42

5: Regression I: OLS

Videregående kvantitative metoder i studiet af politisk adfærd

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

Institut for Statskundskab
Københavns Universitet

10. oktober 2018

1 Opsamling

2 Motivation

3 OLS

4 Implementering i R

5 Mutz (2018)

6 Kig fremad

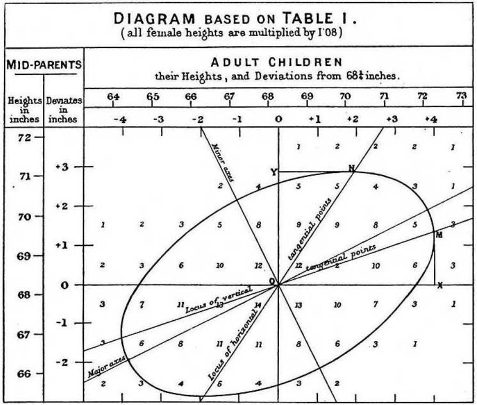
Sidste gang:

- konceptuelt om text as data
- klassifikation I: tf-idf
- klassifikation II: dictionary-metoder
- udestående: skalering m. wordscores

- fra i dag, temaskift: hvilke data kan vi få? → *hvad kan vi lære af data?*
- i dag: kausal inferens m. tværsnitsdata
- næste gang (dvs. i morgen): kausal inferens m. paneldata
- efter efterårsferien: workshop II → input hertil meget velkomne!

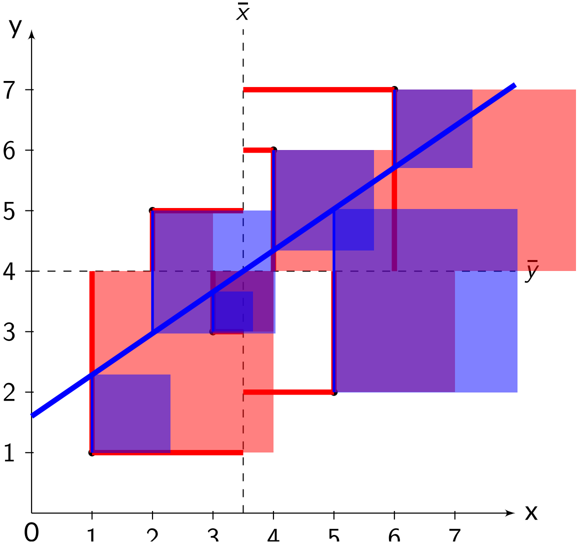
Kan en privat uddannelse betale sig?

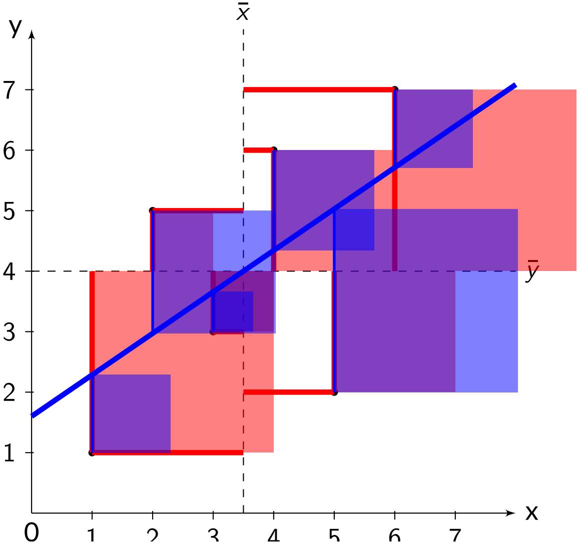
Galton, F. (1886). “Regression towards mediocrity in hereditary stature”. *The Journal of the Anthropological Institute of Great Britain and Ireland*. 15: 246–263



Baggrund

FROM THE DIRECTOR OF THE OTHERS
FEAR ALWAYS FINDS ITS VICTIM
ETHAN HAWKE
EMMA WATSON
REGRESSION
A FILM BY ALEJANDRO AMENÁBAR
COMING SOON


x̄
ȳ
y
x
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7


y
x
$\bar{x}$
$\bar{y}$
0
1
2
3
4
5
6
7
1
2
3
4
5
6
7

- Total Sum of Squares (TSS): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- TSS består af to dele:
  - Explained Sum of Squares (ESS)
  - Residual Sum of Squares (RSS)
- $TSS = ESS + RSS$
- OLS estimerer den linje der minimerer RSS
- centralt her: under de rette forudsætninger har smh. ml. X og Y en kausal fortolkning!

