Introduktion til grupperet data

Sarah Yde Junge [1] *& Emil Bargmann Madsen* [2]

Juni 2023

[1] Institut for Statskundskab, Aarhus Universitet (syj@ps.au.dk).

[2] Dansk Center for Forskningsanalyse, Institut for Statskundskab, Aarhus Universitet (ebm@ps.au.dk.)

Vi vil gerne takke Andreas Videbæk Jensen for hjælpsomme kommentarer og sparring omkring denne note.

Centrale begreber:

Paneldata
Hierakisk data
Selektion
Klyngerobuste standardfejl
Fixed effects
Multilevelmodeller

Formålet med denne note er at introducere en type struktur som et datasæt kan have, hvor observationerne eller enhederne på den ene eller anden måde er indlejret i en gruppe, og hvor viden om denne gruppering er vigtig for at forstå data om den enkelte observation. Vi forsøger først at skabe et overblik over, hvad grupperet data er ved at give en række eksempler på, hvordan en sådan struktur opstår, og hvordan grupperet data kan se ud. Dernæst opridser vi både substantielle og statiske problemer ved at ignorere en mulig gruppestruktur i ens dataanalyse, og afslutter med at kigge på mulighederne for at tage højde for gruppering i data.

Hvad er grupperet data?

Data kan siges at være grupperet, når de enkelte observationer ikke er uafhængige af hinanden, men er indlejret i grupper, i hvilke der er en indbyrdes afhængighed.[3]

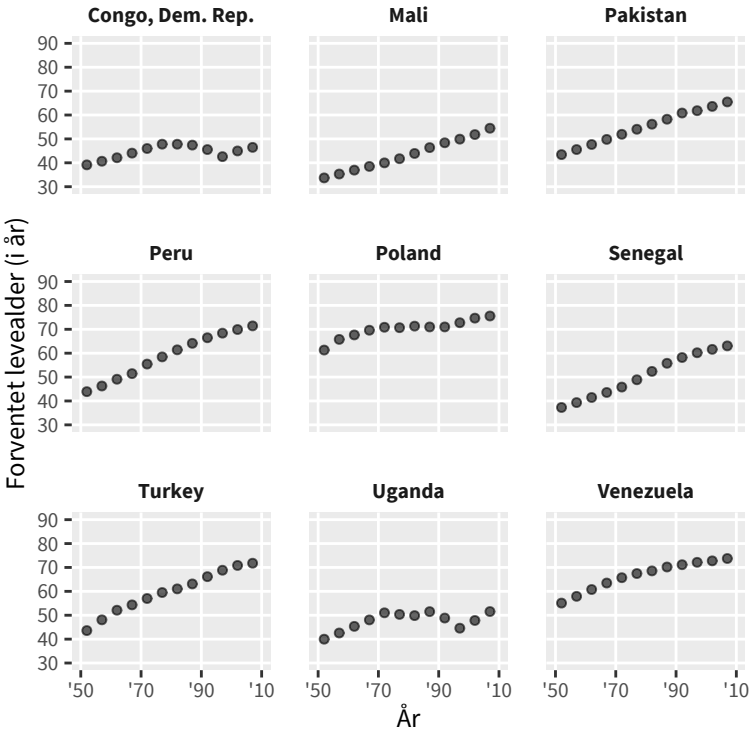
Hvis vi har to observationer Y_1 og Y_2 af en (stokastisk) variabel **Y**, kan vi forstå uafhængighed mellem disse ved at forestille os, at vi kender værdien på Y_1. Hvis vi bliver bedre til at gætte værdien af Y_2 ved at basere vores gæt på værdien af Y_1 fremfor blot, som vi ellers ville gøre, at gætte på, at værdien af Y_2 er lig gennemsnittet af Y, betyder det, at der er en afhængighed mellem Y_1 og Y_2.

Afhængighed mellem observationer kan skabe både selektion i vores model, forårsage fejlestimering af vores standardfejl og svække undersøgelsens eksterne validitet. Derfor er vi nødt til at forholde os til en eventuelt grupperet struktur i vores data. Afhængigheder kan opstå på mange måder, og nedenfor vil vi forsøge at skitsere forskellige eksempler på dette.

[3] Vi bruger her betegnelsen "grupperet data", men denne type datastruktur kaldes også typisk hierarkisk data, paneldata og på engelsk "clustered"eller "nested"data.

Enheder målt gentagne gange

Den første type grupperet data, vi vil introducere, er paneldata. Paneldata består af n enheder målt t gange over tid. Hver observation udgør altså ikke bare værdien for én enhed, men for én enhed på ét specifik tidspunkt. I dette tilfælde betegner vi en observation Y_{it}, hvor Y er observationens værdi på variablen **Y**, i angiver enhed $1, ..., n$, og t angiver tidspunkt $1, ..., t$. Figur 1 viser et eksempel på denne type data. Her er den forventede levealder målt for 9 lande mellem 1950 og 2010.



Figur 1: Forventet levealder for 9 lande, 1950-2010.

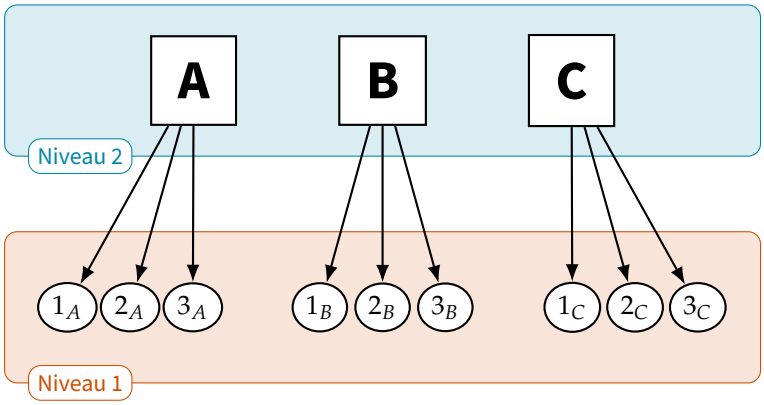
Hvert enkelt punkt viser den gennemsnitlige levealder for et land på et bestemt tidspunkt (f.eks. $Y_{\text{Mali, 1977}}$). Figuren viser, at næsten alle lande oplever en positiv udvikling i den forventede levealder, men også at der er tydelig variation mellem lande både i udgangspunktet, og i hvor meget og hvor hurtigt levealderen stiger. Samtidig kan vi også se, at den forventede levealder i et land på et bestemt tidspunkt bedst kan forudsiges ved at kigge på de tidligere år for det givne land. F.eks. giver den forventede levealder for Venezuela i 1992 os dårlige forudsætninger for at forudse nedgangen i levealder i Uganda i 1997. Kender vi derimod den gennemsnitlige levealder i Uganda i 1987-1992 har vi langt bedre mulighed for at forudsige den stigende dødelighed i 1997. Det er altså tydeligt her, at observationerne over

tid ikke er uafhængige af hinanden. De afhænger af, hvilket land vi observerer, og vi taler derfor om observationerne i dette datasæt som grupperet i lande.

