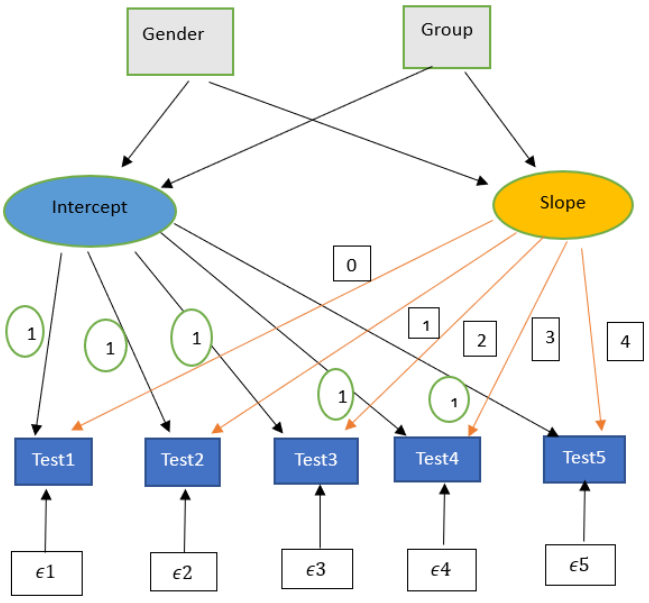
Longitudinel Data Modeling (Lineære Mixed Modeller og Latente Kurve Modeller)

Semesterprojekt, 10. semester (Specialet)



Hayam Alabd

Aalborg Universitet

Institut for Matematiske Fag
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

AALBORG UNIVERSITET
STUDENTERRAPPORT

Institut for Matematiske Fag
Aalborg Universitet
Skjernvej 4A
9220 Aalborg Øst
http://math.aau.dk

Titel:

Longitudinel Data Modeling (Lineære Mixed Modeller og Latente Kurve Modeller)

Projekt:

10. semester (Specialet)

Projektperiode:

1. febuar 2023 - 2. juni 2023

Projektgruppe:

STAT 10 4111c

Deltagere:

Hayam Alabd

Vejleder:

Rasmus Waagepetersen

Sidetal: 54 + forside
Afsluttet 2. juni 2023

Synopsis:

Dette semesterprojekt omhandler longitudinel data af grupper af elever fra folkeskole (4. klasse), som analyseres på baggrund af de statistiske metoder og resultater, som udledes og beskrives igennem projektet.
Formålet med dette er blandt andet at undersøge forskellen mellem lavt- og højt-præsterende elevers udviklingen af deres færdigheder i brøkregning i løbet et skoleår, og hvordan disse forskelle er relaterede til undervisningen af andre matematiske emner.
Dette gøres ved for eksempel at anvende en lineær mixed model, som er en statistisk metode, der tager højde for den indbyrdes afhængighed mellem målinger fra den samme elev over tid. Ved at tilpasse denne model til data, kan der vurderes den individuelle udvikling af elevernes færdigheder i brøkregning og eventuelle forskelle mellem lavt- og højtpræsterende elever identificeres.
Dernæst introduceres en alternativ metode til lineær mixed model. Det er *Latente kurve modeller*, som er en statistisk metode, der tillader at estimere de underliggende udviklingskurver for elevernes færdigheder i brøkregning baseret på deres målte præstationer over tid.
Slutteligt udføres en velovervejet statistisk analyse af datasættet.

Forord

Det indeværende semesterprojekt er udarbejdet af studerende i STAT-gruppen 4111c på tiende semester, på matematikuddannelsen ved det Ingeniør- og Naturvidenskabelige Fakultet på Aalborg Universitet, i foråret 2023. Projektets omfang er afgrænset af, at det henvender sig til matematikstuderende med godt kendskab til sandsynlighedsteori og statistik inferens.

Projektet følger konventionerne for Vancouver-formatet til referering af kilder. Alle kilder nummereres med et tal, [n], som da kan findes i litteraturlisten under dette nummer. Definitioner, sætninger, lemmaer, propositioner, eksempler, figurer og tabeller nummereres løbende efter kapitel. Ydermere vil det statistiske programmeringssprog R blive brugt til at udføre statistiske analyser i projektet.

Slutteligt ønsker projektets forfatter at rette en stor tak til Rasmus Waagepetersen, for behjælpelige råd og vejledning, samt til Institut for Matematiske Fag for at stille et grupperum til rådighed i projektperioden.

Underskrifter

Hayam Alabd

Indholdsfortegnelse

1 | Introduktion

Longitudinelle dataanalyser er en vigtig metode inden for forskning inden for medicin, psykologi og samfundsvidenskab. Ved at følge de samme individer over tid, kan longitudinelle data give en mere nuanceret forståelse af udviklingen af sygdomme, adfærdsmæssige mønstre og andre fænomener.
Dette projekt vil undersøge, hvordan lineære mixede modeller og latent growth modeller kan anvendes til at analysere longitudinelle data, og hvordan disse metoder kan bruges til at give en mere præcis og nuanceret forståelse af udviklingen af, at elevers færdigheder i brøkregning i folkeskoler udvikler sig over tid. Projektet vil også undersøge, hvilke faktorer der kan påvirke udviklingen af elevers færdigheder i brøkregninger over tid. I projektet vil vi benytte longitudinelle data for en gruppe elever i folkeskolen, der er blevet fulgt i løbet af et skoleår. Eleverne var testet i brøkregning færdigheder på 5 tidspunkter, og deres udvikling er blevet registreret og analyseret ved hjælp af lineære mixede modeller og latent growth modeller. Projektet har undersøgt, hvordan elevernes færdigheder i brøkregning udvikler sig over tid, og hvilke faktorer der har påvirket denne udvikling; herunder deres køn, andre matematiske færdigheder, skoler, klasser og undervisningsmetoder.

Lineære mixede modeller og latent growth modeller er avancerede statistiske metoder, der er specielt designet til at analysere longitudinelle data. Lineære mixede modeller tager højde for både faste og tilfældige faktorer i modellen, og er en alsidig analysemetode, der kan anvendes til at analysere en bred vifte af data; herunder gentagne målinger, longitudinelle data.

Latent growth modeller er en type af lineære mixede modeller, der er specielt udviklet til at analysere udviklingen af latente variabler over tid. Ved at anvende latent growth modeller kan forskere undersøge, hvordan en given latent variabel udvikler sig over tid, og hvilke faktorer der kan påvirke denne udvikling.

Resultaterne af projektet kan bruges til at identificere effektive undervisningsmetoder og interventioner, der kan hjælpe eleverne med at forbedre deres færdigheder i brøkregninger over tid. Projektet kan også bidrage til en mere omfattende forståelse af, hvordan matematiske færdigheder udvikler sig over tid, og hvilke faktorer der spiller en rolle i denne udvikling.

Mere specifikt består projektet af to kapitaler. Det første kapitel i projektet vil fokusere på lineære mixede modeller og deres anvendelse i analysen af longitudinelle data. Kapitlet vil først introducere konceptet af lineære mixede modeller og forklare, hvordan de anvendes til at analysere longitudinelle data. Der vil også gives eksempler på, hvordan disse modeller kan bruges. Estimationsmetoder vil også blive diskuteret i dette kapitel; herunder maximum likelihood (LM), restricted maximum likelihood (REML) og penalized likelihood (PLUP). Endelig vil kapitlet omhandle den faktiske dataanalyse, hvor lineære mixede modeller anvendes til at

analysere datasættet om færdigheder i brøkregning hos skoleelever. Analyseresultaterne vil blive diskuteret, og deres implikationer for undervisning og pædagogik vil blive overvejet.

Det anden kapitel i projektet vil fokusere på latent growth modeller og deres anvendelse i analysen af longitudinelle data om udviklingen af færdigheder i brøkregning hos skoleelever. Kapitlet vil først introducere konceptet af latent growth modeller og forklare, hvordan de anvendes til at analysere udviklingen af latente variabler over tid. Der vil også gives eksempler på, hvordan disse modeller kan bruges. Estimationsmetoder vil også blive diskuteret i dette kapitel, herunder maximum likelihood (LM) estimation. Endelig vil kapitlet omhandle den faktiske dataanalyse ved hjælp af latent growth modeller. Der vil blive præsenteret et case-studie, hvor latent growth modeller anvendes til at analysere en datasæt om færdigheder i brøkregning hos skoleelever. Analyseresultaterne vil blive diskuteret, og deres implikationer for undervisning og pædagogik vil blive overvejet.

2 | Lineære Mixed modeller

Vi vil i dette kapitel introducere *Lineære mixed modeller med stokastiske effekter*, samt nogle grundlæggende begreber og definitioner, som knytter sig til sådanne modeller; såsom *Hierarkisk data* og *longitudinelle data*. Vi vil også introduesere nogle etimations metoder , blandt andet *Maksimun Likelihood estimat(ML)* og *Restricted miksimum likelihood estimat(REML)*, samt hypotesetests og konfidensintervaller.
Dette kapitel tager udgangspunkt i kapitel 5 i [4], medmindre andet fremgår af teksten.