Regressionsmodel med af outcome Y_i treatment-variabel P_i og kontrolvariabel A_i:

$$Y_i = \alpha + \beta P_i + \gamma A_i + e_i \tag{1}$$

Alternativ notation: CEF (Conditional Expectation Function)

$$E[Y_i | P_i, A_i] \tag{2}$$

Koefficienter kan udtrykkes som forskelle mellem CE's:

$$E[Y_i | P_i = 1, A_i] - E[Y_i | P_i = 0, A_i] = \beta \tag{3}$$

Den fittede Y_i, $\widehat{Y}_i$, omfatter ikke fejlleddet:

$$\widehat{Y}_i = \alpha + \beta P_i + \gamma A_i \tag{4}$$

Dermed:

$$e_i = Y_i - \widehat{Y}_i = Y_i - \alpha + \beta P_i + \gamma A_i \tag{5}$$

Hvad forklarer e_i?

- udeladte variable (omitted variables)
- målefejl
- fundamental tilfældig variation (MM: ‘serendipitous variation’)

Kontroller kan også være kategoriske (fx. specifikke kombinationer af skoler) eller intervalskalerede (fx. SAT) (jf. s. 61)

$$ln(Y_i) = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 lnPI_i + e_i \qquad (6)$$

Standardfejlen for β:

$$SE(\beta) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_\beta} \tag{7}$$

Implikation: små fejlled (= præcise estimater) kræver

- ↓ σ_e og/eller
- ↑ n og/eller
- ↑ σ_β

Kort vs. lang form:

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l \tag{8}$$

$$Y_i = \alpha^s + \beta^s P_i + e_i^s \tag{9}$$

$\rightarrow$ hvor forskellige er β^l og β^l?

$$\beta^s - \beta^l = \pi_1 \times \gamma \tag{10}$$

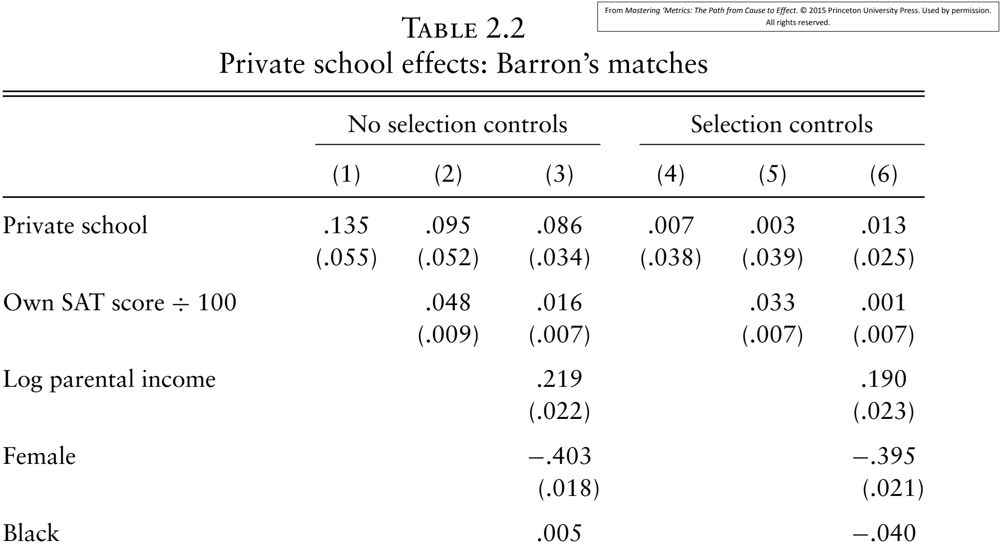
hvor π_1 er koefficienten af P_i på A_i:

$$A_i = \pi_0 + \pi_1 P_i + u_i \tag{11}$$

Når vi har kontrolleret for alle confounders:

- $\rightarrow$ residualet ukorreleret med P_i og X_i
- $\rightarrow$ koefficienten på P_i har en kausal fortolkning
- a.k.a. 'selection-on-observables' antagelsen

Output



From *Mastering 'Metrics: The Path from Cause to Effect*. © 2015 Princeton University Press. Used by permission. All rights reserved.

