

Multilevel analyse

Metode 3

Link til slides: kortlink.dk/2acds

Søren Damsbo-Svendsen
sdas@ifs.ku.dk

Institut for Statskundskab
Københavns Universitet

Uge 11

Oversigt

Uge	Holdtime	Emne	Øvelsesopgave
6		Kvantitativ indholdsanalyse	1
7		Diskursanalyse I	2
8	25-26 februar 2021	Diskursanalyse II	3
9	04-05 marts 2021	Interaktioner og modelspecifikation	4
10	11-12 marts 2021	Logistisk regression	5
11	18-19 marts 2021	Multilevel analyse	6
12		Kausal inferens I: Kausalitet og instrumentvariable (IV)	7
13		Påskeferie	
14	08-09 april 2021	Kausal inferens II: Paneldata	8
15	15-16 april 2021	Kausal inferens III: Eksperimentelle designs	9
16		Kausal inferens IIII: Regression Discontinuity (RD)	10
17		Social Data Science: Big Data	11
18		Process tracing	12
19		Kriterier for god videnskab	
20		Opsamling/spørgetime	
23		Aflevering af skriftlig hjemmeopgave (07 juni 2021)	

Recap fra sidste gang

- Vi bruger **logistisk regression** i stedet for OLS, når Y er binær, pga. OLS' problemer med funktionel form (linearitet), fejlledd, absurde sandsynligheder mv.
- I logistisk regression er der ikke bare én effekt af $X \rightarrow Y$ (afhænger af, hvor på X-aksen, vi står)
- **Beta-koefficienterne** angiver effekten af at X vokser med +1 på *den naturlige logaritme til oddset* for, at $Y=1$
 - umuligt at forstå intuitivt og **odds ratio** gør det ikke meget bedre
- Vi fortolker vha. tricks
 - **Average Marginal Effect** (AME): alle marginale effekter opsummeret i et gennemsnit
 - **Marginal Effect at the Mean** (MEM): marginal effekt ved gennemsnittet af X
 - **Forudsagte sandsynligheder** (S-kurven): sandsynligheden for $Y=1$ for alle niveauer af X

Dagens program

- **Multilevel-analyse** → nyt emne!
- Hvad er det, hvornår og hvorfor bruger vi det?
- Hvordan udfører vi multilevel-analyse i **Stata** og fortolker output?
 - Eksempler fra Mehmetoglu & Jakobsen (2016)
- **Øvelsesopgaver** i Stata

Dagens formål

- At få **bedre forståelse** for multilevel-analyse
 - hvilke spørgsmål det besvarer
 - hvilke problemer det afhjælper
 - hvilke muligheder det giver
- At træne **udførelse og fortolkning** i Stata

Pensum

Hox, J. J. (2010). *Multilevel analysis: techniques and applications* (2. edition). Routledge. Kapitel 1

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using Stata: a guide for the social sciences*. Sage. Kapitel 9

Hvad er multilevel-analyse?

Vi kan se på det som endnu en **udvidelse af OLS**:

"Multilevel analysis can be seen as a generalization of OLS regression to accommodate the complexities of estimating regression models with two or more levels"

(Mehmetoglu & Jakobsen, 2016, p. 197)

Multilevel research

"The general concept is that **individuals interact with the social contexts** to which they belong, that individual persons are influenced by the social groups or contexts to which they belong, and that those groups are in turn influenced by the individuals who make up that group. The individuals and the social groups are conceptualized as a **hierarchical system of individuals nested within groups**, with individuals and groups defined at **separate levels** of this hierarchical system [...] This leads to research into the **relationships between variables characterizing individuals and variables characterizing groups**, a kind of research that is generally referred to as **multilevel research**"

(Hox, 2010, p. 1)

Hvorfor multilevel-analyse?

