3: 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

20. september 2017

1 Opsamling

2 Motivation

3 OLS

4 Implementering i R

5 Solt (2017)

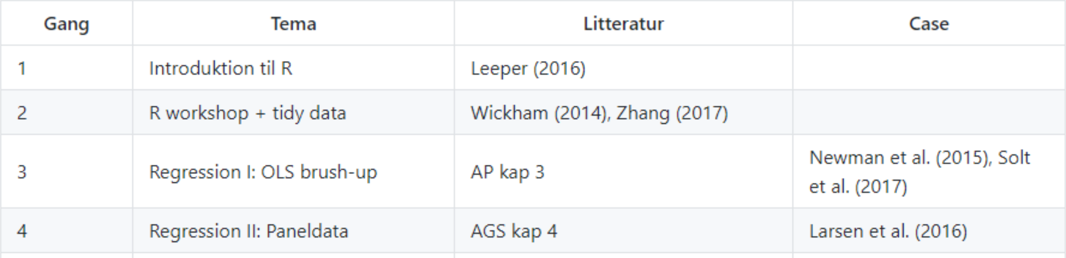
6 Kig fremad

Sidste gang:

- data frames
- kriterier for tidy data
- de fire verber i databehandling
- piping

Fagets opbygning

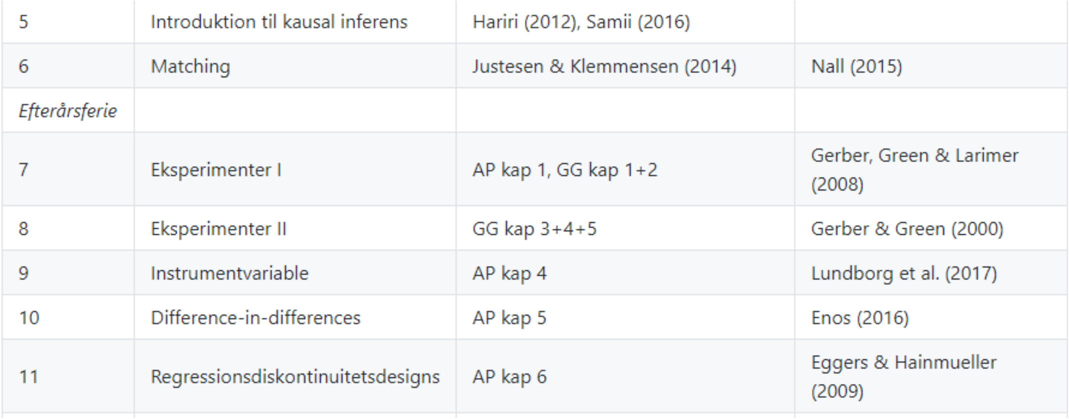
Blok 1



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

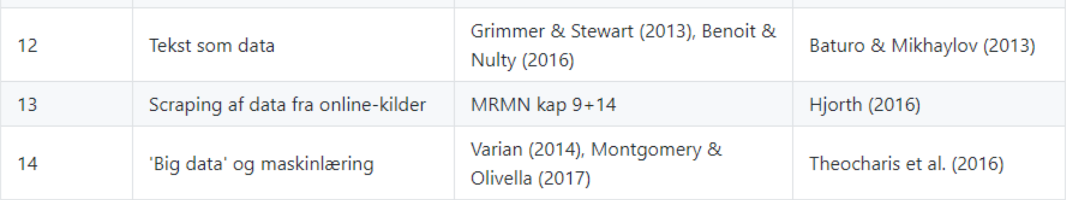
Blok 2



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)

Motivation: Newman om konsekvenser af synlig, lokal ulighed

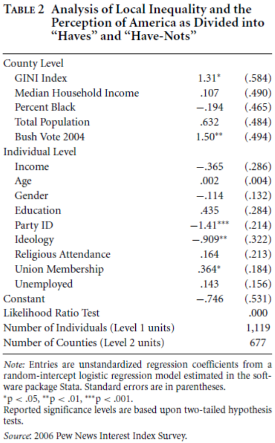


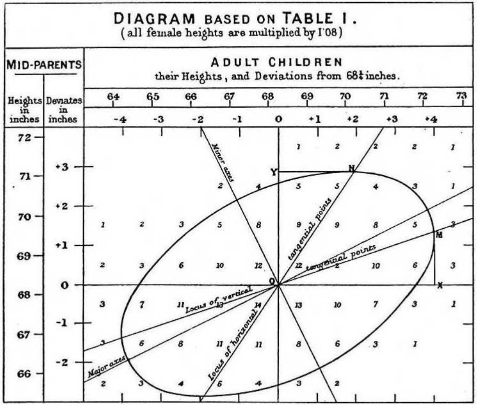
Table 2 Analysis of Local Inequality and the Perception of America as Divided into "Haves" and "Have-Nots"

