4: Regression III: Multileveldata, interaktioner

Videregående kvantitative metoder i studiet af politisk adfærd

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

Institut for Statskundskab
Københavns Universitet

29. september 2016

- permanent lokaleændring: faste holdtimer nu i **lokale 1.0.10**
- frivillig R-workshop mandag d. 10. oktober 13-16, lokale 2.0.30
- justering: midterm udleveres på workshoppen kl. 13

Uge	Dato	Tema	Litteratur	Case
1	5/9	Introduktion til R	Imai kap 1	
2	12/9	Regression I: OLS	GH kap 3, MM kap 2	Gilens & Page (
3	26/9	Regression II: Panelmodeller	GH kap 11	Larsen et al. (20
4	29/9	Regression III: Multilevelmodeller, interaktioner	GH kap 12	Berkman & Plut
5	3/10	Introduktion til kausal inferens	Hariri (2012), Samii (2016)	
6	10/10	Matching	Justesen & Klemmensen (2014)	Ladd & Lenz (2
	17/10	*Efterårsferie*		

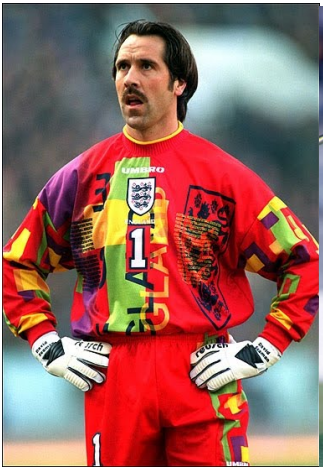
Uge	Dato	Tema	Litteratur	Case
	17/10	*Efterårsferie*		
7	24/10	Eksperimenter I	MM kap 1, GG kap 1+2	Bond et al. (2012)
8	31/10	Eksperimenter II	GG kap 3+4+5	Gerber & Green (2000)
9	7/11	Instrumentvariable	MM kap 3	Arunachalam & Watso
10	14/11	Regressionsdiskontinuitetsdesigns	MM kap 4	Eggers & Hainmueller
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?

- opsamling: subsetting, plotting
- formlen for en regressionskoefficient (kovarians, varians)
- motivation: hjalp boligboblen Fogh?
- introduktion til paneldata
- fixed effects modeller
- vigtigt forbehold: clustering

Spørgsmål?

Eksempel til illustration: to målmænd



Vigtig KPI for målmænd: redningstilbøjelighed π_i

- $\bar{\pi} = .1$
- $\pi_S = \frac{150}{1000}$
- $\pi_C = \frac{2}{5}$

$\rightarrow$ hvilken målmand bør vi foretrække?

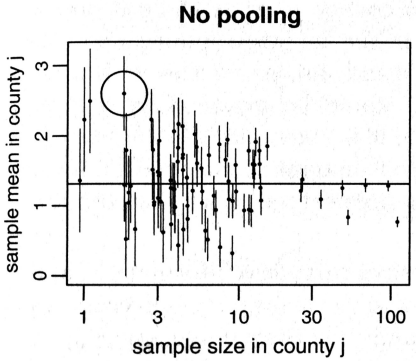
- complete-pooling: $\pi_S = \pi_S = \bar{\pi} = .1 \rightarrow$ indifferent
- no-pooling: $\pi_S = .15$, $\pi_C = .2 \rightarrow$ foretrækker Campos
- er disse tilfredsstillende? hvorfor/hvorfor ikke?

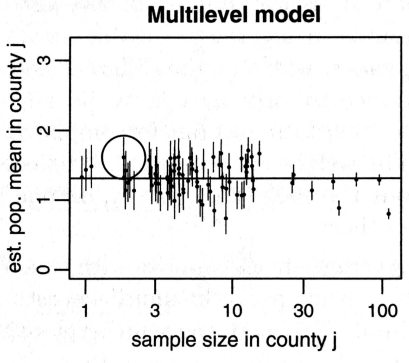
$$\hat{\alpha}_j^{multilevel} \approx \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{all}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \quad (1)$$

central indsigt: partial pooling $\rightarrow \hat{\alpha}_j^{multilevel}$ estimeres som et vægtet gennemsnit af $\bar{y}_j$ og $\bar{y}_{all}$

bemærk: multilevelaspektet vedrører her kun α, dvs. konstantleddet $\rightarrow$ *varying intercepts*

Illustration i GH:





complete-pooling i modelform:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{2}$$

no-pooling i modelform:

$$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \tag{3}$$

partial pooling: som no-pooling, men $\alpha_{j[i]}$ modelleres som i (1)
$\rightarrow$ konstantled følger en (normal)fordeling, hvis varians estimeres i modellen

Spørgsmål?

Observationers 'klumpethed' (clustering) udtrykkes typisk ved *intraclass correlation coefficient* (ICC):

$$ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_y^2} \tag{4}$$

fortolkning: hvor stor andel af samlet variation afspejler gruppeforskelle?
$\rightarrow$ ICC bruges også i psykometri som reliabilitetsmål

Spørgsmål?