Enheder indlejret i fælles grupper eller institutioner

En anden type af grupperet data er det, vi kalder hierarkisk data. I hierarkisk data er enhederne modsat paneldata kun observeret én gang. I stedet opstår den grupperede struktur, fordi observationerne ved denne type data tilhører fælles kontekster, grupper eller institutioner. Et godt eksempel på hierarkisk data er skoleelever. Hvis vi har en stikprøve med elever fra flere forskellige skoler, er der en stor sandsynlighed for, at de elever, der går på den samme skole har en række karakteristika til fælles. Det kunne være, hvor gode undervisningsfaciliteter de har, hvor dygtige deres lærere er, men også hvilket nabolag, de bor i. Det vil betyde, at observationer af elever, der går på den samme skole, ikke er uafhængige af hinanden - eleverne deler en række karakteristika, som får dem til at ligne hinanden. Dermed bidrager hver observation af en elev fra den samme skole med mindre ny information - det er det, vi kalder uafhængig variation - til vores data. Vi kalder denne type data hierarkisk data, fordi de enkelte analyseenheder, i dette eksempel elever, er indlejeret i en overliggende kontekst, her en skole, og at data indeholder flere af sådanne grupperinger, i vores eksempel flere skoler.

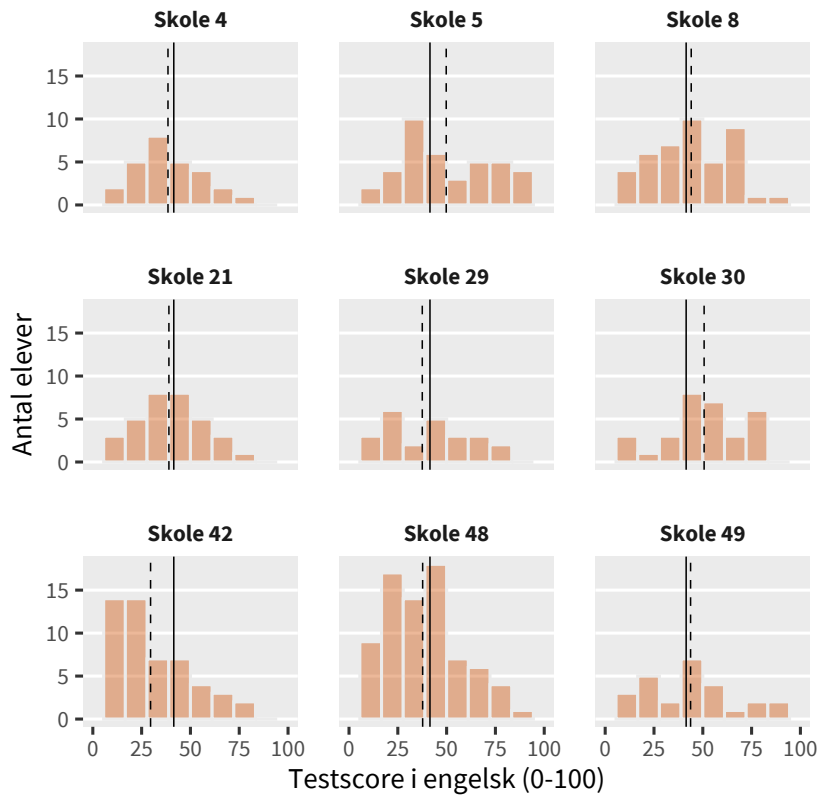
I figur 2 er den hierarkiske struktur illustreret. Analyseenhederne på niveau 2 er i eksemplet ovenfor skolerne, mens enhederne på niveau 1 er elever, der går på disse skoler.



Figur 2: Illustration af hierarkisk data med to niveauer.

Figur 3 viser, hvorfor det, ligesom med panledata, er væsentligt at forholde sig til den grupperede struktur, når man arbejder med hierarkisk data. I figuren ses fordelingen af testscorer i en engelskprøve på tværs af ni skoler. Den sorte streg angiver den gennemsnitlige

score på tværs af skoler, mens den stiplede streg angiver den gennemsnitlige score på den enkelte skole.



Figur 3: Fordelingen af testscorer i engelsk på ni skoler.
Den solide sorte streg viser den gennemsnitlige testscore på tværs af skoler, mens den stiplede streg viser den gennemsnitlige testscore på en skole.

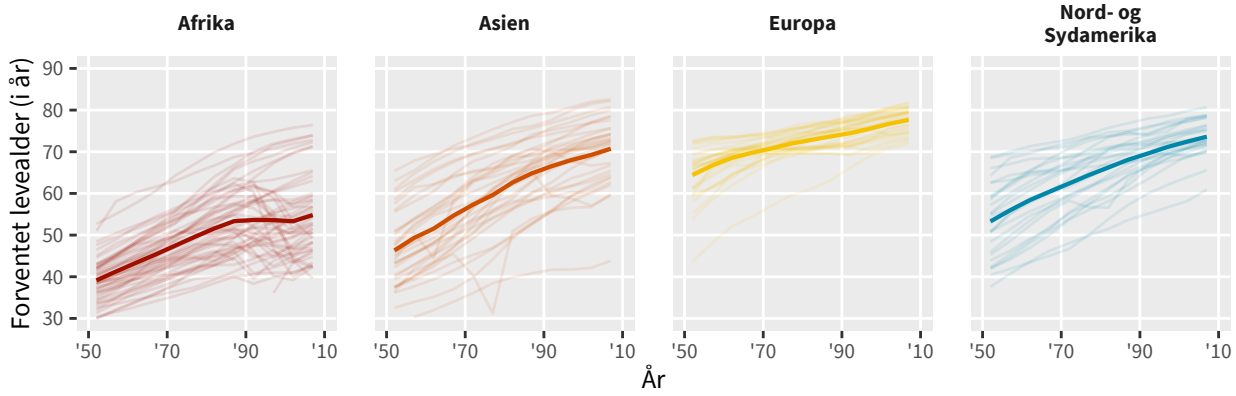
Ønskede vi at evaluere, hvordan det står til med engelskkundskaberne i en række skoler, kunne den gennemsnitlige testscore være et interessant mål. Det kan vi udregne med nedestående formel:

$$\begin{aligned}\hat{\mu}^{Testscore} &= \frac{1}{n}\sum_{i=1}^{n} Testscore_i \\ &= 41.5\end{aligned}$$

Tallet, vi får ud, er den gennemsnitlige testscore for alle elever på tværs af skoler. Dette tal vil sjældent være et retvisende billede af niveauet på den enkelte skole, fordi det gennemsnitlige niveau kan variere en del fra skole til skole. For nogle skoler (f.eks. skole 4, 21 og 49) er det overordnede gennemsnit retvisende for den gennemsnitlige elev på skolen. Andre skoler har derimod et gennemsnitligt niveau der ligger væsentligt over (f.eks. skole 30) eller under (f.eks. skole 42). Variationen mellem skoler indikerer, at der måske er systematiske forskelle skolerne imellem. En evaluering af elevernes faglige niveau bør derfor tage højde for, at de enkelte elever er indlejret i skoler.