2.1 Lineære Mixede modeller

I dette afsnit vil vi introducere *Lineære mixed modeller*, som er en kendt statisk modeller anvendt i analysen for data med indlejret eller hierarkisk struktur, som kaldes for *hierarkiske data*. I en sådan type af data er individuelle observationer grupperet indenfor højere niveauer som fx. skoler, regioner eller lande. I hierarkisk data er der ofte en vis afhængighed mellem observationerne inden for grupperne. Og derfor er en klassisk lineær regressionsmodel(OLS) irrelevant; blandt andet fordi disse modeller er bygget på forudsætningen om, at observationer er uafhængige [1, s. 294]. Fordelen ved at bruge lineære mixed modeller er, at disse modeller tager højde for variansen både indenfor og imellem grupperne i data.
Mixed modeller er også anvendt i tilfældet af *longitudinelle data*, som er data med gentagne målinger for det samme individ på forskellige tidspunkter. De kan eksempelvis være brugt til at undersøge udviklingen af elever i skoler i løbet af et skoleår.
Nu vil vi definere Lineære mixed modeller. Disse modeller indeholder faste/systematiske effekter og stokastiske/tilfældige effekter. De beskriver generelt relationen/afhængigheden mellem observationerne både indenfor og imellem grupperne ved antagelsen af, at der eksisterer en eller flere uobserverede latente variable for hver gruppe af data. De latente variable er stokastiske, og dermed refererer de til den stokastiske del i modellen. Vi lader nu Y være en $n \times 1$ vektor af data, X en $n \times p$ design matrix for systematiske effekter, $\beta = (\beta_1, \ldots, \beta_p)^T$ en $p \times 1$ vektor af ukendte regressions koefficienter for systematiske effekter, Z en $n \times q$ design matrix for tilfældige effekter, hvor $q \leq p$, $U = (U_1, \ldots, U_q)^T$ en $n \times q$ vektor af tilfældige effekter og $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ en vektor af uafhængig og identisk fordelte fejl. Vi kan konstruere den følgende definition.

Definition 2.1 Lineære mixed modeller

Betragt en stokastisk vektor $Y = (Y_1, ..., Y_n)$. Da definerer lineær mixed model som følgende

$$Y = X\beta + ZU + \epsilon \tag{2.1}$$

hvor X og Z er kendte matricer, og $\epsilon \sim N(0, \Sigma)$ og $U \sim N(0, \Psi)$ er uafhængige. Σ og Ψ er kovariansmatricerne, β er en parameter vektor, som kaldes for faste eller systematiske effekter, og U betegner stokastiske effekter i modellen.

Bemærk at (2.1) viser en lineær kombination mellem faste effekter $X\beta$ og tilfældige effekter ZU. Bemærk også at Σ og Ψ er afhængige af ukendte parametre ψ, som skal estimeres. Altså Σ og Ψ er funktioner af ψ.

Da tilfældige effekter U i modellen i(2.1) er uafhængige af fejlene ϵ, gælder det at

$$\begin{pmatrix} \epsilon \\ U \end{pmatrix} = \begin{pmatrix} \Sigma & 0 \\ 0 & \Psi \end{pmatrix} \tag{2.2}$$

Bemærk, at modellen i (2.1) kan fortolkes som en hierarkisk model på følgende måde,

$$Y \mid U = u \sim N(X\beta + Zu, \Sigma(\psi)) \tag{2.3}$$

$$U \sim N(0, \Psi(\psi)) \tag{2.4}$$

med tæthedsfunktioner,

$$f_U(u; \psi) = \frac{1}{\sqrt{(2\pi)^q det(\Psi(\psi))}} \exp(-\frac{1}{2} u^T \Psi(\psi) u) \quad for \quad u \in \mathbb{R}^q \tag{2.5}$$

$$f_{Y|U}(y; \psi) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma(\psi))}} \exp\left(-\frac{1}{2}(y - X\beta - Zu)^T \Sigma(\psi)^{-1}(y - X\beta - Z)\right) \quad for \quad y \in \mathbb{R}^n \tag{2.6}$$

Da tilfældige effekter U og fejlene ϵ er uafhængige, og begge følger en normal fordeling med middelværdien 0 og kovariansmatricer i henholdesvis Ψ og Σ, så følger Y en multivariat normalfordeling med

$$\mathbb{E}[Y] = X\beta \tag{2.7}$$

$$V = \Sigma + Z\Psi Z^{-1} \tag{2.8}$$

Og dermed den marginale tæthed for Y er

$$f_Y(y; \beta, \psi) = \frac{1}{(2\pi)^{n/2} det(V(\psi))} \exp(-\frac{1}{2}(y - X\beta)^T V(\psi)^{-1}(y - X\beta)) \tag{2.9}$$

Lineære mixede modeller kan også være brugbare i situationer, hvor der mangler data, idet lineære mixede modeller er bygget på antagelsen om, at manglende data er *Missing at random (MAR)*. Under antagelsen om, at manglende data er MAR, er inferensen baseret på ML-estimation i lineære mixede modeller valid [6, s. 48] I følgende eksempel vil vi beskrive et eksempl på hierarakiske modeller, *Ensidet varians analyse* eller på engelsik *One-way ANOVA*.

Eksempel 2.2 Ensidet variansanalyse
I dette eksempel vil vi beskrive *Ensidet variansanalyse/ One way ANOVA* som en varianskomponentmodel, som er en slags hierarkisk model, hvor grupper, der analyseres indeholder forskellige populationer, og de aktuelle grupper betragtes som tilfældige resultater af at analysere et antal grupper indenfor disse populationer. Variansen mellem grupperne er beskrevet ved paramtre, der karakteriserer variationerne mellem grupperne [4, s.48].
For at definere en sådan model, betragter vi en population med n observationer opdelt i k grupper, og hvor den i'te gruppe indeholder n_i observationer for $i = 1, \ldots, k$, og $n = \sum_{i=1}^{k} n_i$. Dette kan repræsenteres på formlen

$$\begin{array}{ll} \text{Group} & \text{Observations} \\ 1 & Y_{11}, Y_{12}, \ldots, Y_{1n_1} \\ 2 & Y_{21}, Y_{22}, \ldots, Y_{2n_2} \\ \vdots & \vdots, \vdots \\ k & Y_{k1}, Y_{k2}, \ldots, Y_{kn_k} \end{array} \tag{2.10}$$

Denne opdeling eller gruppering kan være resultatet af en eller flere faktorer, og hvert sæt af faktorsniveauer definerer en enkel behandling. Så hvis faktoren eksempelvis er *køn*, er niveauerne *mand* og *kvinde*. I tilfælde af at $n_1 = n_2 = \cdots = n_k = m$ siger vi, at eksperimentet er i *balance*.

Definition 2.3 Betragt stokastiske variable Y_{ij} hvor $i = 1, \ldots, k$ og $j = 1, \ldots, n_i$. Da definerer ensidet variansanalyse model med stokastiske effekter for Y_{ij} som

$$Y_{ij} = \mu + U_i + \epsilon_{ij} \tag{2.11}$$

hvor $\epsilon_{ij} \sim N(0, \sigma^2)$, $U_i \sim N(0, \tau^2)$, ϵ_{ij} er uafhængige af hindanden, og det samme gælder for U_i, ligeledes erϵ_{ij} uafhængige af hinanden.

Bemærk at, modellen i (2.11) kan udtrykkes som en hierakisk model ved at lade $\mu_i = \mu + U_i$, og så får vi at,

$$Y_{ij} \mid \mu_i \sim N(\mu_i, \sigma^2) \tag{2.12}$$

$$\mu_i \sim N(\mu, \tau^2) \tag{2.13}$$

hvor $\sigma^2 = \text{Var}(\epsilon_{ij})$ og $\tau^2 = \text{Var}(U_i)$, og μ_i'er er uafhængige og hvorY_{ij} er betinget uafhængigt givet μ_i. Bemærk, at faste parametre i modellen specificeret i (2.11) er (μ, σ^2, τ^2). Vi kan da definere *Signal noise ratio* γ som $\gamma = \frac{\tau^2}{\sigma^2}$. Parameteren γ udtrykker således inhomogeniteten imellem grupperne i forhold til den interne variation i grupperne.
Modellen i (2.11) kan også udtrykkes på en vektorform som følgende,

$$Y = X\beta + ZU + \epsilon \tag{2.14}$$

Hvor $X = 1_n$, som er en søjlevektor af 1'er, $\beta = \mu$, $U = (U_1, U_2, \ldots, U_k)^T$, $\Sigma = \sigma^2 I_n$ og $\Psi = \sigma_u^2 I_k$. Derudover er Z en $(n \times k)$-matrix, hvor det i, j'te element er 1, hvis y_{ij} tilhører den i'te gruppe og ellers 0.

