2: Regression I: OLS, interaktioner

Videregående kvantitative metoder i studiet af politisk adfærd

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

Institut for Statskundskab
Københavns Universitet

12. september 2016

- To undervisningsgange flyttes pga. anden undervisning:
  - Gang 7 (24/10) finder sted kl. 16-18 i **lokale 2.2.42**.
  - Gang 4 finder sted **29/9 kl. 12-14**, lokale tbd.
- Frivillig R-workshop tirsdag d. 11. oktober
- Midterm er bestået/ikke bestået
- Udleveres tirsdag d. 11. oktober kl. 18, frist 7 dage

Uge	Dato	Tema	Litteratur	Case
1	5/9	Introduktion til R	Imai kap 1	
2	12/9	Regression I: OLS, interaktioner	GH kap 3, MM kap 2	Gilens & Page (2014)
3	26/9	Regression II: Panelmodeller	GH kap 11	Larsen et al. (2016)
4	29/9	Regression III: Multilevelmodeller	GH kap 12	Dinesen & Sønderskov (2012)
5	3/10	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	
6	10/10	Matching	Justesen & Klemmensen (2014)	Ladd & Lenz (2009)
	17/10	*Efterårsferie*		

Uge	Dato	Tema	Litteratur	Case
	17/10	*Efterårsferie*		
7	24/10	Eksperimenter I	MM kap 1, GG kap 1+2	Bond et al. (2012)
8	31/10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	7/11	Instrumentvariable	MM kap 3	Arunachalam & Watso
10	14/11	Regressionsdiskontinuitetsdesigns	MM kap 4	Eggers & Hainmueller
11	21/11	Difference-in-difference designs	MM kap 5	Enos (2016)
12	28/11	'Big data' og maskinlæring	Grimmer (2015), Varian (2014)	
13	5/12	Scraping af data fra online-kilder	MRMN kap 9	
14	12/12	Tekst som data	Grimmer & Stewart (2013), Imai kap 5	

Spørgsmål?

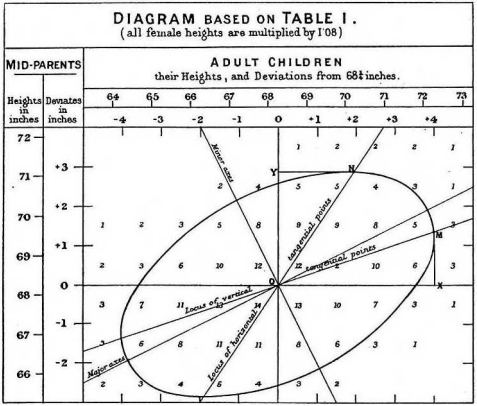
- Imai: 2 revolutioner i kvantitativ samfundsvidenskab
- fordele og ulemper ved R
- objekter
- vektorer
- funktioner
- data frames
- import/export

Opsamles senere i dag:

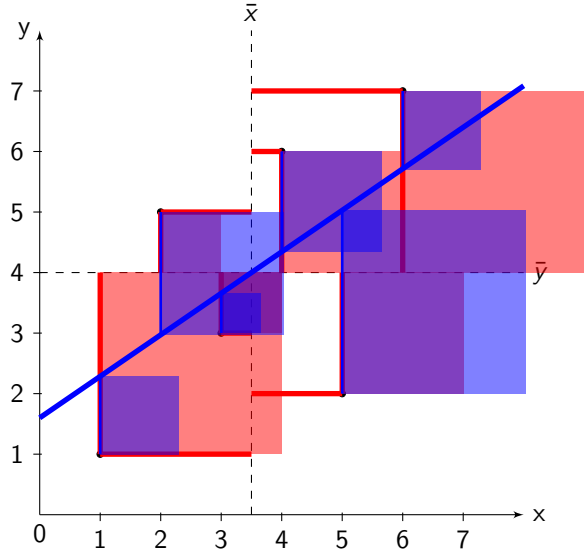
- subsetting af data frames

Ezra Klein om Gilens & Page

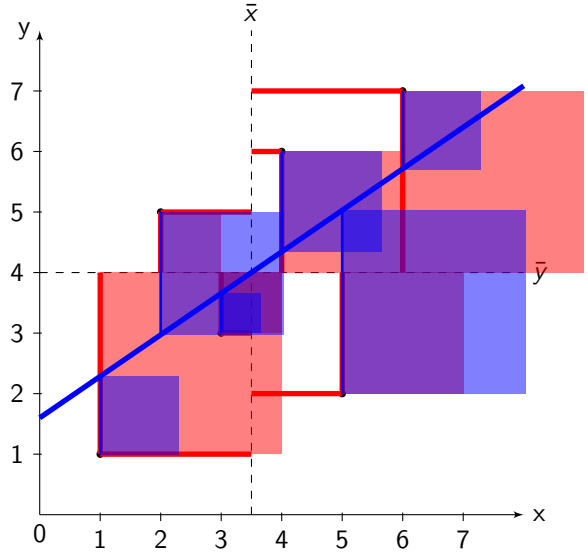
Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland*. 15: 246–263