Denne hierarkiske struktur kan genfindes mange steder. Et andet og meget klassisk eksempel er respondenter i tværnationale undersøgelser. Forestil dig, at vi ville vide noget om befolkningens tillid til de europæiske institutioner. For at undersøge dette, trækker vi en stikprøve med respondenter fra Danmark, Tyskland, Italien og Ungarn og stiller dem alle de samme spørgsmål. Der er stor sandsynlighed for at data fra denne undersøgelse ville vise, at danskerne ligner hinanden mere, end de ligner de andre, simpelthen fordi de er danskere, bor i Danmark, har de danske institutioner som deres referenceramme etc. Andre eksempler er mennesker indlejret i husholdninger, byrådsmedlemmer indlejret i kommuner, og medarbejdere indlejret i virksomheder eller organisationer. Når man først har fået øjene op for dette, vil man se, at rigtig meget af det data, vi beskæftiger os med i statskundskaben, har en eller anden form for grupperet struktur.

Her er en væsentlig pointe, at fordi grupperinger er så almindeligt forekommende, kan data også let være grupperet på flere niveauer end blot to eller være grupperet både i et hierarki og over tid på én gang. Figur 4 viser det samme data som figur 1, men for 140 lande fordelt på fire kontinenter. Her er det klart, at hvert enkelt lands gennemsnitlige levealder på et tidspunkt er afhængig af tidligere års forventede levealder, men *også* at udviklingen i lande på samme kontinent er stærkt forbundne. Borgere i de fleste europæiske lande har som udgangspunkt en længere forventet levetid i 1952 end i Asien, men også en mindre forbedring af levealderen over tid. Data kan derfor siges at være grupperet, både fordi lande indgår i forskellige fælles kontekster (kontinenternes fælles historie) og er observeret over tid.



Figur 4: Forventet levealder i 140 lande på fire kontinenter.
De tykke linjer angiver udviklingen i den gennemsnitlige levealder for hvert kontinent, mens de tynde streger angiver udviklingen for hvert enkelt land.

Netop fordi grupperinger findes alle steder og på mange niveauer, er det vigtigt, at man bruger sin teoretiske viden om emnet og sit

forskningsspørgsmål som guideline for, hvilke grupperinger, det er relevant at være opmærksom på i sin analyse.

Eksperimenter og kvasieksperimenter

Fra de ovenstående afsnit er det tydeligt, at observationelt data kan være grupperet i tid og rum. Det kan eksperimentielt data også, og nedenfor vil vi ufolde, hvad man skal være særligt opmærksom på i forhold til grupperinger i eksperimentielt data.

Eksperimenter er kendetegnet ved, at deltagere tilfældigt tildeles enten treatment eller kontrol status. Den tilfældige tildeling sikrer (hvis stikprøven er stor nok jf. store tals lov), at treatment og kontrolgruppen ikke er systematisk forskellige i udgangspunktet. Dermed kan vi estimere effekten af treatment ved at estimere forskellen i outcome mellem treatment og kontrolgruppen.

I samfundsvidenskaben er der dog en række forhold, vi ikke kan manipulere på individniveau, selvom vi er interesserede i, hvordan de virker på individet. Det gælder for eksempel alle de spørgsmål, vi måtte have omkring betydningen af institutionelle forhold, for eksempel, hvad betyder lærerkvalifikationer for elevernes læring? Hvad betyder en øget autonomi i forvaltningen for medarbejdernes sygefravær? Eller hvilken effekt har antikorruptionstiltag på tilliden til de offentlige institutioner?

Vil man lave eksperimenter, der kan belyse ovenstående spørgsmål, er man nødt til at acceptere, at alle elever i den samme klasse får den samme treatment - de har alle den samme lærer, ligeledes er graden af autonomi i forvaltningen nødt til at være den samme for alle medarbejdere i den samme forvaltning, og antikorruptionstiltag er nødt til at implementeres på instutitions-, regions- eller måske endda landeniveau for at kunne virke. Treatment vil altså i disse (og mange andre) tilfælde ikke kunne randomiseres til de individer, vi er interesserede i at måle effekten på, men i stedet tildeles på gruppeniveauet, niveau 2.

I tråd med dette, kan man også forestille sig eksperimenter, hvor man i princippet kan tildele treatment på individniveau, men hvor det er meget besværligt at administrere, eller hvor risikoen for at nogle respondenter modtager en anden treatment, end den de burde, er stor, og man derfor vælger at tildele treatment på niveau 2 alligevel.

Eksperimenter designes derfor ofte bevidst med en grupperet struktur. Data, der opstår som følge af naturlige eksperimenter eller kvasieksperimenter, er også typisk grupperet, selvom disse ikke er designet i et forskningsøjemed. Kvasieksperimenter og naturlige eksperimenter er i stedet karakteriseret ved at treatment skyldes

udefrakommende forhold, såsom lodtrækninger, migration, naturkatastrofer, ændret lovgivning, konkurser osv. Mange af de udefrakommende forhold, vi kan være interesserede i at undersøge effekten af, rammer ikke vores observationer på niveau 1 tilfældigt, men kan nogle gange antages at ramme observationer på niveau 2 tilfældigt. Et eksempel på dette er Card og Krugers 1994 berømte undersøgelse af effekten af mindstelønninger, hvor en ændring i lovgivningen i New Jersey berørte alle medarbejdere på fastfoodrestauranter i New Jersey, men ikke havde betydning for medarbejdere i fastfood restauranter lige på den anden side af Delaware River i Pennsylvania. Det er ikke tilfældigt, at alle medarbejere på den samme restaurant oplever den samme lovgivning, men man kan argumentere for, at det er tilfældigt, om restauranten ligger lige på den ene side eller lige på den anden side af floden, altså at det er tilfældigt på niveau 2.

Der kan således være flere grunde til, at treatment ikke tildeles på det niveau, vi er interesserede i at måle effekten på. Det behøver ikke være et problem i sig selv - nogle gange er det faktisk en forudsætning for at undersøge de spørgsmål, vi er interesserede i, men det har nogle konsekvenser for vores analyse, som vi vil introducere nedenfor.

Udfordringer med grupperet data

Grupperet data kan skabe en række udfordringer for vores analyse. Nedenfor vil vi forsøge at forklare, hvorfor grupperinger kan medføre, at vi fejlestimerer vores koefficienter, hvorfor grupperinger har betydning for estimeringen af standardfejl og sidst, hvordan grupperinger kan svække undersøgelsers eksterne validitet.