Marginale fordelinger i den tilfældige model for ensidet variansanalyse.

Den marginale fordeling af Y_{ij} er en normal fordeling med $\mathbb{E}[Y_{ij}] = \mu$ og kovariansen

$$Cov[Y_{ij}, Y_{hl}] = \begin{cases} 0 & \text{for } i \neq h \\ \tau^2 & \text{for } i = h, \quad j \neq l \\ \tau^2 + \sigma^2 & \text{for } (i,j) = (h,l) \end{cases} \tag{2.15}$$

Dette gælder på grund af følgende

- For $i \neq l$ betragter vi to forskellige gruppe, og derfor gælder det, at

$$Cov[Y_{ij}, Y_{hl}] = Cov[U_i, U_h] + Cov[U_h, \epsilon_{ij}] + Cov[U_i, \epsilon_{hl}] + Cov[\epsilon_{ij}, \epsilon_{hl}] = 0 \tag{2.16}$$

- For $i = h, i \neq l$ betragter vi to observationer fra den samme gruppe, og derfor gælder det at

$$Cov[Y_{ij}, Y_{hl}] = Cov[U_i, U_h] = \text{Var}[U_i] = \tau^2 \tag{2.17}$$

- For $i = h$ og $j = l$ betragter vi den selvsamme observation, så

$$Cov[Y_{ij}, Y_{hl}] = \text{Var}[Y_{ij}] = \text{Var}[U_i] + \text{Var}[\epsilon_{ij}] = \tau^2 + \sigma^2 \tag{2.18}$$

Vi kan på samme måde for kovariansen, beregne korrelationen mellem observationerne i de tre tilfælde $((i,j) = (h,l), (i = h, j \neq l)$ og $i \neq h)$. Ved at benytte relationen

$$Corr[Y_{ij}, Y_{hl}] = \frac{Cov[Y_{ij}, Y_{hl}]}{\sqrt{Var[Y_{ij}]}\sqrt{\text{Var}[Y_{hl}]}} \tag{2.19}$$

, så får vi at

$$Corr[Y_{ij}, Y_{hl}] = \begin{cases} 1 & \text{for } (i,j) = (h,l) \\ \frac{\text{Var}[U_i]}{\sqrt{Var[Y_{ij}]}\sqrt{\text{Var}[Y_{hl}]}} = \frac{\tau^2}{\tau^2+\sigma^2} = \frac{\gamma}{\gamma+1} & \text{for} \quad i = h, \quad j \neq l \\ 0 & \text{for} \quad i \neq h \end{cases} \tag{2.20}$$

Bemærk her, at den simultane fordeling for gruppegennemsnittene er karakteriseret ved

$$Cov[\bar{Y}_{i.}, \bar{Y}_{h.}] = \begin{cases} \tau^2 + \sigma^2/n_i & \text{for} \quad i = h \\ 0 \quad \text{for} \quad \text{ellers} \end{cases} \tag{2.21}$$

Dette betyder, at de k gruppegennemsnitte $\bar{Y}_i$ for $i = 1, 2, \ldots, k$ er indbyrdes uafhængige, og variansen for gruppegennemsnittet er givet ved

$$V[\bar{Y}_{i.}] = \tau^2 + \sigma^2/n = \sigma^2(\gamma + 1/n) \tag{2.22}$$

Bemærk at, en forøgelse af stikprøvestørrelsen i de enkelte grupper således vil forøge præcisionen ved bestemmelse af gruppeforventningsværdien α_i, men variationen mellem de enkelte gruppeforventningsværdier formindskes naturligvis ikke ved denne gennemsnitsdannelse.
En vigtig bemærkning her er, at observationerne fra den samme gruppe er korrelerede, og at

der er en positiv kovarians mellem observationerne. Denne positive kovarians udtrykker netop, at observationer indenfor en gruppe vil afvige i samme retning fra den marginale middelværdi μ, hvor korrelationskoefficienten er givet ved

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{\gamma}{\gamma + 1} \tag{2.23}$$

som beskriver korrelationen indenfor gruppen. ρ kaldes ofte for *intraklassekorrelationen.* Bemærk yderligere, at hvis vi betragter observationssættet svarende til den $i$te gruppe som en n_i-dimensional søjlevektor,

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} \tag{2.24}$$

kan vi skrive, at korrelationsmatricen i den marginale fordeling af Y_i er en *equikorrelations-matrix* af formen

$$E_n = (1-\rho)I_n + \rho J_n = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix} \tag{2.25}$$

hvor J_{n_i} er en $n_i \times n_i$-matrix bestående af et 1-tal, og Y_i for $i = \{1, \dots, k\}$ kan således beskrives som k uafhængige observationer af en n_i dimensional variabel $Yi \sim N_{n_i}(, \sigma^2 I_{n_i} + \tau^2 J_{n_i})$, og dermed er dispersionsmatricen for Y_i givet ved

$$\begin{aligned} V_i =& D\left[Y_i\right] = E\left[\left(Y_i - \mu\right)\left(Y_i - \mu\right)^T\right] \\ =& \begin{pmatrix} \tau^2 + \sigma^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \tau^2 + \sigma^2 & \dots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \dots & \tau^2 + \sigma^2 \end{pmatrix} \end{aligned} \tag{2.26}$$

For at finde dispersionsmatricen $D[Y]$ for hele observationssættet, opstiller vi samtlige observationer i én søjle, organiseret efter grupperne, således

$$V = D[Y] = \text{ Blok diag}\left\{V_i\right\} \tag{2.27}$$

hvor V_i er givet i (2.26). Tilsvarende finder man, at korrelationsmatricen for hele observationssættet er en $n \times n$-dimensional blokmatrix med matricerne E_{n_i} i diagonalen, og 0'er udenfor, hvilket illustrerer, at observationer fra forskellige grupper er uafhængige, mens observationer fra samme gruppe er korrelerede.

◀

2.2 Estimation for faste effekter og variansparametrerne i Lineære mixede modeller

De faste parametre β og variansparametre ψ kan vi estimere ved at benytte den marginale tæthedsfunktion i (2.9). Udfra denne tæthed, kan vi skrive likelihoodfunktionen,

$$L(\beta, \psi; y) = \frac{1}{(2\pi)^{1/2} det(V(\psi))} \exp(-\frac{1}{2}(y - X\beta)^T V(\psi)^{-1}(y - X\beta)) \tag{2.28}$$

og dermed er log-likelihoodfunktionen givet ved,

$$\ell(\beta, \psi; y) = -\frac{1}{2}\log|V(\psi)| - \frac{1}{2}(y - X\beta)^T V(\psi)^{-1}(y - X\beta) \tag{2.29}$$

Ved at fastholde ψ og differentiere ℓ med hensyn til β, får vi scorefunktionen,

$$s(\beta; \psi) = X^T[V(\psi)^{-1}y - V(\psi)^{-1}X\beta] \tag{2.30}$$

og dermed får vi maksimum likelihood-estimatet for β ved at løse ligningen $s(\beta; \psi) = 0$. Altså er estimatet for β en løsning til den følgende ligning

$$(X^T V(\psi)^{-1}X)\beta = X^T V(\psi)^{-1}y \Rightarrow \hat{\beta} = (X^T V(\psi)^{-1}X)^T X^T V(\psi)^{-1}y \tag{2.31}$$

Dette genkender vi som vægtede mindste kvadrater (WLS). WLS er også kendt som generaliserede mindste kvadrater (GLS) [1, s. 295]. Yderligere finder vi den observerede Fisher informationmatrix for $\hat{\beta}$, ved at aflede scorefunktionen med hensyn til β, så

$$i(\hat{\beta}) = X^T V(\psi)^{-1}X \tag{2.32}$$

hvoraf vi får en dispersionmatrix på formlen

$$\text{Var}[\hat{\beta}] = X^T V(\psi)^{-1}X \tag{2.33}$$

Her er det vigtigt at bemærke, at estimatet for β er afhængigt af variansparametrene ψ, og derfor er vi nødt til at estimere ψ. Dette kan vi gøre ved at anvende den modificerede profil likelihood ($\beta = \hat{\beta}$), som er givet ved

$$\ell(\psi) = -\frac{1}{2}\log|V(\psi)| - \frac{1}{2}\left(Y - X\hat{\beta}(\psi)\right)^T V(\psi)^{-1}\left(Y - X\hat{\beta}(\psi)\right) \tag{2.34}$$

hvor $\hat{\beta}$ er WLS-estimatet for β.