- Når vores data har en hierarkisk struktur, giver det både problemer og muligheder
- **Metodemæssige grund: Statistisk og kausal inferens**
 - Vi vil gerne have **korrekte standardfejl** (usikkerhed), så vi ikke får for høje/lave **p-værdier** og drager forkerte konklusioner
 - Vi vil gerne kunne kontrollere for gruppe-variable, der både påvirker X og Y (på niveau 1), for at imødegå *omitted variable bias* (spuriøsitet)
- **Substancial/teoretisk grund: Viden om multilevel-dynamikker og konteksteffekter**
 - Vi kan være substancialt interesserede i konteksteffekter, herunder
 - direkte effekt af niveau 2-variabel (X) på niveau 1-variabel (Y)
 - cross-level interaktioner (*hvordan konteksten betinger niveau 1-effekter*)
 - Vi kan altså arbejde med multilevel-spørgsmål/-hypoteser/-problemer

Begreber

- **Niveauer** (levels)
 - niveau 1 = det "laveste" niveau \approx individniveau
 - niveau 2 = gruppeniveau = klynger = noget der *indeholder* niveau 1-observationer \approx landeniveau
- **Multilevel-analyse** = multilevel-model = multilevel-regression
- **Intercept** = Y intercept = konstant = $\beta_0 = \alpha$
 - **varying intercept** = random intercept \approx forskellige gruppe-gennemsnit
- **Slope** = hældning = effekt = (beta)koefficient = β_1
 - **varying slope** = random effect (RE) = random coefficient \approx forskellige effekter på tværs af grupper

Spørgsmål

- Break-out i 10 min.
 - Diskutér så mange spørgsmål som muligt
 - Bagefter hører vi svar fra **gruppe 1/6 på spgm. 1**, fra **gruppe 2/7 på spgm. 2**, fra **gruppe 3/8 på spgm. 3**, fra **gruppe 4/9 på spgm. 4** og fra **gruppe 5/10 på spgm. 5**
-

1. Hvad er et multilevel-problem? Giv et opfundet eksempel på et ML-problem eller en ML-hypotese (se evt. Hox, 2010, pp. 6-7)
2. Hvordan adskiller multilevel-analyse sig fra almindelig regression? (behøver ikke være teknisk)
3. Giv et par eksempler på data med multilevel-struktur
4. Hvad menes der med "aggregering" og "disaggregering"? (se evt. Hox, 2010, pp. 2-4)
5. Hvad er *det statistiske problem* ved at aggregere hhv. disaggregere data? Hvad sker der med N ? (se evt. Hox, 2010, p. 3)

Multilevel-regression

Multilevel-modellen begynder med at opdele **variansen i den afhængige variabel** på de to niveauer

- **Variansen i den afhængige variabel** er et mål for spredningen eller afvigelserne *rundt om variablen's gennemsnit*
- Variansen opdeles i:
 - (1) spredningen **inden for hver gruppe** rundt om *gruppens gennemsnit* (lagt sammen)
 - (2) spredningen **mellem grupperne**, dvs. gruppegennemsnittenes spredning om *det samlede gennemsnit*
- Denne opdeling kan ses i **Null-modellen** uden uafhængige variable (vi ser det om lidt)
- Matematisk ser det således ud:
 - $Y_{ij} = \beta_0 + u_{0j} + e_{ij}$
 - *dependent_variable = total_mean + error_term_level_2 + error_term_level_1*

Multilevel-regression

Bottom line

- Takket være variansopdelingstricket, kan vi inddrage uafhængige variable fra begge niveauer
 - fx $\beta_1 X_{1ij}$ (niveau 1) og $\beta_2 X_{2j}$ (niveau 2)
- Vær opmærksomme på **antal variable ift. antal observationer** på niveau 2
 - tommelfingerregel* om min. **10 obs.** per uafhængig variabel
- Multilevel-modellen bruger automatisk det korrekte N for hvert niveau til at beregne usikkerheder og p-værdier → smart!