Selektion og misspecifikation af regressionsmodeller

Prøv at tænke tilbage til eksemplet fra figur 3. Hvad kunne vi gøre, hvis vi ønskede at undersøge effekten af elevernes køn på deres testscore i engelskprøven? Et naturligt udgangspunkt ville være, at estimere en simpel regressionsmodel:

$$\hat{Y}_i^{Testscore} = \underbrace{\hat{\beta}_0}_{\substack{\text{Gennemsnitlig} \\ \text{testscore for} \\ \text{piger}}} + \underbrace{\hat{\beta}_1}_{\substack{\text{Gennem-} \\ \text{snitlig} \\ \text{kønsforskel}}} \times \text{Køn}_i + \hat{u}_i$$

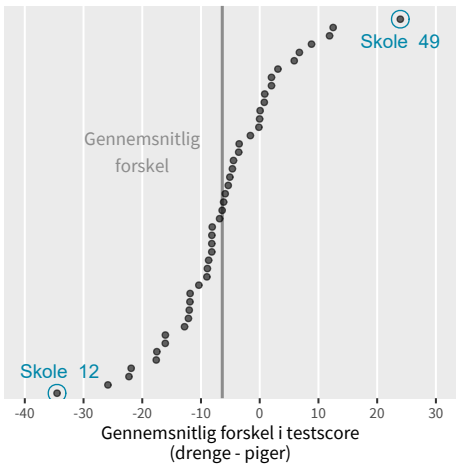
hvor $\hat{\beta}_0$ er den gennemsnitlige testscore for piger, $\hat{\beta}_1$ er forskellen i testscore mellem piger og drenge og $\hat{u}_i$ er fejlleddet. Vi så ovenfor, at vores data bestod af elever fra en række skoler med meget forskelligt *gennemsnitligt* niveau af engelskkundskaber. Samtidigt er der også en stor forskel på kønssammensætningen mellem skolerne.

Skole 21 har kun omkring 38 % piger, mens skole 30 har hele 68 %. Vores simple regression lider derfor af det problem, at variationen i testscorer mellem piger og drenge både skyldes forskelle internt på skolerne, men også mellem skoler med forskelligt niveau og forskellig kønssammensætning. Vi kan derfor ikke på baggrund af modellen afgøre, om forskellen alene skyldes forskelle i evner mellem piger og drenge, eller at skoler med lavere gennemsnitligt niveau, måske også har en overrepræsentation af elever med et bestemt køn. Vores problem skyldes i høj grad, at vi ikke har indtænkt alle relevante faktorer i vores model, fordi der kunne være selektionsbias på vores gruppeniveau. Specifikt kan vi forestille os, at faktorer, som påvirker variationen i testscorer på tværs af skoler også påvirker kønssammensætningen. Sammenligner vi andelen af piger og den gennemsnitlige testscore på tværs af 47 skoler[4] finder vi også en moderat korrelation på $r = 0.38$. Noget tyder altså på, at der er en moderat korrelation mellem kønssammensætningen på den enkelte skole og det gennemsnitlige faglige niveau.

Tager vi ikke højde for selektionsbias på gruppeniveauet risikerer vi, at vores resultater fra den simple lineære model er et udtryk for "Simpson's Paradox". I Simpson's Paradox kan en sammenhæng mellem to variable se enten positiv eller negativ ud, hvis vi sammenligner data på tværs af grupper, men være enten ikke-eksisterende eller med omvendt fortegn, når vi sammenligner inden for grupper. Figur 5 viser, at dette kan være tilfældet i vores skoledata.

Den gennemsnitlige forskel mellem drenge og piger ($\hat{\beta}_1$) er -6.4 ($p \leq 0.001$), således at piger i gennemsnit scorer over seks point mere i deres engelskprøve. Men for en lang række af skoler er den forskel mindre, større eller med modsat fortegn. På skole 49 scorer drengene i gennemsnit 23.9 point højere, mens pigerne på skole 12 er næsten 35 point bedre end drengene. Gruppestrukturen, hvor eleverne går på skoler med forskelligt niveau og kønsssammensætning skjuler altså, at en del af forskellen mellem drenge og piger skyldes forskelle mellem skoler, som vi ikke har taget højde for i vores model. En anden måde at anskue denne problematik på er, at skolernes kønssammensætning korrelerer med kønsforskellen i testscore. Hvis ikke vi tager højde for gruppestrukturen, så indgår denne i modellens fejlled og vores uafhængige variabel, køn, kommer derfor til at korrelere med fejlleddet.

[4] De ni skoler i figur 3 er blot et tilfældigt udsnit af en større undersøgelse blandt 1119 engelske skoleelever i 47 skoler.



Figur 5: Gennemsnitlig forskel i testscore mellem drenge og piger på 47 skoler. *Den solide sorte streg viser den gennemsnitlige testscore på tværs af skoler.*

Afhængige observationer og forkerte standardfejl

En anden vigtig, statistisk, motivation for at tænke over mulige gruppestrukturer i ens data handler om usikkerheden forbundet med vores estimater. Hvis data er grupperet på den ene eller den anden

måde, er der risiko for, at observationer af en variabel er kollerede med hinanden inden for grupperingerne. Når vores observationer korrelerer inden for grupper, vil dette ofte betyde, at fejlledene i vores regressionsanalyse også er korrelerede inden for grupperne. Har vi data over tid, omtaler vi ofte dette som autokorrelation. I dette ligger at observationer tæt på hinanden i tid har fejlled, der ligner hinanden. Med hierarkisk data er strukturen på afhængigheden mellem vores fejlled givet ved grupperingen - der er en risiko for, at fejlledene for observationer i den samme gruppering ligner hinanden. Hvis fejlleddene korrelerer indenfor grupperingerne, betyder det, at vores observationer ikke er uafhængige.

Når vores observationer er afhængige af hinanden i tid eller rum, betyder det, at hver ny observation bidrager med mindre ny information, end hvis den havde været uafhængig af de andre observationer. Tager vi ikke højde for denne afhængighed, vil vores estimater af standardfejlen blive for små, og dermed vil p-værdier også blive or små og vores konfidensintervaller for smalle. Det betyder, at vil vil overvurdere det inferentielle potentiale. Med andre ord kommer vi til at have større tiltro til, at vores resultater i stikprøven kan infereres til populationen, end vi har belæg for.

Afhængighed mellem vores observationer har således ikke betydning for vores koefficienter, så længe afhængigheden mellem vores fejlled er ukorreleret med værdierne på vores uafhængige variable. Dog kan afhængighed mellem vores observationer have ret store konsekvenser for vores estimat af standardfejlen, altså koefficienternes usikkerhed, og dermed for, hvor stor tiltro vi kan have til, at vores estimat i stikprøven kan infereres til den underliggende population.

Kausal heterogenitet og generaliserbarhed

En sidste problematik omhandler vores evne til at generalisere fra en kausaleffekt observeret i én kontekst til en anden. Når forskere og andre gennemfører randomiserede, kontrollerede forsøg, så foregår disse ofte i en bestemt kontekst, for eksempel i et bestemt land eller på en bestemt type forsøgspersoner. Nogle gange foretages forsøg endda i foreskellige kontekster og analyseres måske samlet i én analyse. Her kan problemet være, at en kausaleffekt måske kun er til stede i nogle typer af kontekster, eller at effekten varierer på tværs af de grupper, hvori forsøget er foretaget.