En vigtig bemærkning her er, at vi kan få WLS estimat på en alternativ måde end ved at benytte maksimum likelihoodfunktionen. Vi kan benytte ortogonel projektionsmetoden, som går ud på at projicere data på et underrum, som er ortogonel på underummet, der indeholder alle lineære kompinationer af faste og tilfældige effekter i modellen. Vi vil illustrere ortogonel projektionsmetoden i følgende eksempel.

Eksempel 2.4 Estimationen ved brug af ortogonelle projektioner

Vi betragter en lineær model, hvor $Y \sim N(\mu, \sigma^2 I)$, hvor $\mu = X\beta$. Vi lader P være ortogonal projektion på $M = span(X)$ og $P = X(XX^T)^{-1}X^T$. Da M og P er ortogonale, kan vi bruge Pythagoras' sætning til at skrive

$$\|Y - X\beta\|^2 = \|Y - PY\|^2 + \|PY - X\beta\|^2 \tag{2.35}$$

Da $\hat{\mu} = Py = X(X^TX)^{-1}X^Ty$, medfører det at $\hat{\beta} = (X^TX)^{-1}X^Ty$. Og variansen er dermed,

$$\hat{\sigma}^2 = \|Y - PY\|^2/n = \|Y - X\hat{\beta}\|^2/n \tag{2.36}$$

Men kovariansmatricen er ikke altid lig I. Vi antager derfor at $Y \sim N(\mu, \sigma^2 W)$, hvor $W = LL^T$, og hvor L er Cholesky-faktorisering af W. Bemærk her, at W er positiv semi-definit, og dermed er L invertibel og W er derfor positiv definit). Bemærk også, at MLE-estimat baseret på Y er ækvivalent med MLE-estimat baseret på $\tilde{Y} = L^{-1}Y$. Yderligere er kovariansen givet ved $Cov(\tilde{Y}) = \sigma^2 I$ og middelværdien $\mathbb{E}[\tilde{Y}] = L^{-1}X\beta = \tilde{X}$, og dermed får vi GLS-estimat for β og σ^2 som,

$$\hat{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y} = (X^TW^{-1})^{-1}X^TW^{-1}y \tag{2.37}$$

Og

$$\hat{\sigma}^2 = \frac{1}{n}\left(y - X\hat{\beta}\right)^T W^{-1}\left(y - X\hat{\beta}\right) \tag{2.38}$$

◀

Vi vil introducere et eksempel på et specielt tilfælde af den modificerede profil likelihood i (2.34), hvor kovariansmatricen for ϵ er en diagonal matrix i form af $\mathrm{Var}(\epsilon) = \sigma^2 I$.

Eksempel 2.5 Profil likelihood med ukorreleret støj
Vi betragter modellen

$$Y = X\beta + ZU + \epsilon \tag{2.39}$$

hvor U og ϵ begge følger en normalfordeling med middelværdi 0 og varians $\mathrm{Var}(\epsilon) = \Sigma = \sigma^2 I_n$ og $\mathrm{Var}(U) = \Psi = \tau^2 L(\theta)L(\theta)^T$. $L(\theta)$ er $k \times k$-matrix.
Vi ved, at varians-kovariansmatricen for Y er givet ved

$$V(\psi) = \Sigma + Z\Psi Z^T = \sigma^2 I_n + \tau^2 ZL(\theta)L(\theta)^T Z^T \tag{2.40}$$

$$= \sigma^2\left(I_n - \frac{\tau^2}{\sigma^2}ZL(\theta)L(\theta)^T Z^T\right) = \sigma^2 W(\phi, \theta) \tag{2.41}$$

hvor $\phi = \frac{\tau^2}{\sigma^2}$. Dermed er $Y \sim N_n(X\beta, \sigma^2 W)$, hvorfor $\hat{\beta}$ kan udregnes ved brug af normalfordelingen (ortogonal projektion)

$$\hat{\beta}(\phi, \theta) = \left(X^TW(\phi, \theta)^{-1}X\right)^{-1} X^TW(\phi, \theta)^{-1}y \tag{2.42}$$

Bemærk, at $\hat{\beta}$ er WLS estimat, defineret i (2.31). Vi ved fra ortogonal projektioneksemplet, at

$$\sigma^2(\phi, \theta) = \frac{1}{n}\left(y - X\beta(\phi, \theta)^T\right)^T W(\phi, \theta)^{-1}\left(y - X\beta(\phi, \theta)\right) \tag{2.43}$$

Dermed er profile log-likelihood for (ϕ, θ)

$$\begin{aligned}
\ell(\phi,\theta) &= -\frac{1}{2}\log \mid V(\hat{\sigma},\phi,\theta) \mid -\frac{1}{2}\left(y - X\hat{\beta}(\phi,\theta)\right)^T V(\hat{\sigma},\phi,\theta)^{-1}\left(y - X\hat{\beta}(\phi,\theta)\right) \\
&= -\frac{1}{2}\log\left(\mid \hat{\sigma}^2(\phi,\theta)W(\phi,\theta) \mid\right) - \frac{1}{2}\left(y - X\hat{\beta}(\phi,\theta)\right)^T \left(\hat{\sigma}^2(\phi,\theta)W(\phi,\theta)\right)^{-1}\left(y - X\hat{\beta}(\phi,\theta)\right) \\
&= -\frac{1}{2}\log\left(\mid \hat{\sigma}^2(\phi,\theta)W(\phi,\theta) \mid\right) - \frac{1}{\hat{\sigma}^2(\phi,\theta)}\left(y - X\hat{\beta}(\phi,\theta)\right)^T W(\phi,\theta)^{-1}\left(y - X\hat{\beta}(\phi,\theta)\right) \\
&= -\frac{1}{2}\log\left(\mid \hat{\sigma}^2(\phi,\theta)W(\phi,\theta) \mid\right) - \frac{n}{2}
\end{aligned} \tag{2.44}$$

Hvis $k < n$, kan vi benytte det følgende resultat,

$$\mid W(\phi,\theta) \mid = \mid I_n + \phi Z L(\theta) L(\theta)^T Z^T \mid = \mid I_k + \phi L(\theta)^T Z^T Z L(\theta) \mid \tag{2.45}$$

hvilket vil spare os for beregninger, da vi skal finde determinant af en matrix med mindre dimension. Vi får, at

$$\ell(\phi,\theta) = -\frac{1}{2}\log\left(\hat{\sigma}^{2n}(\phi,\theta) \mid I_k + \phi L(\theta)^T Z^T Z L(\theta) \mid\right) - \frac{n}{2} \tag{2.46}$$

$$= -\frac{n}{2}\log\hat{\sigma}^2(\phi,\theta) - \frac{1}{2}\log \mid I_k + \phi L(\theta)^T Z^T Z L(\theta) \mid - \frac{n}{2} \tag{2.47}$$

For at minimere $\ell(\phi,\theta)$, fjerner vi konstante led, hvormed

$$\ell(\phi,\theta) \equiv -\frac{n}{2}\log\hat{\sigma}^2(\phi,\theta) - \frac{1}{2}\log \mid I_k + \phi L(\theta)^T Z^T Z L(\theta) \mid \tag{2.48}$$

Når $k < n$ benytter vi vi os af Woodbury til at omskrive den inverse af $W(\phi,\theta)$ til den inverse af en $k \times k$-matrix, hvilket vil gøre udregningerne for $\hat{\sigma}^2$ mere simple. Vi vil da få

$$W(\phi,\theta)^{-1} = \left(I + \phi L(\theta) L(\theta)^T Z^T Z\right)^{-1} \tag{2.49}$$

$$= I^{-1} - I^{-1} Z\left(\phi^{-1}\left(L(\theta)L(\theta)^T\right)^{-1} + Z^T I^{-1} Z\right)^{-1} Z^T I^{-1} \tag{2.50}$$

$$= I - Z\left(\phi^{-1}\left(L(\theta)\right)^{-1} L(\theta)^{-1} + Z^T Z\right)^{-1} Z^T \tag{2.51}$$

◀

Bemærk, at i Lineære modeller og herunder lineære mixed modeller får vi generelt et biased variansestimat [8, s. 16]. F.eks hvis vi betragter en normal stikprøve, hvor $Y_i \sim N(\mu, \sigma^2)$, er ML-estimat $\hat{\sigma}^2$, $\mathbb{E}[\hat{\sigma}^2]) = \frac{n}{n-1}\sigma^2$ biased (unbiased estimat er $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$). Biasedness kommer af at vi skal estimere μ. Hvis μ er kendt, så ML-estimat for σ^2 er $\hat{\sigma}^2 = \sum_i (Y_i - \mu)^2$. Dog opstår der et problemt, idet vi i praksis ikke kender μ, og vi benytter derfor gennesnittet af Y, $\bar{Y}_\cdot$ i stedet for μ. Dermed er ML-estimat for σ^2 er $\hat{\sigma}^2 = \sum_i (Y_i - \bar{Y}_\cdot)^2$. Idet $\bar{Y}_\cdot$ er gennemsnittet af data Y_i, er det meget tæt på data Y_i, og dermed vil estimatet $\hat{\sigma}^2 = \sum_i (Y_i - \bar{Y}_\cdot)^2$ underestimere $\hat{\sigma}^2 = \sum_i (Y_i - \mu)^2$, hvorfor vi får $\mathbb{E}[\hat{\sigma}^2]) = \frac{n}{n-1}\sigma^2$ som er mindre end σ^2.
I følgende afsnit vil vi introducere en anden salgs maksimum likelihoodestimat, som løser problematikken med biasedness. Det er såkaldt *Residual Maksimum Likelihood (REML)* [8, s. 17].