County Level		
GINI Index	1.31*	(.584)
Median Household Income	.107	(.490)
Percent Black	−.194	(.465)
Total Population	.632	(.484)
Bush Vote 2004	1.50**	(.494)
Individual Level		
Income	−.365	(.286)
Age	.002	(.004)
Gender	−.114	(.132)
Education	.435	(.284)
Party ID	−1.41***	(.214)
Ideology	−.909**	(.322)
Religious Attendance	.164	(.213)
Union Membership	.364*	(.184)
Unemployed	.143	(.156)
Constant	−.746	(.531)
Likelihood Ratio Test		.000
Number of Individuals (Level 1 units)		1,119
Number of Counties (Level 2 units)		677

Note: Entries are unstandardized regression coefficients from a random-intercept logistic regression model estimated in the software package Stata. Standard errors are in parentheses.
*p < .05, **p < .01, ***p < .001.
Reported significance levels are based upon two-tailed hypothesis tests.

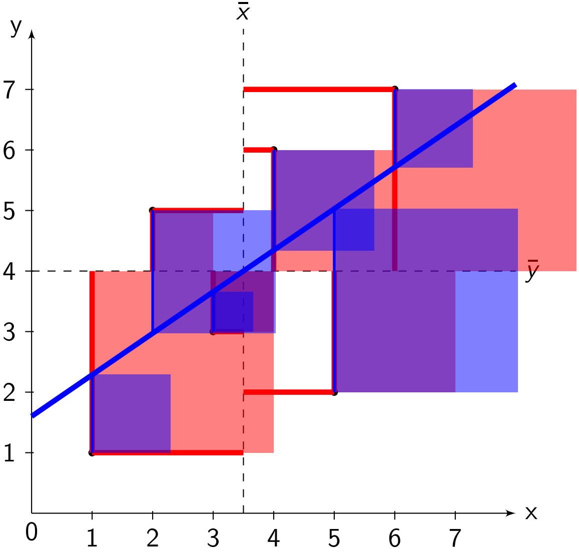
Source: 2006 Pew News Interest Index Survey.

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

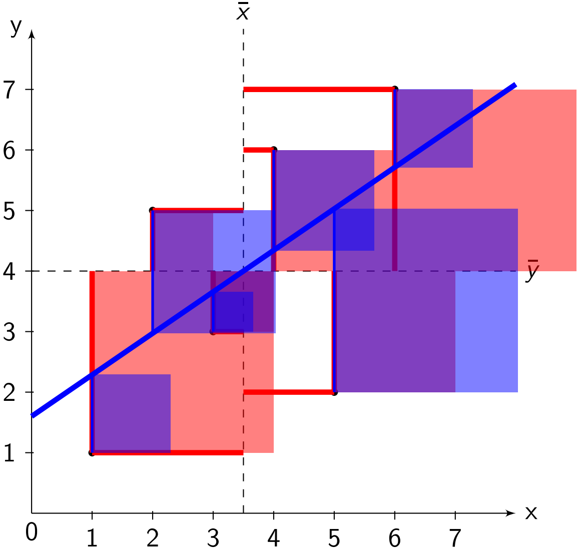


Baggrund

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


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

- Total Sum of Squares (SST): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- SST består af to dele:
  - Explained Sum of Squares (SSE)
  - Residual Sum of Squares (SSR)
- $SST = SSE + SSR$
- OLS estimerer den linje der minimerer SSR