Et eksempel på det kommer fra fattigdomsbekæmpelse, hvor organisationer i samarbejde med en række økonomer har undersøget effekten af mikrolån til primært kvindelige iværksættere i udviklingslande. Idéen er, at let adgang til små lån vil øge sandsynligheden for, at fattige husholdninger starter virksomheder. Problemet er dog, at

de enkelte forsøg med mikrolån er implementeret i meget forskellige lande såsom Bosnien, Etiopien, Indien, Mexico, Mongoliet og Marokko, og resultaterne af de enkelte forsøg viser både positive, ingen og negative effekter på husstandenes profit og forbrug (Pritchett & Sandefur, 2015).

En samlet evaluering af de forskellige forsøg ville vise et meget mudret billede af effekten af mikrolån på fattigdomsbekæmpelse, hvis ikke vi tager højde for konteksten omkring det enkelte forsøg. Her er altså et eksempel på, at kausaleffekten er *heterogen* og varierer på tværs af grupper af forsøgspersoner indlejret i forskellige kontekster. I sådanne tilfælde er gruppering i data ikke nødvendigvis et problem for den interne validitet. Vi kan stadig estimere den gennemsnitlige kasaule effekt selvom effekten nogle steder er positiv, nogle steder negativ og andre steder helt fraværende. Gruppestrukturen kan dog udgøre et problem for den eksterne validitet, fordi generaliserbarheden af et forsøg er usikker. Uanset hvad er det godt at kende til en eventuel kausal heterogenitet på tværs af grupper for i hvert fald at kende begrænsningerne ved de interventioner, man kunne ønske at bruge i andre kontekster (Meager, 2019).

Håndtering af grupperet data

En grupperet datastruktur kan altså have en række uønskede konsekvenser for vores resultater, med mindre vi aktivt forsøger at tage højde for disse. For det første udgør grupperet data en udfordring for vores evne til at kvantificere usikkerheden i vores estimater. Grupperingerne kan lede til forkerte standardfejl - et centralt element i udregningen af både p-værdier og konfidensintervaller. For det andet kan gruppering i data medføre selektionsproblemer, der kan forhindre os i at estimere en retvisende kausal effekt. For det tredje kan grupperinger svække undersøgelsen eksterne validitet. I dette afsnit introducerer vi først klyngerobuste standardfejl som en mulig løsning på fejlestimerede standardfejl. Dernæst introducerer vi to måder at inkludere eller tage højde for gruppering i data gennem modellering med henblik på at minimere selektionsproblemer: Fixed effects- og multilevelmodeller. Vi præsenterer ikke løsninger på udfordringer relateret til den eksterne validitet, men bevidstheden om kausal heterogenitet er relevant at have i baghovedet, når man konkluderer på sine analyse af grupperet data.

Klyngerobuste standardfejl

Som beskrevet ovenfor vil en grupperet struktur i data medføre, at hver observation bidrager med mindre ny information, end hvis

observationerne var uafhængige. Det betyder også, at der er en risiko for, at vores standardfejl bliver forkerte. For at korrigere for denne afhængighed kan man anvende klyngerobuste standardfejl, der en særlig type af robuste standardfejl.

Robuste standardfejl korrigerer for manglende varianshomogenitet. Er der ikke varianshomogenitet vil den estimerede standardfejl være for stor for nogle værdier af x og for lille for andre værdier af x. De robuste standardfejl tillader, at variansen er forskellig for forskellige værdier af x, og giver os dermed (såfremt modellen ellers er specificeret korrekt) et retvisende estimat af standardfejlen, selvom der ikke er varianshomogenitet (Stock & Watson, 2020, afsnit 5.4)

De klyngerobuste standardfejl er en særlig variant af robuste standardfejl, der udover at korrigere for variansheterogenitet også korrigerer vores estimat af den statistiske usikkerhed for afhængighed mellem observationerne. De klyngerobuste standardfejl tillader korrelation mellem fejlleddene indenfor grupperingerne i data og korrigerer dermed for, at hver observation bidrager med mindre information om den sammenhæng, vi er interesserede i, end hvis observationerne havde være uafhængige. Lidt karikeret kan man sige, at de klyngerobuste standardfejl "renser"observationerne for den information, der er forklaret af de andre observationer i gruppen og dermed estimerer standardfejlen på baggrund af den uafhængige variation, hver observation bidrager med. De klyngerobuste standarfejl er en matematisk kompliceret størrelse, men hvis man ønsker at forstå dem på et dybere plan, kan man læse om dem i Stock og Watson, 2020 appendix 10.2 og afsnit 16.4.

Indtil for nyligt var tommelfingerreglen, at man kunne lave sine analyser med både almindelige og klyngerobuste standardfejl, og så afrapportere de største standardfejl (Cameron & Miller, 2015). På den måde ville man være sikker på, at man ikke overvurderede sikkerheden på sine estimater.

Af to grunde er vores anbefaling dog en smule mere nuanceret. For det første, og dette relaterer sig til både de robuste og de klyngerobuste standardfejl, vidner en stor forskel mellem almindelige standardfejl og robuste standardfejl ikke nødvendigvis blot om grupperinger i data. De kan også være et tegn på, at der er noget underliggende galt med vores modelspecifikation, som vi burde håndtere i stedet for blot at anvende robuste standardfejl (King & Roberts, 2015). Det kunne eksempelvis være, at den funktionelle form er forkert, det vil sige, at vi modellerer en sammenhæng linæert som ikke er lineær, at en eller flere variable bør tranformeres, eller at en væsentlig variabel er udeladt af modellen.

For det andet, og mere specifikt i relation til de klyngerobuste standardfejl, risikerer vi ved brug af klyngerobuste standardfejl at

gøre vores estimater alt for konservative. Det betyder, at vi risikerer at forkaste sammenhænge, som faktisk findes i den population vi gerne vil inferere til, og derved begrænser vi vores egen mulighed for at blive klogere. Klyngerobuste standardfejl bør derfor ikke bruges, hver gang data er grupperet, men i stedet i de tilfælde, hvor data har et af nedenstående karakteristika (Abadie m.fl., 2023):

- Data, hvor den uafhængige variabel, vi er interesserede i at måle effekten af, er tildelt på niveau 2
- Data, hvor den uafhængige variabel, vi er interesserede i at måle effekten af, er på niveau 1, og stikprøven kun indeholder observationer fra en mindre del af grupperne i den population, vi gerne vil sige noget om. For eksempel hvis vi gerne vil udtale os om uddannelsespolitiske tiltag og vores stikprøve kun indeholder elever fra få skoler.
- Data, hvor den uafhængige variabel, vi er interesserede i at måle effekten af, er på niveau 1, og stikprøven indeholder mange grupperinger fra den population vi gerne vil inferere til, men kun en mindre del af hver grupperings enheder. Det kunne være data fra mange lande, men kun få borgere fra hvert land.