Stata, fremgangsmåde og eksempler

Multilevel-analyse i Stata

`mixed Y_lvl1 || ID_lvl2:` (*Null-model*)

`mixed Y_lvl1 X1_lvl1 X2_lvl1 X3_lvl1 X4_lvl2 X5_lvl2 || ID_lvl2:`

- **ID_lvl2** er en variabel, der identifierer grupperne (fx *landenavn*). Husk kolon i enden
- **_lvl1** og **_lvl2** indikerer, hvilket niveau variablen hører til i eksemplet (Stata finder selv ud af dette)
- Eventuelle *varying slopes* tilføjes efter **ID_lvl2**:

Options

- "ml variance" specifiserer, at modellen skal fittes med *maximum likelihood estimation*, og at vi vil se *opdelingen af variansen*. **Begge er standard** og derfor ikke nødvendige

Fremgangsmåde

Formål

- Vi er interesserede i **indflydelsen fra konteksten (niveau 2)**, fordi vores teori, problemformulering og/eller hypoteser tilsiger dette
- Vi vil gerne estimere effekten af en eller flere **uafhængige variable (X) på niveau 2** på en **afhængig variabel (Y) på niveau 1**
 - foruden effekten af evt. uafhængige variable (X) på niveau 1

Fremgangsmåde

Seks trin

1. Lav tom ("Null") model, der alene skelner mellem niveauerne
2. Tilføj uafhængige **niveau 1**-variable
3. Tilføj uafhængige **niveau 2**-variable
4. Tilføj evt. **varying slopes** for uafhængige niveau 1-variable
5. Tilføj evt. **interaktion** - enten på samme niveau eller *cross-level*
6. **Fortolk** på helheden med fokus på de "fulde" modeller