2.2.1 Residual Maksimum Likelihood (REML)

Vi kan overkomme problematikken med biased variansestimat i MLE ved at benytte såkaldt *Residual Maksimum Likelihood (REML)*. Da biasedness kommer af, at vi skal estimere middelværdivektoren, eliminerer vi middelværdien i REML-estimat. Dette gør vi ved at transformere data [8, s. 17].
Vi betragter modellen i (2.1), hvor X igen er en $n \times p$-designmatrix. Vi lader A være en anden matrix med dimension $n \times (n-p)$ og har søjler, der spænder over det ortogonale komplement $M^{\perp} = span(A)$ af $M = span(X)$, hvilket medfører at $A^T X = 0$. Vi transformerer nu data Y i modellen (2.1) således

$$\tilde{Y} = A^T Y = A^T ZU + A^T \epsilon \tag{2.52}$$

som har middelværdi 0 og kovarians matrix $A^T V(\psi) A$. Det vil sige, at vi kun har brug for at estimere variansen ψ. Bemærk, at A^T ikke er invertibel, så vi får derfor forskellige estimater for ψ afhængig af, om vi benytter Y eller $\tilde{Y}$. Vi kan benytte MLE-estimat baseret på $\tilde{Y}$ til at estimere ψ, og indsætte det opnåede estimat i WLS for at opnå et REML-estimat for β,

$$\hat{\beta} = \left(X^T V(\hat{\psi})^{-1} X \right)^{-1} X^T V(\hat{\psi})^{-1} y \tag{2.53}$$

Bemærk også, at for hvert lineært underrum, er der uendeligt mange basis, så hvis vi vælger en anden matrix B således, $M^{\perp} = span(B)$, får vi det samme REML-estimat for ψ uanset hvilken matrix A eller B, vi bruger til at transformere data.

2.3 Estimation for stokastiske effekter (BLUP)

I dette afsnit er vi interesserede i at bestemme et estimat for de tilfældige effekter. Til dette formål vil vi introducere nogle estimationsmetoder, såsom *Den bedste lineære unbiased prædiktor (BLUP)* og *Faktoriseret Likelihoodestimat.*

BLUP

Indledningsvis introducerer vi det såkaldt *Det bedste lineære unbised estimat (BLUE).* Vi antager at Y en stokastisk vektor med middelværi $\mu = X\beta$ og en kendt varians. Vi genkalder WLS-estimat for β, som er givet ved

$$\hat{\beta} = \left(X^T V^{-1} X \right)^{-1} X^T V^{-1} Y \tag{2.54}$$

og som minimerer den følgende ligning,

$$(Y - X\beta)^T V^{-1} (Y - X\beta) \tag{2.55}$$

Ovenstående estimat kaldes for den bedste lineære unbiased estimat (BLUE), som har den mindste varians i forhold til alle de andre *lineære unbiased estimater (LUE).* Altså $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta})$ er positiv semi-definit for hver LUE $\tilde{\beta}$.
Da tilfældige effekter i mixed modeller er ukendte normaltfordelte stokastiske variable, prædikterer vi disse effekter mere, end vi estimerer dem [6, s.45]. Til dette formål, benytter vi

den betingede forventede værdi af stokastiske effekter, givet observeretrespons værdier som følgende,

$$\hat{U} = \mathbb{E}[U \mid Y] = \Psi Z^T V^{-1}(Y - X\beta) \tag{2.56}$$

Da tilfældige effekter $u \sim N(0, \Psi)$ og $Y \mid U = u \sim N(X\beta + Zu, \Sigma)$ med kovarians $Cov(U, Y) = \Psi Z^T$ og varians $\text{Var}(Y) = V = Z\Psi Z^T + \Sigma$, kan vi skrive at

$$\hat{U} = \Psi Z^T (Z\Psi Z^T + \Sigma)^{-1}(Y - X\beta) \tag{2.57}$$

$$= (\Psi^{-1} + Z^T\Sigma^{-1}Z)^{-1} Z^T\Sigma^{-1}(Y - X\beta) \tag{2.58}$$

hvor den sidste lighed kommer af Woodbury. Estimatet for u kaldes for den bedste lineære unbiased prædikter (BLUP). BLUP-estimater er de bedste, idet de har den mindste varians blandt alle andre lineære unbiased estimater, således at $\text{Var}(tildeu) - \text{Var}(\hat{u})$ er positiv semi-definit [7, s. 2]. Det vil sige, at BLUP-estimater er de meste præcise lineære unbiased estimater [6, s. 45]. BLUP-estimater er lineære, idet de er lineære funktioner af data Y. BLUP-estimater er unbiased, idet deres forventede værdi er lig med den forventede værdi for tilfældige effekter for et enkelt individ i, altså $\mathbb{E}[\hat{U}_i] = U_i$.
For at se, hvorfor BLUP er den bedste prædikter, præsenterer vi de følgende resultater.

Generelle resultater

betragt X og Y som to stokastiske variabler, og g som en realfunktion. Da

$$Cov\left(Y - \mathbb{E}[Y \mid X], g(X)\right) \tag{2.59}$$

$$= Cov\left(\mathbb{E}[Y - \mathbb{E}[Y \mid X] \mid X], \mathbb{E}[g(X) \mid X]\right) + \mathbb{E}[Cov(Y - \mathbb{E}[Y \mid X], g(X) \mid X)] \tag{2.60}$$

$$= 0 \tag{2.61}$$

For hver prædikter $\tilde{Y} = f(X)$ af Y, kan vi jævnføre (2.59) skrive,

$$\mathbb{E}[(Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - f(X)] = 0 \tag{2.62}$$

hvor vi har valgt g til at være $g(X) = \mathbb{E}[Y \mid X] - f(X)$. Vi kan beregne *Mean squre prediction error (MSPE)* således

$$MSPE = \mathbb{E}[Y - \tilde{Y}]^2 \tag{2.63}$$

$$= \mathbb{E}\left(Y - \mathbb{E}[Y \mid X]\right)^2 + \mathbb{E}\left(\mathbb{E}[Y \mid X] - \tilde{Y}\right)^2 \tag{2.64}$$

$$+ 2\mathbb{E}\left((Y - \mathbb{E}[Y \mid X])(\mathbb{E}[Y \mid X] - \tilde{Y})\right) \tag{2.65}$$

$$= \mathbb{E}\left(Y - \mathbb{E}[Y \mid X]\right)^2 + \mathbb{E}\left(\mathbb{E}[Y \mid X] - \tilde{Y}\right)^2 \geq \mathbb{E}\left(Y - \mathbb{E}[Y \mid X]\right)^2 \tag{2.66}$$

hvor den sidste lighed kommer af (2.59), og dermed $\mathbb{E}[Y \mid X]$ minimerer MSPE. Altså er den bedste prædikter af $Y \mid X$ dens forventede værdi $\mathbb{E}[Y \mid X]$.

Vi kan også benytte resultatet i (2.59) til at beregne den prædikterede varians. Vi kan skrive Y som $Y = \mathbb{E}[Y \mid X] + (Y - \mathbb{E}[Y \mid X])$, hvor $\mathbb{E}[Y \mid X]$ og $Y - \mathbb{E}[Y \mid X]$ er ukorrelerede. Så

$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y \mid X]) + \text{Var}(Y - \mathbb{E}[Y \mid X]) = \text{Var}(\mathbb{E}[Y \mid X]) + \mathbb{E}[\text{Var}(Y \mid X)] \tag{2.67}$$

hvormed

$$\mathrm{Var}(Y - \mathbb{E}[Y \mid X]) = \mathbb{E}[\mathrm{Var}(Y \mid X)] \tag{2.68}$$

Det vil sige, at den prædikterede varians er lig med den betingede forventede værdi af variansen af Y givet X.