Omvendt, hvis vores stikprøve indeholder observationer fra de fleste af grupperne i den population vi gerne vil inferere til, eller hvis vi udtrækker størstedelen af de mulige observationer i hver gruppe i stikprøven, er der en risiko for, at brugen af klyngerobuste standardfejl giver os uforholdsmæssigt store standardfejl.

Fixed effects modeller

Klyngerobuste standardfejl kan således løse udfordringen med at estimere standardfejlen, når vores observationer ikke er uafhængige, men hvis grupperingerne også skaber selektion i den model, vi gerne vil estimere, kan vi ikke løse det med klyngerobuste standardfejl. I stedet kan vi inkludere grupperingen i vores model. Den mest simple af sådanne modeller kaldes *fixed effects estimatoren*, fordi den fastholder (fixer) eller holder konstant variationen mellem grupper og kun efterlader variationen inden for grupperne. Fixed effects estimatoren analyserer udelukkende variation indenfor grupperne, mens alle variable på niveau 2, både observerede og uobserverede, holdes konstant. Derved undgår vi at bekymre os om, hvordan vores observationer selekterer ind i den ene eller den anden gruppe, fordi vi kontrollerer for alt på gruppeniveau.

Det er her væsentligt at huske på, at vi med "grupper"både referer til reelle grupper af eksempelvis elever i en skole eller borgere i et

land, og abstrakte grupper som eksempelvis observationer af den samme person eller det samme land over tid.

Konkret estimeres fixed effects modeller som en velkendt, lineær sammenhæng, med et forskelligt skæringspunkt med y-aksen (α_i) for hver gruppe (i) i data således:

$$Y_{it} = \underbrace{\beta_1}_{\substack{\text{Koefficient} \\ \text{for variabel}}} X_{it} + \underbrace{\alpha_i}_{\substack{\text{Gruppespecifik} \\ \text{konstant}}} + u_{it}, \tag{1}$$

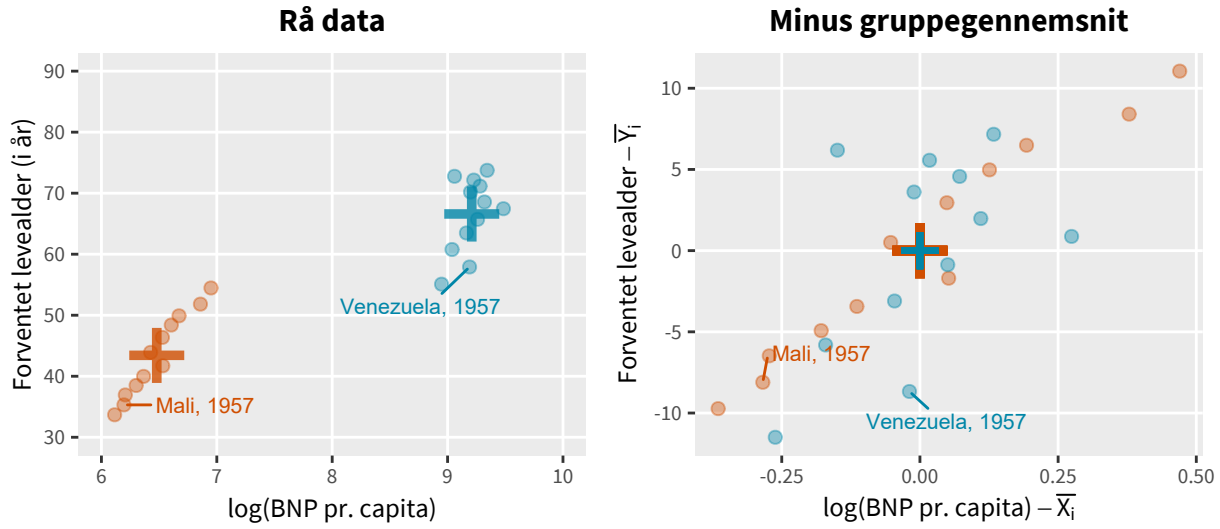
hvor β_1 er koefficienten for variable, der også varierer inden for grupper. Hvis grupperingen her er lande ($i = 1, \ldots, N$) som er observeret over tid ($t = 1, \ldots, T$), så eliminerer fixed effects estimatoren al variation mellem lande, som ikke også varierer over tid. Det kunne f.eks. være deres geografisk placering, hvilket kontinent landet er placeret på, eller hvorvidt landet har været koloniseret.

α_i repræsenterer således alle de både *uobserverede* og observerede, tidsinvariante, effekter som kunne tænkes at påvirke både Y_{it} og X_{it}. Der findes flere metoder til at estimere model (1). Hvis vi kun er interesserede i at estimere effekten af X_{it}, når denne ændres indenfor lande, så kan vi formulere fixed effects estimatoren som en *within group transformation* således:

$$Y_{it} - \bar{Y}_i = \beta_1(X_{it} - \bar{X}_i) + (\alpha_i - \bar{\alpha}_i) + (u_{it} - \bar{u}_i) \tag{2}$$

hvor f.eks. $\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$, det vil sige gennemsnittet for hvert enkelt land i. Vi trækker således gruppegennemsnittene fra hver enkelt led i vores model for at eliminere den variation, som måtte være mellem grupperne. For vores gruppespecifikke konstant α_i er værdien for α_i og $\bar{\alpha}_i$ den samme således, at $(\alpha_i - \bar{\alpha}_i) = 0$. Det er derfor man kalder modellen for en *fixed effects model*, da den netop fastholder den gennemsnitlige forskel mellem grupperne på 0. Figur 6 viser et eksempel på denne logik for to af landene i figur 1 og sammenhængen mellem deres BNP pr. capita (logaritmetransformeret) og den forventede levealder.

På venstre side ser vi de rå data med krydser, som angiver hvert lands gennemsnit på de to variable (dvs. $\bar{X}_i$ og $\bar{Y}_i$), mens det højre plot viser variablene efter de er fratrukket de respektive landegennemsnit. Læg mærke til, at det røde og blå kryds nu overlapper og er centreret omkring nul. Vi har altså fixet variationen mellem lande til at være 0 og kan nu estimere vores model på data, der kun varierer inden for lande. På den måde har vi taget højde for eventuelle forskelle i den gennemsnitlige BNP pr. capita og forventet levealder mellem lande. Vores estimat af effekten af at øge BNP pr. capita,



Figur 6: Sammenhængen mellem log(BNP pr. capita) og forventet levealder i Mali og Venezuela.
Hvert punkt angiver en årlig observation og krydser angiver hvert lands gennemsnitlige værdier på x- og y-aksen.

$\hat{\beta}_{FE} = \hat{\beta}_1(X_{it} - \bar{X}_i)$, er altså kun effekten ved at øge inden for et givent land, i gennemsnit.