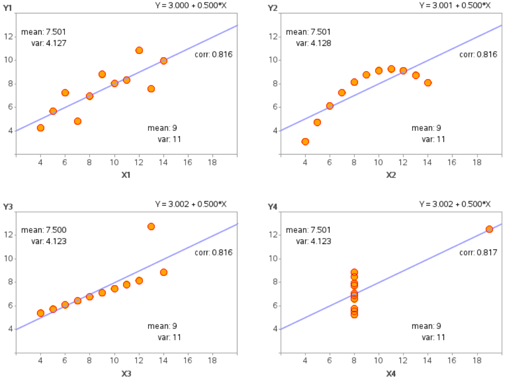
TABLE 2.2
Private school effects: Barron's matches

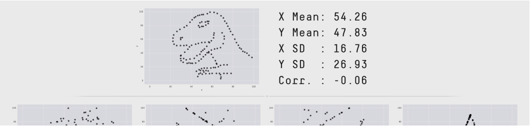
	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040

Typiske faldgruber v. regression:

1. omitted variable bias (jf. ovenfor)
2. kontrol for post-treatment / ‘bad controls’ (mere herom i uge 8)
3. outliers
4. multikollinearitet
5. ikke-lineær funktionel form

Ad 3-5: jf. *Anscombe's Quartet*



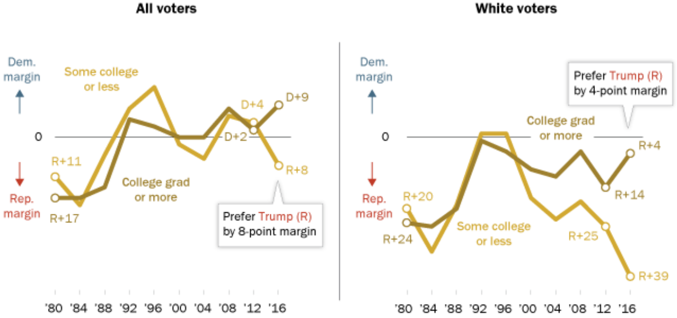


```
ols <- lm(y~x+z,data=df)
```

https://eu.freep.com/story/news/local/michigan/2017/05/28/michigan-donald-trump-voters/344246001/

Wide education gaps in 2016 preferences, among all voters and among whites

Presidential candidate preference, by educational attainment



Source: Based on exit polls conducted by Edison Research for the National Election Pool, as reported by CNN. Data from prior years from national exit polls. In 1980, race was coded by the interviewer instead of being asked of the respondent.

PEW RESEARCH CENTER

→ hvad forklarer den stærke smh. ml. uddannelse og Trump-støtte?

Mutz: betydningen af ‘status threat’ (ctr. ‘left behind’-tesen)

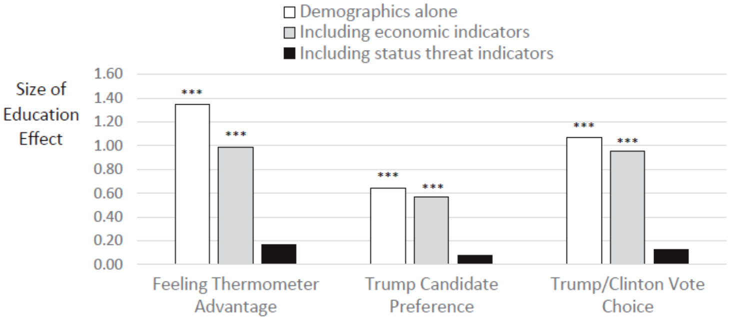


Fig. 3. Status threat accounts for the impact of education on the 2016 presidential election. Note that bars represent the predictive strength of education on each of three different outcome measures after taking into account (*i*) demographics alone, (*ii*) demographics and economic predictors only, and (*iii*) demographics and threat indicators only. Details are in Table S5. ****P* < 0.001.