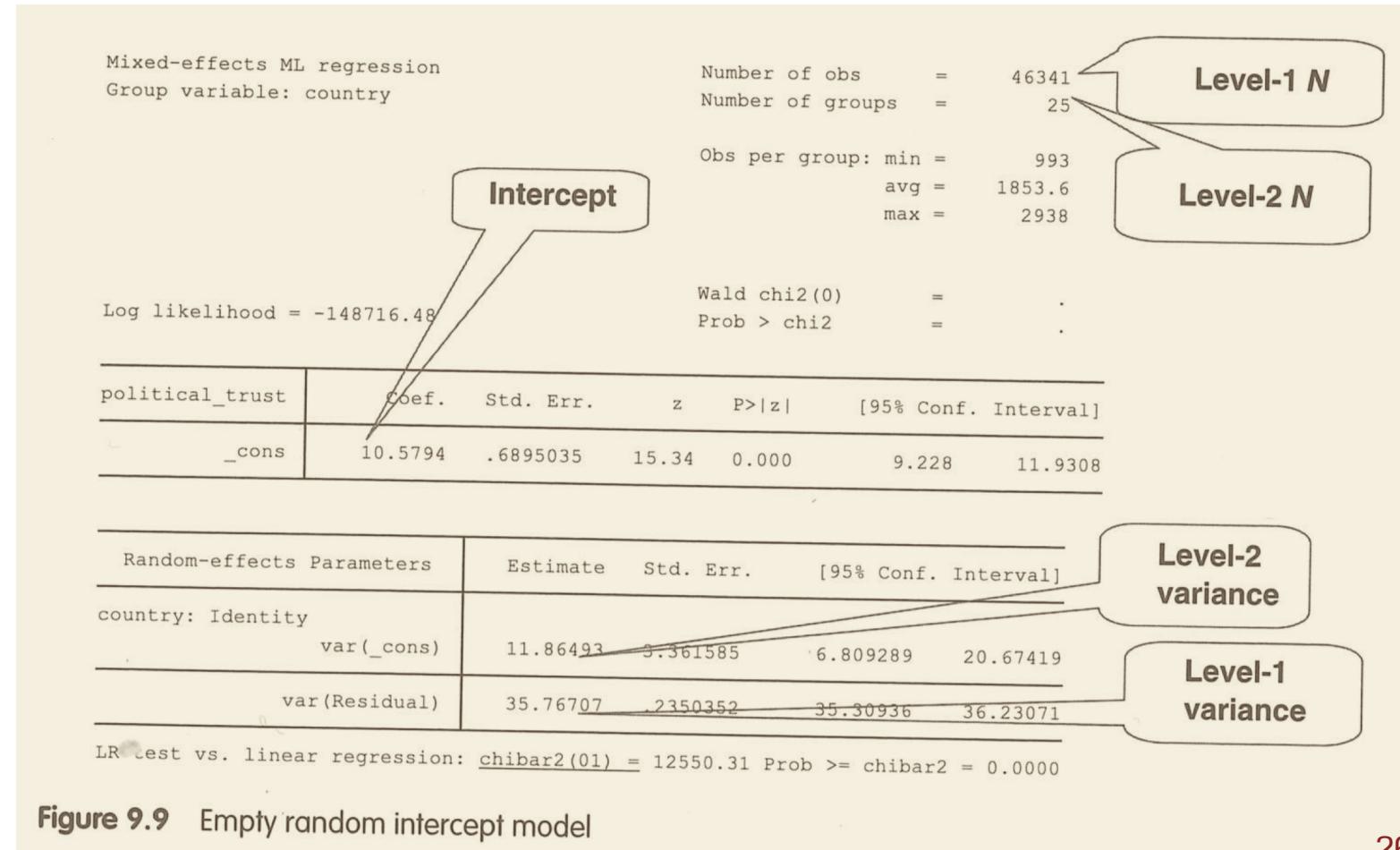
1. Lav tom ("Null") model

Eksempel fra Mehmetoglu & Jakobsen, 2016, pp. 201ff

mixed political_trust ||
country:

Giver os bl.a. **antal obs.**
(N) per niveau og den
gennemsnitlige tillid
(_cons)

Desuden **opdelingen af**
den uforklarede varians i
tillid på gruppeniveau,
var(_cons), og
individniveau,
var(Residual)



1. Lav tom ("Null") model

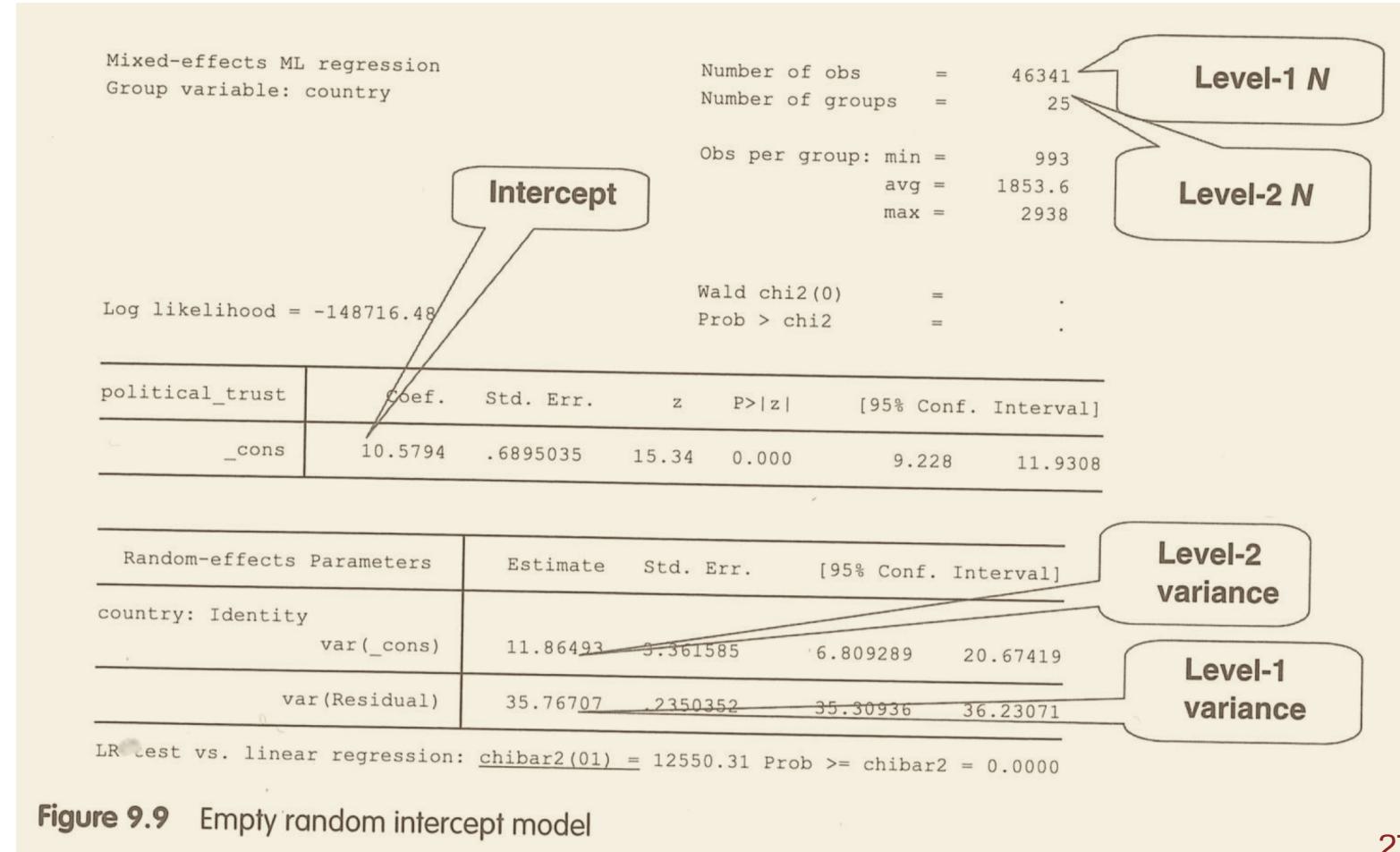
Eksempel fra Mehmetoglu & Jakobsen, 2016, pp. 201ff