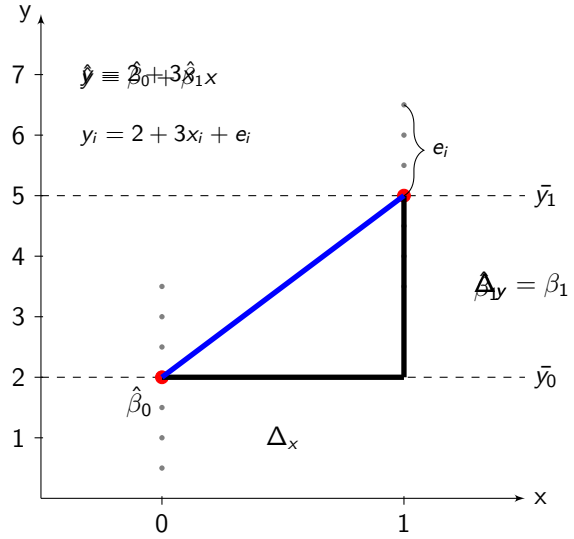



FROM THE DIRECTOR OF THE OTHERS
FEAR ALWAYS FINDS ITS VICTIM
ETHAN HAWKE
EMMA WATSON
REGRESSION
A FILM BY ALEJANDRO AMENÁBAR
COMING SOON


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

- 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


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


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

Regressionsmodel med 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)

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

Standard fejl i model med K koefficienter:

$$Y_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ki} + \gamma A_i + e_i \tag{7}$$

$$SE(\widehat{\beta_k}) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\tilde{X}_k}} \tag{8}$$

Implikation: små fejlled (det er godt!) kræver

- $\downarrow \sigma_e$ og/eller
- $\uparrow n$ og/eller
- $\uparrow \tilde{X}_k$

Kort vs. lang form:

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l \tag{9}$$

$$Y_i = \alpha^s + \beta^s P_i + e_i^s \tag{10}$$

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

$$\beta^s - \beta^l = \pi_1 \times \gamma \tag{11}$$

hvor π_1 er koefficienten af P_i på A_i:

$$A_i = \pi_0 + \pi_1 P_i + u_i \tag{12}$$

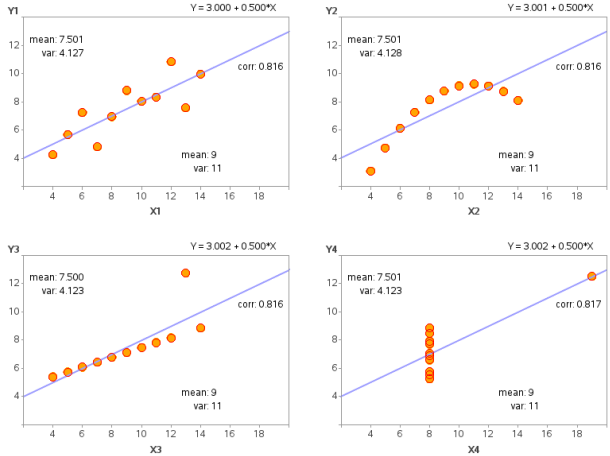
TABLE 2.2
Private school effects: Barron's matches

	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)

Typiske faldgruber v. regression:

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

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



$\rightarrow$ kig altid på data først!

Fire demokratiteoretiske traditioner:

1. Majoritarian Electoral Democracy
2. Economic-Elite Domination
3. Majoritarian Pluralism
4. Biased Pluralism

Table 1
Theoretical predictions concerning the independent influence of sets of actors upon policy outcomes

	Sets of Actors				
Theory (ideal type)	**Average Citizens**	**Economic Elites**	**All Interest Groups**	**Mass Interest Groups**	**Business Interest Groups**
Majoritarian Electoral Democracy	Y	n	n	n	n
Dominance by Economic Elites	y	Y	y	n	y
Majoritarian Pluralism	y	n	Y	Y	Y
Biased Pluralism	n	n	y	y	Y

n = little or no independent influence

y = some independent influence

Y = substantial independent influence

Næste gang:

- regression II: paneldata
- opsamling på interaktioner (GH kap 3)
- GH kap 11: fokus på non-nestede strukturer (afsnit 11.3)
- case: helt nyt working paper - fokus på aggregate-level analyse

Tak for i dag!