> »regardless of which outcome measures I examined, including indicators of economic status did not eliminate the impact of education. (...) However, after the relationship between Trump support and perceived status threat is taken into account, even lack of a college education no longer predicts Trump support for any of the measures. These findings strongly suggest that group-based status threat was the main reason that those without college educations were more supportive of Trump.« (8)

> »these results speak to the importance of group status in the formation of political preferences. Political uprisings are often about downtrodden groups rising up to assert their right to better treatment (...). The 2016 election, in contrast, was an effort by members of already dominant groups to assure their continued dominance (...)« (9)

Tabel S5 i SI:

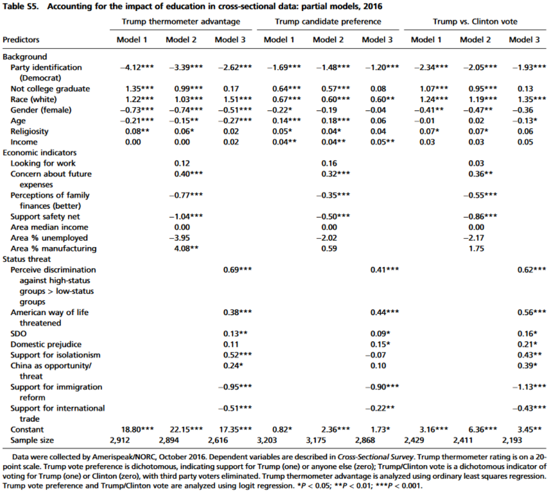


Table S5. Accounting for the impact of education in cross-sectional data: partial models, 2016

	Trump thermometer advantage			Trump candidate preference			Trump vs. Clinton vote		
Predictors	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Background									
Party identification (Democrat)	−4.12***	−3.39***	−2.62***	−1.69***	−1.48***	−1.20***	−2.34***	−2.05***	−1.93***
Not college graduate	1.35***	0.99***	0.17	0.64***	0.57***	0.08	1.07***	0.95***	0.13
Race (white)	1.22***	1.03***	1.51***	0.67***	0.60***	0.60**	1.24***	1.19***	1.35***
Gender (female)	−0.73***	−0.74***	−0.51***	−0.22*	−0.19	−0.04	−0.41**	−0.47**	−0.36
Age	−0.21***	−0.15**	−0.27***	0.14***	0.18***	0.06	−0.01	0.02	−0.13*
Religiosity	0.08**	0.06*	0.02	0.05*	0.04*	0.04	0.07*	0.07*	0.06
Income	0.00	0.00	0.02	0.04**	0.04**	0.05**	0.03	0.03	0.05
Economic indicators									
Looking for work		0.12			0.16			0.03	
Concern about future expenses		0.40***			0.32***			0.36**	
Perceptions of family finances (better)		−0.77***			−0.35***			−0.55***	
Support safety net		−1.04***			−0.50***			−0.86***	
Area median income		0.00			0.00			0.00	
Area % unemployed		−3.95			−2.02			−2.17	
Area % manufacturing		4.08**			0.59			1.75	
Status threat									
Perceive discrimination against high-status groups > low-status groups			0.69***			0.41***			0.62***
American way of life threatened			0.38***			0.44***			0.56***
SDO			0.13**			0.09*			0.16*
Domestic prejudice			0.11			0.15*			0.21*
Support for isolationism			0.52***			−0.07			0.43**
China as opportunity/threat			0.24*			0.10			0.39*
Support for immigration reform			−0.95***			−0.90***			−1.13***
Support for international trade			−0.51***			−0.22**			−0.43***
Constant	18.80***	22.15***	17.35***	0.82*	2.36***	1.73*	3.16***	6.36***	3.45**
Sample size	2,912	2,894	2,616	3,203	3,175	2,868	2,429	2,411	2,193

Data were collected by Amerispeak/NORC, October 2016. Dependent variables are described in *Cross-Sectional Survey*. Trump thermometer rating is on a 20-point scale. Trump vote preference is dichotomous, indicating support for Trump (one) or anyone else (zero); Trump/Clinton vote is a dichotomous indicator of voting for Trump (one) or Clinton (zero), with third party voters eliminated. Trump thermometer advantage is analyzed using ordinary least squares regression. Trump vote preference and Trump/Clinton vote are analyzed using logit regression. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Næste gang:

- regression II: paneldata
- læs AGS 3.1+3.2+3.6.1 (datastruktur og OVB)
- læs AGS 4 t.o.m. 4.1.2.1 (FE-modeller)
- øvelse: genskab Mutz' overordnede resultat som vist i figur 3, jf. script

Tak for i dag!