Estimation for tilfældige effekter ved faktoriseret likelihood

Vi betragter nu tilfældige effekter som parametre, og formulerer en såkaldt *hierakisk likelihood* svarende til alle parametrene som følgende

$$f(y, u; \beta, \psi) = f_{Y|u}(y; \beta) f_U(u; \psi) \tag{2.69}$$

hvor $f_{Y|u}(u, \beta)$ og $f_U(u; \psi)$ er givet ved (2.3) og (2.4), hvorfor den hierakiske log-likelihood kan udtrykkes som

$$\begin{aligned} \ell(\beta, \psi, u) = &-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(y - X\beta - Zu)^T\Sigma^{-1}(y - X\beta - Zu) \\ &-\frac{1}{2}\log|\Psi| - \frac{1}{2}u^T\Psi^{-1}u \end{aligned} \tag{2.70}$$

Differentierer vi $\ell(\beta, \psi, u)$ med hensyn til u, får vi at,

$$\frac{\partial}{\partial u}\ell(\beta, \psi, u) = Z^T\Sigma^{-1}(y - X\beta - Zu) - \Psi^{-1}u \tag{2.71}$$

Sætter vi det lig 0 og udskifter β med $\hat{\beta}$, får vi den følgende ligning

$$\left(Z^T\Sigma^{-1}Z + \Psi^{-1}\right)u = Z^T\Sigma^{-1}(y - X\beta) \tag{2.72}$$

hvor esimatet $\hat{u}$ er en løsning til den. Ligeledes kaldes estimatet $\hat{u}$ for *den bedste lineære unbiased prædiktor, BLUP*. Navnet kommer af, at dette estimat minimerer

$$\mathbb{E}\left[\left(\widehat{U}_i - \mathbb{E}\left[U_i\right]\right)^2\right] \tag{2.73}$$

blandt alle lineære estimatorer, $\hat{U}_i$, der er centrale (unbiased) for $\mathbb{E}[U_i]$, og dermed opfylder $\mathbb{E}[\widehat{U}_i] = \mathbb{E}\left[U_i\right]$.
En vigtig bemærkning her er, at i BLUP-estimater erstatter vi tilfældige effekter U_i med deres betingede middelværdi $\hat{u}$ givet dataerne, og benytter derefter disse værdier til at lave prædiktion således

$$\widehat{Y} = X\widehat{\beta} + Z\widehat{u} \tag{2.74}$$

Den anden afledte for likelihood funktionen er

$$\frac{\partial^2}{\partial u \partial u^T}\ell(\beta, \psi; u) = -Z^T\Sigma^{-1}Z - \Psi^{-1} \tag{2.75}$$

og dermed er Fisher informationmatrix,

$$i(\hat{u}) = Z^T\Sigma^{-1}Z + \Psi^{-1} \tag{2.76}$$

Dette vil vi bruge senere i projektet til at undersøge nøjagtigheden af $\hat{u}$.

2.3.1 Likelihood ratio tests (LRTs)

Vi vil introducere *Likelihood ratio test (LRT)*, som er en klasse af tests baseret på at sammenligne værdier af likelihoodfunktioner for to modeller (nul-model, som opfylder nulhypotesen, og referensmodel, som omfatter både nul- og alternativhypotese), der definerer en hypotese, der skal testes. Vi vil benytte LRTs til at teste hypotesen om varians- eller parametre om de faste effekter. LRTs er opnået ved trække $-2log$ likelihoodfunktionen for en nul-model fra den for en referensmodel [6, s. 35], som kan ses i følgende ligning

$$-2\log\left(\frac{L(\hat{\beta}_0,\hat{\psi}_0;y)}{L(\hat{\beta},\hat{\psi};y)}\right) = -2\log\Big(L(\hat{\beta}_0,\hat{\psi}_0;y)\Big) - 2\log\Big(L(\hat{\beta},\hat{\psi};y)\Big) \tag{2.77}$$

hvor $L(\hat{\beta}_0,\hat{\psi}_0;y)$ og $L(\hat{\beta},\hat{\psi};y)$ er værdien af likelihoodsfunktionen opnået ved ML- eller REML-estimater. LRT følger endevidere en asymptotisk χ^2-fordeling med frihedsgrad svarende til forskellen mellem antallet af parametre i nul-modellen og antallet af parametre i referensmodellen. Hvis værdien af LRTs er meget stor, har vi evidens for at forkaste nulhypotesen, og det gælder omvendt, hvis værdien for LRT er lille, har vi ikke evidens for at forkaste null-hypotesen [6, s. 35].
LRT brugt til at teste lineære hypotestests for faste parametre i lineære mixed modeller, er baseret på ML-estimation.

$$LRT = -2\Big(\ell_{ML}(\hat{\beta}_0) - \ell_{ML}(\hat{\beta})\Big) \tag{2.78}$$

Bemærk her, at for LRT for faste effekter, har begge modeller (nul-model og referansmodel) samme kovariansparametre, men forskellige faste effekter [6, s. 35].
LRTs brugt til at teste hypotesetests for variansparametre, er dog baseret på REML-estimation til at estimere parametre i begge modeller.

$$LRT = -2\Big(\ell_{REML}(\hat{\psi}_0) - \ell_{REML}(\hat{\psi})\Big) \tag{2.79}$$

2.3.2 Alternative Tests

I dette afsnit vil vi præsentere nogle alternative tests til LRT, som er brugt til at teste hypotese om parametre i lineære mixed modeller. Disse tests, i modsætning til LRT, kræver kun at tilpasses en model (referensmodel).

Alternative tests for parametre for faste effektter

En *t-test* er ofte brugt til at teste en enkelt parameter for fast effekt i lineære mixed modeller [6, s. 37]. Vi kan anvende t-test til at teste eksempelvis nulhypotesen $H_0 : \beta = 0$ mod den alternative hypotese $H : \beta \neq 0$, så t-statistik er givet ved,

$$t = \frac{\hat{\beta}}{se(\hat{\beta})} \tag{2.80}$$

hvor $se(\hat{\beta})$ er standard error for $\hat{\beta}$. Bemærk, at t-statistik i (2.80) generelt ikke følger en eksakt t-fordeling, men en asymptotisk fordeling. Vi benytter derfor *Kenward-Roger (KR)* metoden (beskrevet senere) til at estimere frihedsgrad for t-test [6, s. 37].

En *F-test* kan blive brugt til at teste lineær hypotes om multiple faste effekter i lineære mixed modeller. En lineær hypotese er generelt brugt til at teste $H_0 : L\beta = 0$ mod $H : L\beta \neq 0$, hvor L er en kendt matrix, og den tilhørende F-statistik er defineret ved

$$F = \frac{\hat{\beta}L^T \left(L(X^T\hat{V}^{-1}X)^{-1}L^T\right)^{-1}}{r(L)} \tag{2.81}$$

hvor $r(L)$ er rangen af L. F-test følger desuden en approksimativ F-fordeling [6, s. 37]. Frihedsgrad for F-test estimerer vi derfor ved KR-metoden.

Alternative tests for kovariansparametre

En test, som er brugt til at teste hypotese for kovariansparametre er *Wald-test*. Wald-test bliver anvendt til at teste hypotese af formlen $H_0 : K\beta = b$ mod $H : K\beta \neq b$, og har følgende z-statistik,

$$z = \left(K\hat{\sigma}^2 \left(X^T V^{-1}(\hat{\psi})X\right)^{-1} K^T V\right)^{-1/2} (K\hat{\beta} - B) \tag{2.82}$$

hvor C er en kendt matrix og $b \in \mathbb{R}^p$. Og z-statistik følger en asymptotisk normalfordeling med middelværdi 0 og kovariansmatrix I.

Eksempel 2.6 Konfidens intervaller for faste parametre

Lad $Y \sim N(X\beta, W(\psi))$, hvor $W(\psi)$ er kendt. En ækvivalent inferens baseret på transformerede data $\tilde{Y} = L^{-1}Y$ er $\tilde{Y} \sim N(\tilde{X}\beta, \sigma^2 I)$, hvor $W = LL^T$ som er Cholesky-faktorisering af W.
Derudover er ML for $\mu = X\beta$ og β er

$$\begin{aligned} \hat{\mu} &= X(X^T W(\hat{\psi})^{-1} X^T W(\hat{\psi}))^{-1} Y \\ \hat{\beta} &= (X^T W(\hat{\psi})^{-1} X^T W(\hat{\psi}))^{-1} Y = (X^T W(\hat{\psi})^{-1} X^T W(\hat{\psi}))^{-1} \hat{\mu} \end{aligned} \tag{2.83}$$