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

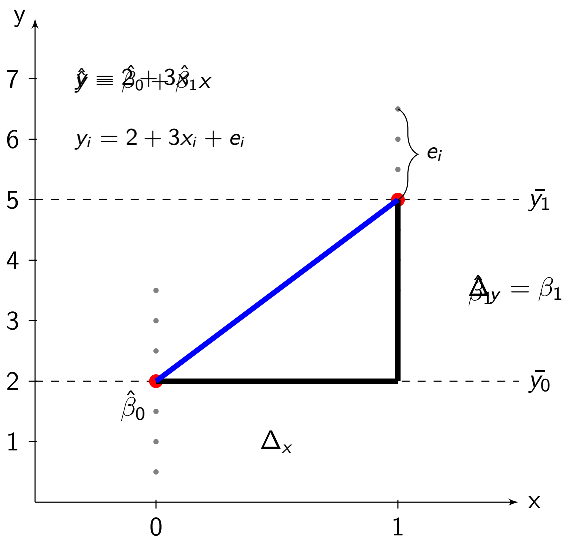
$$\beta = \text{arg min } E[(Y_i - X_i'b)^2] \tag{1}$$

OLS-estimatoren:

$$\beta = E[X_iX_i']^{-1}E[X_iY_i] \tag{2}$$

I den bivariate case:

$$\beta = \frac{Cov(Y_i, X_i)}{Var(X_i)} \tag{3}$$


y
7
6
5
4
3
2
1
$\hat{y} = 2 + 3x$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$y_i = 2 + 3x_i + e_i$
e_i
$\bar{y}_1$
$\Delta_y = \beta_1$
$\hat{\beta}_1$
$\bar{y}_0$
$\hat{\beta}_0$
Δ_x
x
0
1

OLS giver os den bedste lineære approksimation af CEF (Conditional Expectation Function)

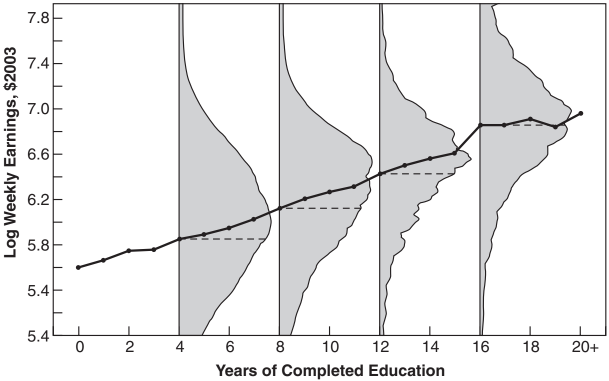


Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

From *Mostly Harmless Econometrics: An Empiricist's Companion*. © 2009 Princeton University Press.

Regressionsmodel med treatment-variabel s_i og kontrolvariabel X_i:

$$Y_i = \alpha + \beta s_i + \gamma X_i + e_i \tag{4}$$

Alternativ notation: CEF

$$E[Y_i | s_i, X_i] \tag{5}$$

Koefficienter kan udtrykkes som forskelle mellem CE's:

$$E[Y_i | X_i, s_i = s] - E[Y_i | X_i, s_i = s - 1] = \beta \tag{6}$$

Når vi har 'conditional independence' er potentielle outcomes for Y_i uafhængige af s_i betinget på X_i (CIA):

$$Y_{si} \perp\!\!\!\perp s_i | X_i, \text{ for all } s \tag{7}$$

- → når CIA er opfyldt er residualet ukorreleret med s_i og X_i
- → koefficienten på s_i har en kausal fortolkning
- a.k.a. 'selection-on-observables' antagelsen

Kort vs. lang form:

$$Y_i = \alpha^l + \rho^l s_i + A_i\gamma + e_i^l \tag{8}$$

$$Y_i = \alpha^s + \rho^s s_i + e_i^s \tag{9}$$

→ hvor forskellige er ρ^l og ρ^s?

$$\rho^s - \rho^l = \gamma \times \delta_{As} \tag{10}$$

hvor δ_{As} er koefficienten af s_i på A_i:

$$A_i = \alpha + \delta s_i + u_i \tag{11}$$

→ OVB er en funktion af udeladte variables korrelation med den uafhængige *og* den afhængige

OVB i human kapital-modeller:

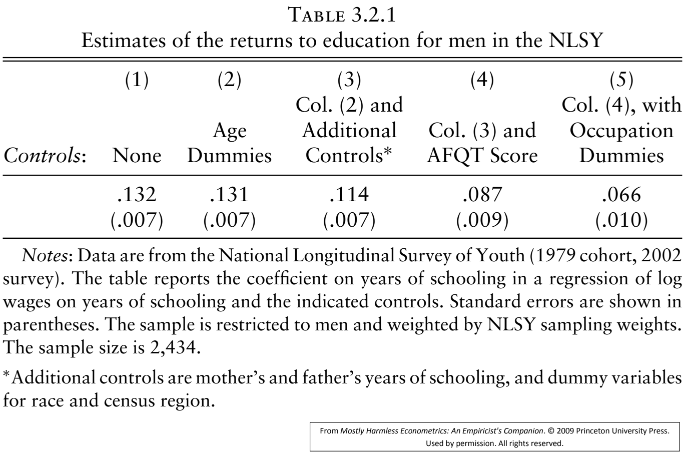


TABLE 3.2.1
Estimates of the returns to education for men in the NLSY

	(1)	(2)	(3)	(4)	(5)
			Col. (2) and		Col. (4), with
		Age	Additional	Col. (3) and	Occupation
Controls:	None	Dummies	Controls*	AFQT Score	Dummies
	.132	.131	.114	.087	.066
	(.007)	(.007)	(.007)	(.009)	(.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey). The table reports the coefficient on years of schooling in a regression of log wages on years of schooling and the indicated controls. Standard errors are shown in parentheses. The sample is restricted to men and weighted by NLSY sampling weights. The sample size is 2,434.

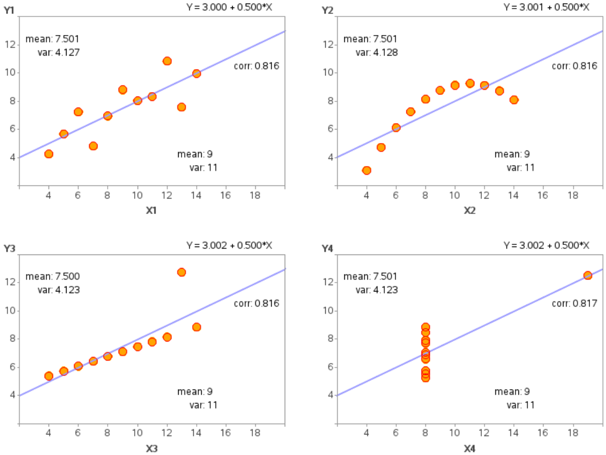
*Additional controls are mother's and father's years of schooling, and dummy variables for race and census region.

From *Mostly Harmless Econometrics: An Empiricist's Companion*. © 2009 Princeton University Press. Used by permission. All rights reserved.

Typiske faldgruber v. regression:

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

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



→ kig altid på data først!

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

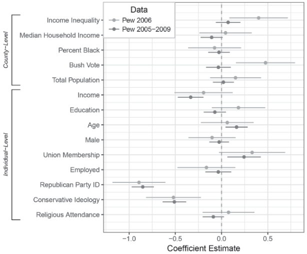


Figure 1. Local inequality and the perception of America as divided into "haves" and "have-nots": results using all available data. The dots represent the estimated change in the logged odds of believing the United States to be divided into "haves" and "have-nots" for a change of two standard deviations in the independent variable; the whiskers represent the 95% confidence intervals of these estimates. The statistically significant result for county income inequality in the 2006 survey presented in table 2 of Newman et al. (2015) is not evident when all of the available data are examined.

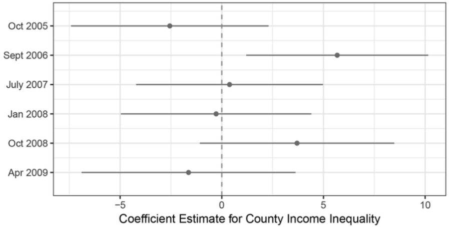


Figure 2. Local inequality and the perception of America as divided into "haves" and "have-nots": results using each available data set. Dots represent the estimated change in the logged odds of believing the United States to be divided into "haves" and "have-nots" for a change of two standard deviations in county income inequality; whiskers represent 95% confidence intervals. The only one of the six available surveys conducted in the time period Newman et al. (2015) examines that yields a statistically significant result is the 2006 survey that article presents.

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)

Tak for i dag!