Hvor stor en andel af den samlede varians i tillid tilskrives gruppeneveauet (hvor stor er VPC/ICC)?

$$\frac{11.9}{11.9+35.8}$$

Tommelfingerregel:
Min. 5 % før niveau 2 er relevant

Kan det betale sig at fortsætte med en multilevel-model?



2. Tilføj uafhængige niveau 1-variable

mixed political_trust

woman age unemployed

eduys || country:

Påvirker uddannelse (X4) politisk tillid (Y)?

Er den uforklarede varians i tillid faldet? Markant?

Mixed-effects ML regression	Number of obs	=	45288		
Group variable: country	Number of groups	=	25		
	Obs per group:	min =	952		
		avg =	1811.5		
		max =	2874		
Log likelihood = -145227.59	Wald chi2(4)	=	232.56		
	Prob > chi2	=	0.0000		
political_trust	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
woman	.0004654	.0564919	0.01	0.993	-.1102566 .1111874
age	.005863	.0016091	3.64	0.000	.0027094 .0090167
unemployed	-1.12698	.127711	-8.82	0.000	-1.377289 -.8766708
eduhrs	.0903774	.0075716	11.94	0.000	.0755374 .1052174
_cons	9.238013	.6914083	13.36	0.000	7.882877 10.59315
Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]		
country: Identity					
var(_cons)	11.43201	3.239449	6.560255	19.92162	Unexplained level-2 variance
var(Residual)	35.59259	.2365936	35.13188	36.05934	Unexplained level-1 variance

Figure 9.10 Stata output for random intercept model

2. Tilføj uafhængige niveau 1-variable

Påvirker uddannelse (X4)
politisk tillid (Y)?

Ja ($p=0.000$). For hvert år
vokser tilliden med 0.09

Er den uforklarede varians
i tillid faldet? Markant?