Da $\hat{\beta} \sim N(\beta, \sigma^2(XW(\hat{\psi})^{-1}X^T)^{-1})$, følger REML-estimat for σ, $\tilde{\sigma}$ en $\frac{\chi^2(n-d)}{(n-d)}$ fordeling. Vi kan desuden opstille approksimative $(1-\alpha)$-konfidensintervaller således

$$\hat{\beta} \pm z_{\alpha/2}\sigma\sqrt{(XW(\hat{\psi})^{-1}X^T)^{-1}} \tag{2.84}$$

hvor $z_{\alpha/2}$ er $\alpha/2$-fraktilen for standard normalfordelingen. ◀

2.3.3 Missing data at random (MAR) og missing data completely at random (MCAR)

Statistisk inferens baseret på data med manglende værdier, er bygget på visse antagelser om manglende datamekanisme. Validiteten af disse antagelser kræver evaluering inden man analyserer data. For eksempel inferens baseret på likelihood er kun valid, hvis missing data er *Missing at random (MAR)* [3, s. 795]. MAR antager at manglende data er afhængig af observerede data, men uafhængig af de uobserverede data. Det vil sige, at det at teste for MAR

kræver information om de manglende data, hvilket gør det svært at teste for MAR. Vi tester i stedet for, at missing data er *Missing completely at random (MCAR)*, hvor man antager at manglende data hverken er afhængig af observerede eller uobserverede data. I dataanalyser er *Little's test* brugbar til at undersøge antagelsen om manglende data for multivariate observerede data er MCAR [3, s. 795].
Vi vil nu definere MAR og MCAR. Vi betragter derfor en population med n observationer, og $y = (y_{i1}, \ldots, y_{ip})^T$ som en p-dimensional vektor for $i = 1, \ldots, n$ og at $Y = (y_1, \ldots, y_n)^T$ er en $n \times p$ data matrix. Vi er interesserede i at teste om Y er MCAR. Vi lader derfor Y_m betegne manglende indgange af Y, og Y_o betegne observerede indgange af Y. Vi kan også i nogle tilfælde have fuldt observerede variable x som q-dimensional vektor. Vi lader derfor X være en $n \times q$ data matrix af variable værdier. Vi lader også $r = (r_{i1}, \ldots, r_{ip})^T$ være en p-dimensional vektor, som betegner en indikator for om en komponent y_{ik} er mangle eller ej, altså

$$r_{ik} = \begin{cases} 1 & \text{hvis } y_{ik} \quad \text{er observeret} \\ 0 & \text{hvis } y_{ik} \quad \text{er missing} \end{cases} \tag{2.85}$$

for $i = 1, \ldots, n$ og $k = 1, \ldots, q$. Den stacked matrix af r er $R = (r_1, \ldots, r_n)^T$. Da definerer MAR antagelsen som,

$$\mathbb{P}\left(R \mid Y_m, Y_o, X\right) = \mathbb{P}\left(R \mid Y_o, X\right) \tag{2.86}$$

Så med andre ord er fordelingen af manglende værdier kun afhængig af de obserevede data [3, s.796].
En stærkere antagelse er MCAR defineret som,

$$\mathbb{P}\left(R \mid Y_m, Y_o, X\right) = \mathbb{P}(R) \tag{2.87}$$

dette medfører at fordelingen af indikatorer hverken er afhængig af observerede eller uobserverede data [3, s.796].

Test for MCAR

I Little's test af MCAR, antager vi at data y_i for $i = 1, \ldots, n$ p-dimensionelle multivarit er normalfordelte med middelværdivektor μ og kovarians matrix Σ, hvor en del komponenter y_i mangler [3, s. 797]. Vi antager at der i alt er J værdi mønstre som mangler (antallet af variable, hvor der mangler nogle værdier) af y_i'er. Vi antager, at for hvert manglende mønster j, er der to indeksmængde, som er observerede o_j og manglende m_j komponenter, og $p_j = |o_j|$ er antallet af observerede komponenter i mønstret j. Vi lader endvidere μ_{o_j} være $p_j \times 1$ middelværdivektor og Σ_{o_j} være $p_j \times p_j$ kovarians matrix for observerede komponenter i det j'te manglende mønster. Vi lader endelig $I_j \subseteq \{1, \ldots, n\}$ være en indeksmængde af j og $n_j =\mid I_j \mid$ hvor $n = \sum_{j=1}^{J} n_j$.
Little's χ^2 teststatistik for MCAR er på følgende formel,

$$d_0^2 = \sum_{j=1}^{J} n_j \left(\overline{y}_{o_j} - \mu_{o_j}\right)^T \Sigma_{o_j}^{-1} \left(\overline{y}_{o_j} - \boldsymbol{\mu}_{o_j}\right) \tag{2.88}$$

Hvis data er MCAR givet r_i, så er den følgende nulhypotese H_0 opfyldt,

$$H_0: \quad y_{o,i} \mid r_i \sim N(\mu_{o_j}, \Sigma_{o_j}) \quad \text{hvis} \quad i \in I_j, 1 \leq j \leq J \quad \text{hvor} \quad \mu_{o_j} \subseteq \mu \tag{2.89}$$

Hvis data dog ikke er MAR givet indikatoren r_i, forventer vi, at middelværdien for y varierer på tværs af forskellige mønstre, altså den alternative hypotese er

$$H_1: \quad y_{o,i} \mid r_i \sim N(\nu_{o_j}, \Sigma_{o_j}) \quad \text{hvis} \quad i \in \boldsymbol{I}_j, 1 \leq j \leq J \tag{2.90}$$

hvor ν_{o_j}, $j = 1, \ldots, J$ er middelværdivektor for hvert mønster j.
Bemærk, at forkastelse af antagelsen om MAR er tilstrækkelig og ikke nødvendig for at kunne forkaste antagelsen om MCAR [3, s. 797].
Det kan bevises, at d_0^2 er en likelihood-ratio teststørrelse brugt til at teste antagelsen (2.86) mod alternativ hypotese H_1 i (2.90). Hvis normalitetsantagelsen er opfyldt, følger d_0^2 en χ^2 fordeling med frihedsgrad $df = \sum_{j=1}^{J} p_j - p$. Men hvis y_i'er ikke er multivariat normalfordelt, dog med den samme middelværdi μ og kovarians matrix Σ, så kan vi bruge *The multivariate central limit theorem (CLT)* til at sige, at under antagelsen om MCAR, følger d_0^2 asymptotisk den samme χ^2 fordeling.
Bemærk, at både μ og Σ i praksis er ukendte. I Little's test foreslåes derfor at benytte deres unbiased estimater $\hat{\mu}$ og $\tilde{\Sigma} = n\hat{\Sigma}/(n-1)$ i stedet, hvor $\hat{\mu}$ og $\hat{\Sigma}$ er maksimum likelihood estimater baserede på nulhypotesen H_0 defineret i (2.89). Yderligere erstatter vi Σ_{o_j} i (2.88) med $\widetilde{\Sigma}_{o_j} \subseteq \widetilde{\Sigma}$, som medfører, at

$$d^2 = \sum_{j=1}^{J} n_j \left(\overline{y}_{o_j} - \widehat{\mu}_{o_j}\right)^T \widetilde{\Sigma}_{o_j}^{-1} \left(\overline{y}_{o_j} - \widehat{\mu}_{o_j}\right) \tag{2.91}$$

hvilket medfører, at d følger asymptotisk en χ^2 fordeling med frihedsgrad $df = \sum_{j=1}^{J} p_j - p$. Endvidere forkaster vi (2.86), hvis $d^2 > \chi^2_{df}(1-\alpha)$, hvor α er signifikans niveau (ofte $\alpha = 0.5$) [3, s. 798].

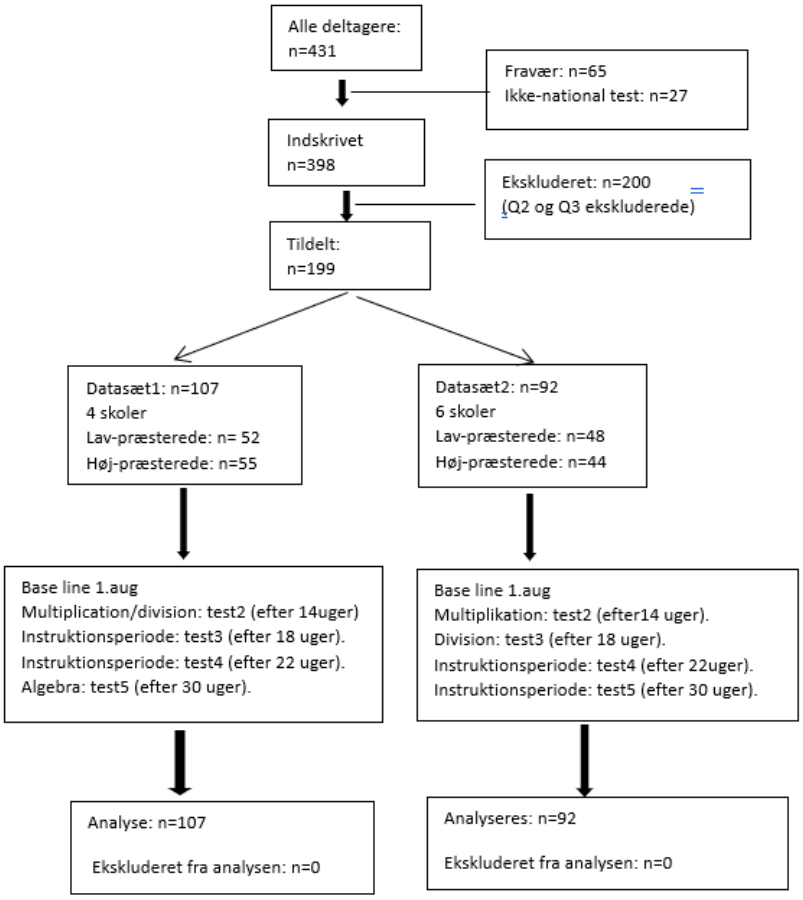
2.4 Dataanalysen (LMM-analysen)

Formålet med dette afsnit er at afprøve de på nuværende tidspunkt beskrivende metoder på et praktisk givet datasæt, som indeholder målinger på 431 elever fra 21 forskellige 4.klasser fra 11 skoler. Dette datasæt indeholder informationer i 10 variabler, som kan ses i følgende tabel.