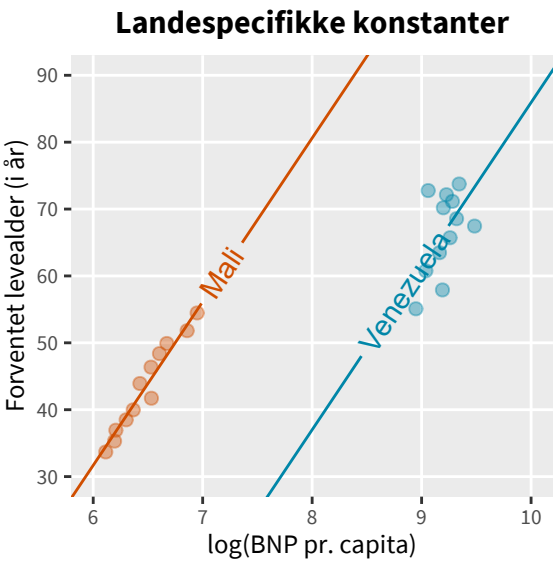
En anden måde at konstruere fixed effects estimatoren på, er ved specifikt at estimere en konstant for hver gruppe i vores data. Konkret kan vi inkludere et sæt af dummyvariable (d_1, ...$d_I - 1$), en for hver gruppe pånær én. På den måde får vi et estimat for effekten af grupperingen (hver koefficient for d_i), og konstanten bliver den gennemsnitlige værdi af Y_i for det udeladte land. Modellen vil i denne situation se lidt anderledes ud end model (1), men resultaterne vil være de samme:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 d_1 + \beta_3 d_2, ..., \beta_j d_{I-1} + u_{it}, \tag{3}$$

hvor β_0 er gennemsnittet for referencegruppen, mens koefficienterne for hver af dummierne d_1 til d_{I-1} angiver den enkelte gruppes afvigelse fra referencegruppen. Således angiver $\beta_0 + \beta_j d_i$ den gennemsnitlige værdi af Y for obsevatioerne i gruppe i. Figur 7 viser logikken for Mali og Venezuela igen.

Hvert land får her en specifik konstant, hvor Malis er $\beta_0 = -115.09$ og Venezuelas er $\beta_0 + \beta_2 = -155.09 + (-157.7) = -273.8$, mens hældnigen $\beta_1 = 24.5$ er den samme. Vi antager dermed, at effekten er den samme i alle landene, men vi tager højde for, at hvert land har sit eget udgangspunkt. Hvis vi estimerede den bedste rette linje for punkterne i det højre plot i figur 6 ville få præcis den samme hældningskoefficient, og dermed effekt af at øge log(BNP pr. capita) med én.

Selvom de to metoder giver samme resultat, så vil man ofte benyt-



Figur 7: Sammenhængen mellem log(BNP pr. capita) og forventet levealder i Mali og Venezuela.
Hvert punkt angiver en årlig observation og linjerne viser den bedste rette linje for hvert land, med hver sin konstant men samme hældning.

te den første. Det skyldes primært, at det er tungere for computeren jo flere grupper, og dermed koefficienter, der skal estimeres. Desuden er outputtet fra en fixed effects model mere overskueligt, fordi der ikke indgår en koefficient for hver gruppe i data. Begge metoder er dog lige rigtige.

Multilevel modeller

Fixed effects modeller er et stærkt redskab til at håndtere grupperet data, men de har også den ulempe, at de smider meget variation i data væk, fordi de udelukkende estimerer variationen indenfor grupperne. Derved har fixed effects modeller ikke særlig stor statistisk power - det betyder, at der skal flere observationer til, før vi kan foretage en meningsfuld statistisk test af, om den sammenhæng vi leder efter, er til stede eller ej. Dertil kommer, at fixed effects modeller ikke kan estimere effekten af variable på gruppeniveau. Der er derfor mange situationer, hvor en fixed effects model ikke er optimal til at modellere grupperingen i data.

Heldigvis er der også løsninger på det. Der er multilevel modeller (også kaldet hierarkiske modeller), der kan bruges til at estimere effekten af niveau 2 variable. Lidt simpelt fortalt "låner"disse modeller information fra alle grupperne i data, når de estimerer effekten i hver enkelt gruppe, og dermed giver sådanne modeller os både et estimat af effekten indenfor grupperne og for effekten af variable mellem grupperne. Multilevel modeller har ydermere den fordel, at de er mere efficiente, idet de udnytter en større del af variationen i data. Dermed bliver vi uden at få flere observationer i stand til at estimere

de sammenhænge vi leder efter mere præcist - vi får større statistisk power.

En multilevel model kan specificeres meget på samme måde som en fixed effects model:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it} \tag{4}$$

Her er α_i igen den gruppespecifikke konstant, som kan variere over grupperne $i = 1, 2, ..., I$. Den store forskel ligger dog i, hvordan vi behandler α_i. Husk på, at vi i fixed effects modellen fastholder den gennemsnitlige forskel mellem grupper på nul. I multilevelmodeller er vi i stedet interesseret i at estimere både variationen mellem og indenfor grupperne. Vores gruppespecifikke konstant består derfor af to led:

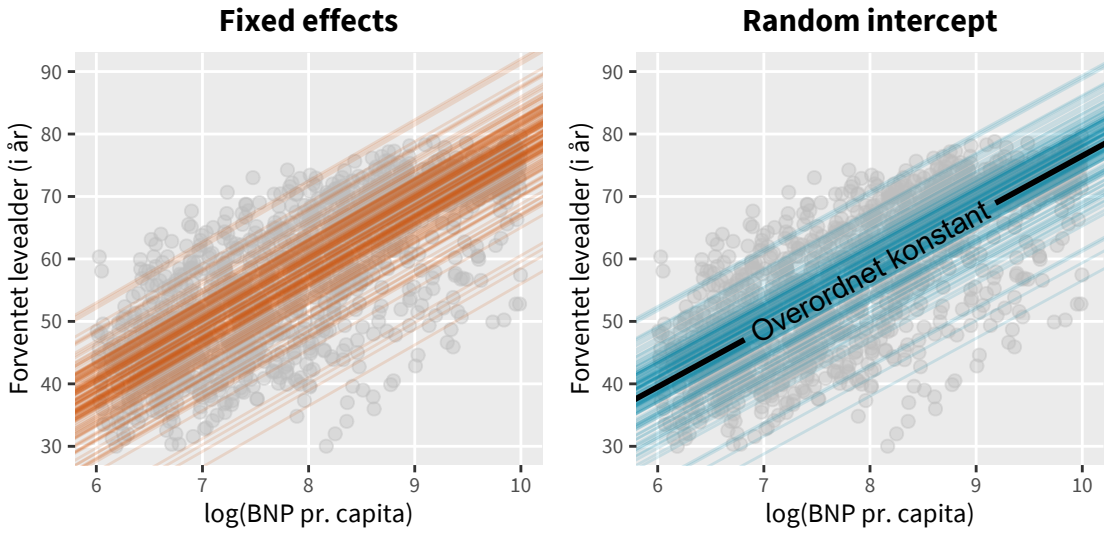
$$\alpha_i = \beta_0 + \gamma_i \tag{5}$$

