightarrow$ Y_i antager altid værdien Y_{0i} *eller* Y_{1i}

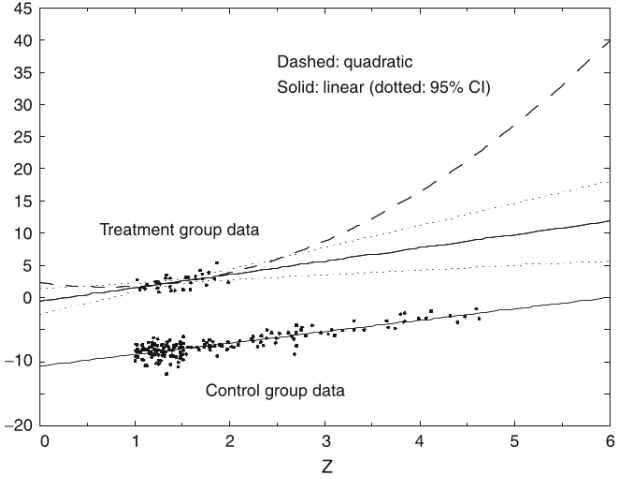
For hvert individ i kan vi definere effekten af treatment::

$$\delta_i = Y_{1i} - Y_{0i} \tag{3}$$

M.a.o.: *effekten* δ_i *er forskellen mellem* Y_i *når* $D_i = 1$ *og* Y_i *når* $D_i = 0$

- kaldes også 'Rubin causal model' efter Donald B. Rubin
- problem: T er altid *enten* 1 *eller* 0
- $\rightarrow$ vi observerer altid kun Y_{i1} *eller* Y_{i1}
- $\rightarrow$ vi kan aldrig observere δ_i
- denne uobserverbarhed kaldes the fundamental problem of causal inference
- POF er central byggesten i rationalet bag eksperimenter (jf. gang 7+8)

Problem i OLS-tilgange: effektestimater beror (potentielt) på interpolation/ekstrapolation
$\rightarrow$ *model dependence*

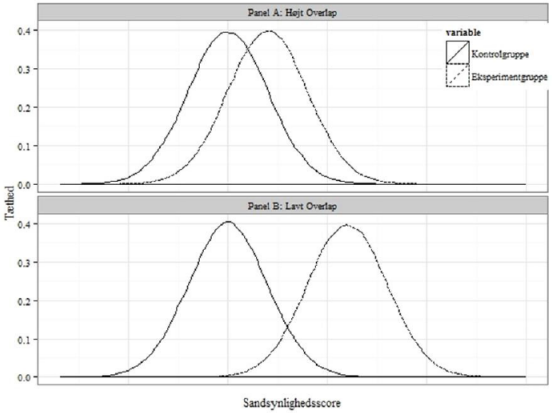


Throwback til Samii (2016):

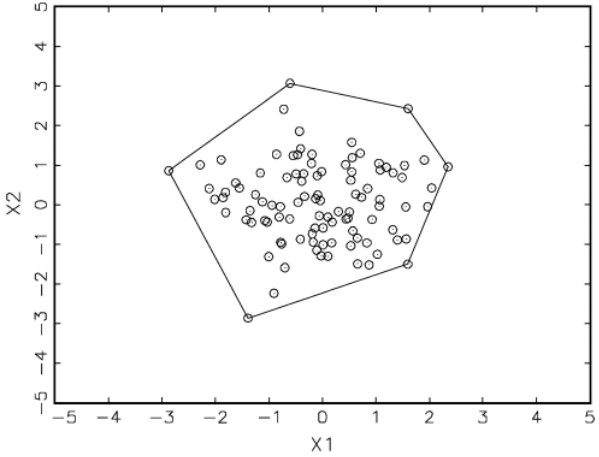
»Where there is no overlap, one can only make comparisons with interpolated or extrapolated counterfactual potential outcomes values. (...) Among those who take positivity and overlap seriously, the common reaction (...) has been to resort to other estimation methods like matching estimators. Matching estimation forces the researcher immediately to confront the reality of limited overlap.« (944)

Matching fokuserer på dele af data med *common support*

Figur 1: Intervaller i data med højt og lavt overlap



I N-dimensionelle data defineres support ud fra datas 'konvekse hylster' (*convex hull*)



Klassisk tilgang i matching: *exact matching*

Men: ofte ikke tilstrækkeligt mange eksakte matches $\rightarrow$ behov for *dimensionalitetsreduktion*

Traditionel løsning: *propensity score matching* (PSM) $\rightarrow$ hver obs. estimeres ssh for treatment

Efter PSM-estimation: fx. *nearest neighbor matching*

Alternativ: *radius* (caliper) matching

Sidenhen skepsis om PSM's fortræffeligheder, jf. fx. King & Nielsen (2016), "Why Propensity Scores Should Not Be Used for Matching" (`https://youtu.be/rBv39pK1iEs`)

Centralt problem ved PSM: misspecifikation i PS-modellen kan give bias (og er inefficient selv ved korrekt specifikation)

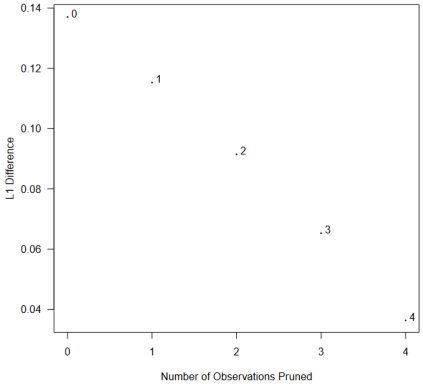
Alternativ: *Coarsened Exact Matching* $\rightarrow$ variable opdeles i grove, *meningsfulde* kategorier og matches eksakt iht. kategori