Faldet marginalt, men
stort set uændret. De nye
variable (den nye model)
har således ikke megen
yderligere forklaringskraft

Mixed-effects ML regression					
Group variable: country		Number of obs	=	45288	
		Number of groups		25	
		Obs per group: min =		952	
		avg =		1811.5	
		max =		2874	
		Wald chi2(4) =		232.56	
		Prob > chi2 =		0.0000	
political_trust	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
woman	.0004654	.0564919	0.01	0.993	-.1102566 .1111874
age	.005863	.0016091	3.64	0.000	.0027094 .0090167
unemployed	-1.12698	.127711	-8.82	0.000	-1.377289 -.8766708
eduhrs	.0903774	.0075716	11.94	0.000	.0755374 .1052174
_cons	9.238013	.6914083	13.36	0.000	7.882877 10.59315
Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
country: Identity					
var(_cons)		11.43201	3.239449	6.560255	19.92162
var(Residual)		35.59259	.2365936	35.13188	36.05934
LR test vs. linear regression: chibar2(01) = 11569.74 Prob >= chibar2 = 0.0000					

Effect of variables

Unexplained level-2 variance

Unexplained level-1 variance

Figure 9.10 Stata output for random intercept model

3. Tilføj uafhængige niveau 2-variable

mixed political_trust

woman age unemployed

eduysr GDPcapita1000 ||

country:

Påvirker uddannelse (X4)

stadic politisk tillid (Y)?

Hvad kan vi sige om
effekten af BNP (X5)?

Er den uforklarede varians
i tillid faldet denne gang?

Markant?

Mixed-effects ML regression		Number of obs = 45288			
Group variable: country		Number of groups = 25			
		Obs per group: min = 952			
		avg = 1811.5			
		max = 2874			
Log likelihood = -145215.39		Wald chi2(5) = 275.58			
		Prob > chi2 = 0.0000			
political_trust		Coef.	Std. Err.	z	P> z
woman		.0009456	.0564918	0.02	0.987
age		.0058665	.001609	3.65	0.000
unemployed		-1.127754	.1277087	-8.83	0.000
eduysr		.0904258	.0075707	11.94	0.000
GDPcapital1000		.1409978	.0219212	6.43	0.000
_cons		4.839403	.8118732	5.96	0.000
[95% Conf. Interval]					
Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
country: Identity					
var(_cons)		4.293036	1.219944	2.459689	7.492881
var(Residual)		35.59258	.2365935	35.13188	36.05933
LR test vs. linear regression: chibar2(01) = 5176.31 Prob >= chibar2 = 0.0000					

Figure 9.13 Random intercept model, including level-2 variable

3. Tilføj uafhængige niveau 2-variable

Effekten af uddannelse
(X4) er praktisk talt
uændret: $\beta \approx 0.09$

Når BNP/cap. vokser med
\$1000, øges hvert lands
gennemsnitlige tillid med
0,14

På landeniveau er den
uforklarede varians faldet
markant fra 11,9 til 4,3

→ ca. 64 % af gruppe-
variansen kan forklares
vha. BNP

$$\frac{11.9 - 4.3}{11.9} \approx 0,64$$

Mixed-effects ML regression		Number of obs = 45288			
Group variable: country		Number of groups = 25			
		Obs per group: min = 952			
		avg = 1811.5			
		max = 2874			
Level-2 variable		Wald chi2(5) = 275.58			
Log likelihood = -145215.39		Prob > chi2 = 0.0000			
political_trust	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
woman	.0009456	.0564918	0.02	0.987	-.1097762 .1116674
age	.0058665	.001609	3.65	0.000	.0027129 .0090201
unemployed	-1.127754	.1277087	-8.83	0.000	1.378059 .8774497
eduhrs	.0904258	.0075707	11.94	0.000	.07555874 .1052641
GDPcapital1000	.1409978	.0219212	6.43	0.000	.0980331 .1839626
_cons	4.839403	.8118732	5.96	0.000	3.248161 6.430645
Random-effects Parameters		Estimate	Std. Err.	[95% Conf. Interval]	
country: Identity					
var(_cons)		4.293036	1.219944	2.459689 7.492881	
var(Residual)		35.59258	.2365935	35.13188 36.05933	
LR test vs. linear regression: chibar2(01) = 5176.31 Prob >= chibar2 = 0.0000					

SE calculated
based
on level-2 N

Figure 9.13 Random intercept model, including level-2 variable

4. Tilføj evt. varying slopes for niveau 1-variable

mixed political_trust

woman age unemployed

eduysr GDPcapita1000 ||

country: eduysr

Hvilken uafhængig
variabel (og fra hvilket
niveau) er specifiseret
med **varying slopes**?