Eksempel (fra Steenbergen & Jones, 2002): hvad betyder et lands samhandel med EU for borgeres holdning til EU?

Table 4 Determinants of EU Support

Parameter	Multilevel Estimate	Regression Estimate
Fixed Effects		
Constant	5.504**	5.016**
	(.220)	(.124)
Tenure	0.014	0.011**
	(.014)	(.002)
Trade	0.032	0.039**
	(.025)	(.003)
Party Cue	0.233**	0.275**
	(.028)	(.018)
Lowest Income Quartile	−.106+	−.181**
	(.064)	(.068)
Highest Income Quartile	0.048	−.001
	(.059)	(.062)
Ideology	0.019	0.023+
	(.015)	(.013)
Opinion Leadership	0.152**	0.166**
	(.028)	(.030)
Male	0.088+	0.093+
	(.050)	(.053)
Age	−.013**	−.014**
	(.002)	(.002)

- data: 6,354 respondenter fra 15 lande i Eurobarometer 1996
- fokus her: *Trade* = landets samhandel med EU-lande
- hvorfor ændrer signifikansniveauet sig når man tager højde for multilevelstrukturen?

»These differences arise precisely because the OLS standard errors are too small. This attenuation is caused by ignoring the clustering of the data. The OLS analysis assumes that we have 6354 independent observations in our data. (...) The problem, of course, is that we do not have 6354 independent observations. (...) To pretend that they are independent is to assume that one has more information than really exists. Thus, the OLS analysis presents too optimistic a view about the significance of the predictor«

– Steenbergen, M. R., & Jones, B. S. (2002). Modeling Multilevel Data Structures. *American Journal of Political Science*, 46(1), 218-237.

Central fordel ved multilevelmodeller: variable på gruppeniveau kan indgå sammen med gruppeindikatorer ($\sim$ gruppe FE) i tværsnitsdata

- i klassisk OLS: gruppe-FE og gruppevariabel er kollineære $\rightarrow$ kan ej estimeres
- i multilevel regression: α_j'er estimeres m. partial pooling
- $\rightarrow$ jf. sammenligning i scriptet

Spørgsmål?

Antag en klassisk regressionsmodel:

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i \tag{5}$$

men: effekterne af X_i og Z_i afhænger af den anden variabels værdi $\rightarrow$ de *interagerer*:

$$\beta_1 = \delta_1 + \delta_2 Z_i \tag{6}$$

$$\beta_2 = \delta_3 + \delta_4 X_i \tag{7}$$

dermed:

$$Y_i = \alpha + \beta_x X_i + \beta_z Z_i + \beta_{xz} X_i Z_i \tag{8}$$

bemærk: $\beta_x \neq \beta_1$, $\beta_z \neq \beta_2$

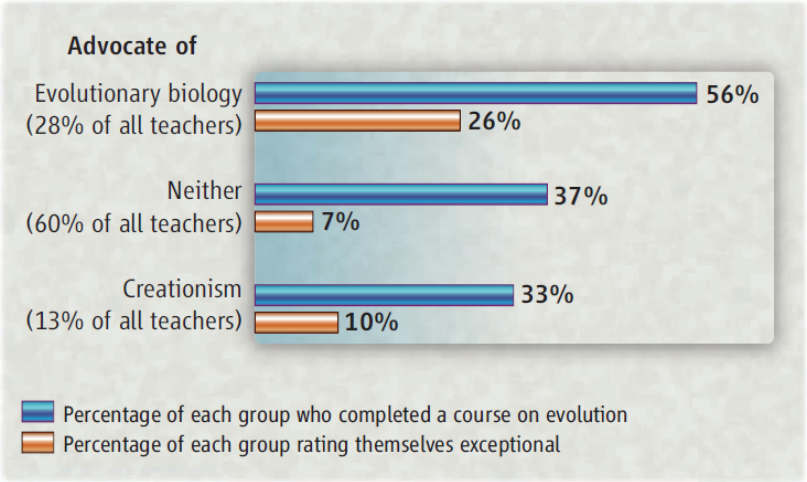
Implementering i R:

```
lm(y~x+z+x:z,data=df)
```

eller

```
lm(y~x*z,data=df)
```

Spørgsmål?


Advocate of
Evolutionary biology
(28% of all teachers)
56%
26%
Neither
(60% of all teachers)
37%
7%
Creationism
(13% of all teachers)
33%
10%
Percentage of each group who completed a course on evolution
Percentage of each group rating themselves exceptional

»[C]hange due to improved standards is likely to be slow, because standards have the greatest impact on the newest teachers – those who were socialized in an era of standards – based education and who take standards and testing for granted«

$\rightarrow$ hvilken slags model kan teste dette udsagn?

Næste gang:

- introduktion til kausal inferens
- Hariri (2012), Samii (2016)
- ingen case-tekst

Tak for i dag!