I CEM måles balance med $\mathcal{L}_1$:

$$\mathcal{L}_1 = \frac{1}{2} \sum_{\ell_1 \ldots \ell_k} |f_{\ell_1 \ldots \ell_k} - g_{\ell_1 \ldots \ell_k}| \tag{4}$$

- hvor f og g er celleandele for hhv. treatede og kontrolenheder
- fuld common support $\rightarrow \mathcal{L}_1 = 0$
- ingen common support $\rightarrow \mathcal{L}_1 = 1$

Nødvendigt tradeoff i matching: balance vs. sample size



Kilde: King, Gary, Christopher Lucas, and Richard Nielsen. 2015. "The Balance-Sample Size Frontier in Matching Methods for Causal Inference."

$\rightarrow$ eksempel på *bias-variance tradeoff*

Fremgangsmåde med CEM:

1. evaluér balance med `imbalance()` (ekskl. treatment og outcome)
2. definér kategorier for matching-variable
3. kør CEM med `cem()` m. information fra (2)
4. lav nyt, 'pruned' datasæt kun m. matchede data
5. evaluér balance igen
6. hvis tilfredsstillende balance, estimér effekt m. pruned data

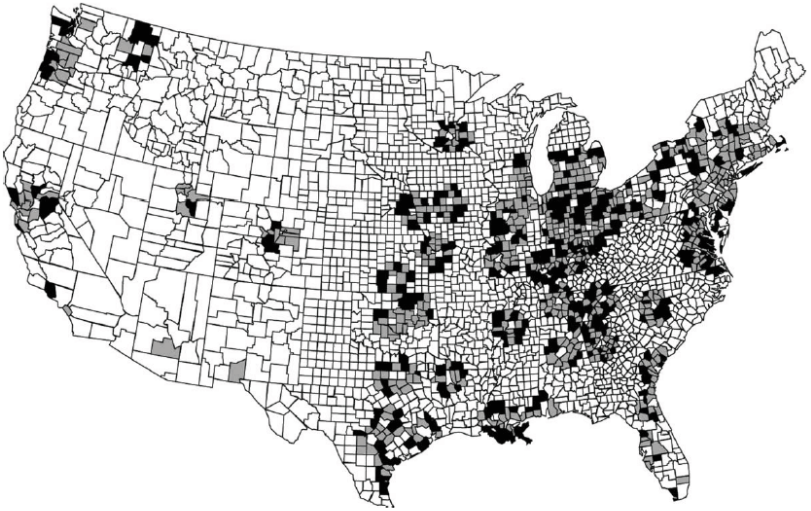


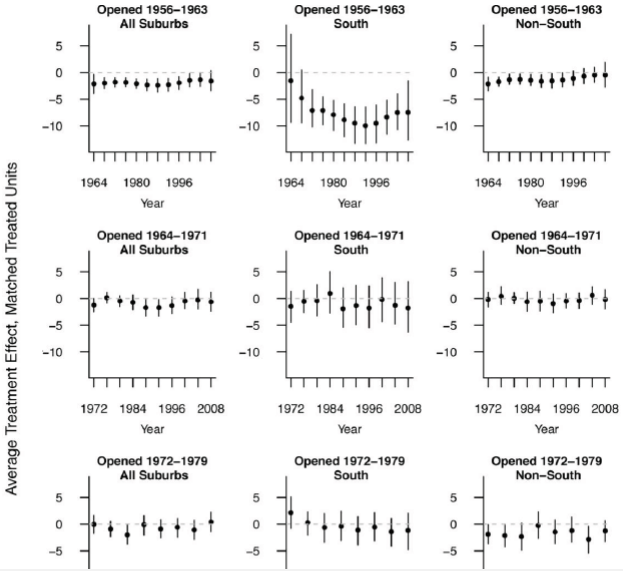
Figure 2. Map of the full suburban county sample. Counties containing an Interstate highway through 1996 are lightly shaded, and those without are shaded black.

Explicit kobling til POF:

»[This article] examines the politics of places where Interstates were built, compared to a counterfactual constructed from comparable units that had no (or fewer) highways. It is assumed throughout that Interstate highways were assigned as if randomly to places, conditional on inclusion of relevant observable confounders.« (396)

»The Interstate program fits well with this research design: a well-documented plan was adopted before construction, enabling one to reconstruct, and control for, the factors leading to nonrandom highway placement. Key planning criteria appear in the 1944 *Interregional Highways* report, which laid out an early version of the present-day Interstate System (United States, Public Roads Administration, 1944). From the postwar period to the late 1960s, highway engineers had substantial latitude to select highway routes, using well-documented technical criteria, most of which appeared in the 1944 report. This differs substantially from present-day roadbuilding and its ad hoc projects as commonly studied in the distributive politics literature.« (396)

→ hvordan påvirker dette designets troværdighed?


Opened 1956–1963
All Suburbs
Opened 1956–1963
South
Opened 1956–1963
Non–South
Opened 1964–1971
All Suburbs
Opened 1964–1971
South
Opened 1964–1971
Non–South
Opened 1972–1979
All Suburbs
Opened 1972–1979
South
Opened 1972–1979
Non–South
Average Treatment Effect, Matched Treated Units
5
0
−5
−10
1964 1980 1996
1972 1984 1996 2008
Year

Næste gang:

- efterårsferie!

Tak for i dag!