Hvad vil det sige?

Hvad kan vi sige om dens
effekt på tillid?

Ekskurs: Hvad er varying slopes?

- Normalt identificeres den (fixed) effekt, β , som passer bedst til hele datasættet
 - der kan godt tages højde for gruppeforskelle, men der findes altså kun **én enkelt effekt, som gælder for alle grupper**
- Sommetider har vi lyst til at slække på denne antagelse og **beregne forskellige effekter for de forskellige grupper**
 - det kalder vi **varying slopes** eller *random effects*
 - vi skal have **god (teoretisk) grund** til at tro, at effekten varierer substancialt
 - varying slopes skal medføre en statistisk signifikans **forbedring af modellens forklaringskraft**, jf. Statas **Irtest**
- Stata præsenterer os for **gennemsnittet af de forskellige effekter**, som vi så kan tolke på
- Man kan sagtens lave en god multilevel-analyse uden varying slopes

4. Tilføj evt. varying slopes for niveau 1-variable

Vi har tilføjet varying
slopes for **uddannelse**
(niveau 1)

Vi mener, at uddannelse
kan påvirke tillid
fundamentalt forskelligt i
forskellige lande. Derfor
beregner vi effekten
indenfor alle lande.

Koefficienten for *eduyrs* er
gennemsnitseffekten

Når individets
uddannelses-niveau
vokser med +1, øges

5. Tilføj evt. interaktion

mixed political_trust

woman unemployed

edugrs GDPcapita1000

i.Nordic##c.age || country:

// OBS! Ingen varying

slopes for *age*, som

Carolin anbefalede?

Er det en alm. same-level

eller en cross-level

interaktion?

Er effekten af alder

betinget af, om landet er

nordisk?

5. Tilføj evt. interaktion

Er det en alm. same-level
eller en cross-level
interaktion?

*Cross-level! Nordic er en
dummy, der angiver om
landet (gruppen) er et af
de nordiske lande*

Er effekten af alder
betinget af, om landet er
nordisk?

*Ja! Effekten af alder er
-0,0262 mindre i nordiske
lande ($p=0,000$) → og den
er 0,0098 i ikke-nordiske
lande*

5. Tilføj evt. interaktion

Er det en alm. same-level
eller en cross-level
interaktion?

*Cross-level! Nordic er en
dummy, der angiver om
landet (gruppen) er et af
de nordiske lande*

Er effekten af alder
betinget af, om landet er
nordisk?

*Ja! Effekten af alder er
-0,0262 mindre i nordiske
lande ($p=0,000$) → og den
er 0,0098 i ikke-nordiske
lande*

Tre ekstra-ting

1. Udvidelser
2. Antagelser
3. Data

Udvidelser

1. Logistisk multilevel regression (logit) → brug det kun, hvis det er nødvendigt
2. Multilevel-model med tre niveauer → brug det kun, hvis det er nødvendigt
3. Cross-classified multilevel-model
 - ikke en klar hierarkisk struktur, fx individer indlejret i sideordnede kontekster
 - → brug det kun, hvis det er nødvendigt

Antagelser

- **Ikke pensum!** - men problematikken kan eventuelt nævnes
- I princippet samme antagelser som OLS - *gange to*
 - linearitet, uafhængige obs., homoskedasticitet, normalfordelte fejller etc. skal i principippet være opfyldt på hvert niveau
 - det meste er svært-til-umuligt at teste