$$\gamma_i \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

hvor β_0 er den overordnede konstant som vi kender fra en almindelig lineær model[5], og hvor γ_i er afvigelsen fra den overordnede konstant for hver gruppe i. Udtrykket $\gamma_i \sim \mathcal{N}(0, \sigma^2)$ beskriver en central antagelse vi gør os: At gruppernes afvigelser fra den overordnede konstant er normalfordelt med gennemsnittet 0 og variansen σ^2. Hver enkelt gruppes konstant vil altså være $\beta_0 + \gamma_i$, dvs. den overordnede konstant plus gruppens afvigelse. En sådan model kaldes ofte en *random intercept model*, fordi vi lader hver enkelt gruppes konstant variere omkring den overordnede konstant. Figur 8 viser en sammenligning mellem vores fixed effects og random intercepts modeller. I begge modeller har vi igen fokuseret på at estimere sammenhængen mellem BNP pr. capita (logaritmetransformeret) og forventet levealder for 142 lande over en periode på næsten 60 år. Fordi hver observation er en årlig måling *indenfor* et land, har vi i begge modeller estimeret en konstant pr. land. Begge modeller giver os meget ens resultater, hvor hvert enkelt land har varierende udgangspunkter. Dog tillader vores multilevel model (random intercepts) os at estimere en overordnet konstant på tværs af lande.

[5] Dvs. gennemsnittet af Y_{it} når alle uafhængige variable er 0.

Formålet er her, at vi kan estimere en effekt på tværs og inden for grupper, fordi modellen "låner"information omkring variationen mellem grupper til at estimere effekten inden for grupper. Dette kaldes ofte *partial pooling*. Partial pooling hjælper os i situationer med et lavt antal observationer pr. gruppe, eller hvor der er stor forskel i antallet af observationer på tværs af grupperne. Det kunne være situationer, hvor vi mangler data for visse grupper, eller hvor der er naturlig variation i gruppestørrelse. En spørgeskemaundersøgelse blandt alle

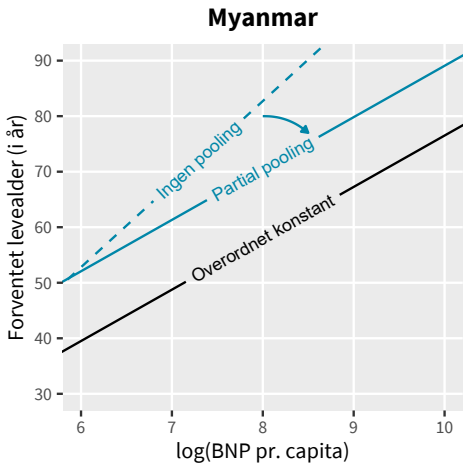


Figur 8: Fixed effects og random intercepts modeller af sammenhængen mellem log(BNP pr. capita) og forventet levealder.
Hvert punkt angiver en årlig observation og linjerne viser den bedste rette linje for hvert land, med hver sin konstant men samme hældning.

9. klasser på danske folkeskoler ville sikkert vise forskel i klassestørrelser (f.eks. mellem mindre landsbyskoler og større byer), og nogle klasser ville måske mangle mange svar, hvis nogle af klassens elever var syge. En multilevel model antager, at informationen fra én gruppe kan bruges til at sige noget substantielt om en anden gruppe, men samtidigt, at der kan være systematiske forskelle i mellem dem. Ved at delvist at "poole"information på tværs af grupper kan modellen korrigere for manglende information i nogle grupper.

For at se, hvordan partial pooling virker, kan vi sammenligne med en situation, hvor hver gruppe ikke bidrager med noget information på tværs af grupper. Det svarer til, at vi estimerede en simpel lineær model for hver enkel gruppe. Figur 9 viser igen et eksempel med BNP pr. capita og forventet levealder for Myanmar. Den stiplede linje er her vores bedste rette linje for en model med kun data fra Myanmar, mens de to andre linjer er den landespecifikke- og den overordnede bedste rette linje fra en random intercepts model med alle 142 lande. Fordi vores multilevelmodel delvist pooler information på tværs af lande, trækker den linjen for Myanmar mod en overordnet linje på tværs af lande, men bibeholder alligevel en landespecifik variation.

Ved delvist at poole information på tværs af grupper mister vi nogle af nuancerne omkring sammenhængen for de enkelte grupper. I Myanmar synes sammenhængen mellem BNP pr. capita og forventet levealder at være noget stærkere, end vores multilevelmodel viser (hældningen er mere stejl!). Men vi vinder et mere præcist estimat, fordi modellen baserer sig på mere information. Et 95 % konfidens



Figur 9: Sammenhængen mellem log(BNP pr. capita) og forventet levealder i Myanmar.
Den stiplede linje er den bedste rette linje for en simpel model med kun data fra Myanmar. De resterende linjer er henholdsvis den landespecifikke og den overordnede bedste rette linje fra en random intercepts model med alle 142 lande.

interval for hældningen i vores simple model er [-1.25, 31.17], mens det i multilevelmodellen er [8.77, 9.73]. Multilevelmodeller er derfor et værktøj, som forsøger at balancere hensynet til både præcision og systematiske forskelle mellem grupper i vores data.[6]

[6] En yderligere mulighed med multilevelmodeller er, at lade hver gruppe have sin egen specifikke hældning, sådan at også koefficienten for X_{it} varierer. Dette kaldes ofte en *random slopes model*. Sådanne modeller er dog også mere komplicerede at estimere og fortolke.

Litteratur

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When Should You Adjust Standard Errors for Clustering?*. *The Quarterly Journal of Economics*, *138*(1), 1–35.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372.

Card, D., & Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, *84*(4), 772–793. Hentet 30. januar 2023, fra https://ideas.repec.org//a/aea/aecrev/v84y1994i4p772-93.html

King, G., & Roberts, M. E. (2015). How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It. *Political Analysis*, *23*(2), 159–179. Hentet 13. januar 2023, fra https://www.jstor.org/stable/24572966

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, *11*(1), 57–91.

Pritchett, L., & Sandefur, J. (2015). Learning from Experiments When Context Matters. *American Economic Review*, *105*(5), 471–475.

Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics* (4th global). Pearson.

Data

Data til brug for denne notes illustrationer er hentet fra to forskellige kilder. Landedata er hentet fra R-pakken *gapminder* (Bryan, 2017), mens data for skoleelever i England er fra *R2MLwiN* (Zhengzheng m.fl., 2016).

Bryan, J. (2017). Gapminder: Data from Gapminder. https://CRAN.R-project.org/package=gapminder

Zhengzheng, Z., Parker, R. M. A., Charlton, C. M. J., Leckie, G., & Browne, W. J. (2016). R2MLwiN: A Package to Run MLwiN from within R. *Journal of Statistical Software*, *72*, 1–43.