Data

- **Data-tilgængelighed** er en fordel ved multilevel-analyse - i hvert fald med *individer* inden for *lande*
- Mange supergode **multinationale survey-datasæt** frit tilgængelige
 - fx European Social Survey og World Values Survey
- Komparative **landedata** også frit tilgængelige, bl.a. hos EU, Verdensbanken og OECD
- Rig mulighed for *replikationsstudier*
- **Hvordan merger man individdata med landedata i Stata?**
 - Jeg lovede Carolin at vise dette
 - Samme fremgangsmåde uanset data; det afgørende er at have **en variabel, der går helt nøjagtigt igen** i begge datasæt
 - Eksempel (do-fil kommer på Absalon) →

Hvordan man merger individdata med landedata: *do-fil*

```
1 * Lav eksempel-data på landeniveau
2 clear *
3
4 input str20 land bnp str
5 "danmark" 5000 10000
6 "sverige" 6000 12000
7 "tyskland" 7000 30000
8 "frankrig" 5000 2500
9 end
10
11 save lande, replace
12
13 * Lav eksempel-data på individniveau *med samme landeidentifier som landedata* (variablen "land")
14 clear *
15
16 input person str20 land alder udd
17 1 "danmark" 20 1
18 2 "danmark" 25 3
19 3 "danmark" 35 5
20 4 "danmark" 45 5
21 5 "sverige" 30 5
22 6 "sverige" 50 4
23 7 "sverige" 55 2
24 8 "tyskland" 80 1
25 9 "tyskland" 65 2
26 10 "frankrig" 25 3
27 end
28
29 * Merge landedata (lande.dta) ind på individdata (åbent datasæt) som "many-to-one" (m:1), fordi hvert land optræder flere
30 gange i individdata, men kun én gang i landedata, baseret på landeidentifier-variablen "land"
31 merge m:1 land using lande.dta
32
33 * Se resultatet
34 sort person // sorter efter personid
bro
```

Hvordan man merger individdata med landedata: *resultat*

	person	land	alder	udd	bnp	str	_merge
1	1	danmark	20	1	5000	10000	matched (3)
2	2	danmark	25	3	5000	10000	matched (3)
3	3	danmark	35	5	5000	10000	matched (3)
4	4	danmark	45	5	5000	10000	matched (3)
5	5	sverige	30	5	6000	12000	matched (3)
6	6	sverige	50	4	6000	12000	matched (3)
7	7	sverige	55	2	6000	12000	matched (3)
8	8	tyskland	80	1	7000	30000	matched (3)
9	9	tyskland	65	2	7000	30000	matched (3)
10	10	frankrig	25	3	5000	2500	matched (3)

Øvelsesopgaver i Stata

Dagens pointer

- Multilevel-analyse er smart, når vi har hierarkiske data og vil undersøge **effekten af konteksten (niveau 2) på et niveau 1-outcome**
- Det fungerer (bl.a.) ved at **opdele variansen** på gruppe- og individniveau og ved automatisk at **justere standardfejl**
- Ofte følger man en **fast fremgangsmåde** med trinvis tilføjelse af uafhængige variable - begyndende fra en tom **Null-model**
- Der er mange muligheder og mulige udvidelser, fx tilføjelse af **varying slopes** eller **(cross-level) interaktioner**
- Er man ikke substantielt interesseret i niveau 2, kan man dog komme langt med OLS

Vi har haft 4/6 holdtimer → Er der noget, der kunne være bedre eller fungerer godt?

- Plz skriv på sdas@ifs.ku.dk

Næste gang

- Vi starter på fire uger med **kausal inferens**
 - begynder at tage kausalitetsspørgsmålet alvorligt frem for blot at antage "effekter" for eksemplernes skyld
- Næste uge
 - **Kausal inferens I:** Kausal inferens og instrumentvariable
 - ingen holdtime
 - derefter **PÅSKEFERIE**
- Vi ses igen efter påske til Kausal inferens II:
Paneldata



Tak for i dag!