Vaiabel	Forklaring
id	elevs id
time	tiden for test
test	testscoren
gender	elevens køn (0:pige, 1:dreng)
age-2018	elevens alder i 2018
class-id	klasser id (21 forskellige klasser)
School-id	skoler id (1-11)
intervention	eleverne i gruppe1 begyndte forløbet først. Eleverne i gruppe2 senere.
reading-score-xtile	baggrundsvariable, vi ikke får brug for i analysen.
$group_1$	$group_1$ er lavt-præsterende elever i interventionsgruppe1 og
$group_2$	$group_2$ er højt-præsterende elever i interventionsgruppe1

Tabel 2.1: Variablene i data

Vi er interesserede i at undersøge forskellen mellem højt-præsterende og lavt-præsterende elever i udviklingen af deres brøkregningsfærdigheder i løbet af et skoleår (2018-2019). Vi er også interesserede i at se, hvordan disse forskelle er relateret til undervisningen af brøkregning og undervisningen af andre matematiske emner, såsom geomtri og algebra. Altså vil vi svare på spørgsmålet; *har lavt- og højt-præsenterende elever forskellige udviklinger i deres brøkregningsfærdigheder i løbet af fjerde klasse? Og hvis der er forskel, hvordan er forskellen relateret til perioderne for undervisning i andre matematisk emner, såsom multiplikation, division, brøker og ligninger, som understøtter udviklingen i brøkregningsfærdigheder?*
Elevernes brøkregningsniveau er målt i otte måneder ved fem målingstidspunkter. Undersøgelsen er designet således, at den giver os mulighed for at observere og sammenligne udviklingen af matematikindlæring af en prøvepopulation i denne periode, idet alle klasserne fulgte den samme pensumstruktur for de emner, der blev introduceret i løbet af skoleåret. Alle de deltagende elever modtog lignende undervisning i multiplikation, division, brøker og ligninger.
Deltagende elever er fra 11 forskellige skoler i Region Midtjylland i Danmark. Disse skoler følger forskellige undervisningsplaner, og derfor var der en forsinkelse /forskydning i brøkregningsundervisning for en periode på fire uger i seks skoler. En af skolerne, hvor der var en klasse med 27 elever er blevet ekskluderet af studiet, idet den er privatskole og derfor følger en helt anden undervisningsplan end de 10 andre almindelige folkeskoler. Endvidere mangler der målinger for 6 elever på grund af fravær i de dage, hvor nationaltesten i brøkregning er blevet lavet. Karakteristisk for nationaltesten er, at testen har et adaptivt design, hvorved hvert nyt spørgsmål blev valgt ud fra elevens tidligere svar, hvilket betyder, at den målte sværhedsgrad for hver elev blev bestemt ud fra den pågældende elevs tidligere svar. Målingerne for de 10 folkeskoler er samlet i to forskellige datasæt ($data_1$ og $data_2$), hvor $data_1$ indeholder målinger fra fire skoler, som var de første, der begyndte med brøkregningsundervisning. Dette datasæt indeholder 55 lavt-præsterende elever og 52 højt-præsterende elever. $Data_2$ indeholder så målinger fra de 6 resterende skoler, som har forsinket deres brøkregningsundervisning. I dette datasæt er der 48 lavt-præsterende elever og 44 højt-præsterende elever. Disse informationer er illustreret i Figur 2.1.



Figur 2.1: Overblik over deltagerne i $data_1$ og $data_2$.

Da planlægning af skoleåret og undervisningsplanerne varierer i de forskellige skoler, er testresultaterne for skolerne i $data_1$ og i $data_2$ ikke helt til at sammenligne. Vi analyserer derfor resultater for de to datasæt hver for sig, men vi benytter den samme strategi for begge analyser. Vi kan altså benytte $data_2$ til at støtte de observerede mønstre i $data_1$. En illustration af de to forskellige undervisningsplaner kan ses i den følgende tabel,

Skoler: n=4	multiplikation/ division	brøker (introduktion)	brøker (komplekse)	ligninger	
Forsinkede skoler: n=6	multiplikation/ division		brøker (introduktion)	brøker (komplekse)	ligninger
	$test_1$ August baseline	$test_2$ November 14 uger efter baseline	$test_3$ December 18 uger efter baseline	$test_4$ Februar 22 uger efter baseline	$test_5$ April 30 uger efter baseline

Tabel 2.2: Brøkregningsfærdighedertests i løbet af et skoleår. 4 skoler danner $data_1$, og de 6 forsinkede skoler danner $data_2$.

Udfra målingerne for disse 10 folkeskoler, konstruerer vi to forskellige grupper baseret på elevernes resultater i dansk nationaltest. Den ene gruppe indeholder 25% af elever med flest point ($n = 99$) svarende til 25%-fraktilen, mens den anden gruppe indeholder 25% af elever med færrest point ($n = 100$) svarende til 75% fraktilen. I alt er der 199 observationer; 106 piger og 93 drenge, og gennemsnittet af elevernes alder er 10 år og 4 måneder med standard deviation på $sd = 0.028$.

Lineære mixed modeller

For at undersøge validiteten af vores lineære mixede model på trods af manglende data, undersøger vi om de manglende data er MAR. Til dette formål benytter vi Little's test for MCAR beskrevet i (2.3.3), hvor vi betragter den kategoriske variabel med niveauer givet af de fire grupper som en uafhængig variabel (højt- og lavt-præsenterende i $data_1$, og højt- og lavt-præsenterende i $data_2$). Resultaterne for Little's test fra R

```
 n        nIncomp        nPattern      chi^2       df        pval
431        179         21             145.97      137       0.284
```

viser, at de manglende målinger er MAR, idet $\chi^2(137) = 145.97$ og p-værdi $p = 0.284$.

Vi forsøger nu blot at lave en passende lineær mixed model for vores datasæt, hvor vi betragter variablene `time`, `group` og `gender` som faste effekter, og variablene `id` og `class-id` som stokastiske effekter i modellen. Vi kan udtrykke modellen som,

$$test_{it} = \beta_0 + \beta^G_{g(i)} + \beta^{Gen}_{\texttt{gen(i)}} + \lambda_t + \delta_{i:t} + U_i + \epsilon_{it} \tag{2.92}$$

hvor β_0 er skæringen, $\beta^G_{g(i)}$ er gruppe effekt for det i'te individ, $\beta^{Gen}_{\texttt{gen(i)}}$ er kønseffekt for det i'te individ, λ_t er effekten af variablen `time`, $\delta_{i:t}$ er vekselvirkningen mellem variablene `time` og `group` for det i'te individ ved tiden t, U_i er stokastiske effekter og ϵ_{it} er et fejlled for det $i$te individ ved tiden t. Derudover er $effekt(\texttt{id}) \sim N(0, \tau_1^2)$ og $effekt(clas_{id}) \sim N(0, \tau_2^2)$. Bemærk, at τ_1^2 er variansen mellem eleverne, τ_2^2 er variansen mellem klasserne og $\tau^2 = \tau_1^2 + \tau_2^2$. Bemærk, at man også kunne inkludere variablen `School` som en stokastisk effekt. Men idet der ikke er observeret variation mellem skolerne, har vi ikke gjort det.

```
   Groups     Name         Std.Dev.
id         (Intercept) 4.2894e+00
clas_id    (Intercept) 1.0858e+00
School_id  (Intercept) 3.8272e-05
Residual               2.9587e+00
```

Vi vil nu analysere de to datasæt ($data_1$ og $data_2$) hver for sig, hvor vi benytter $data_2$ til at validere de observerede mønstre i $data_1$. Analysen laver vi i R og udvalgte kode kan ses i Appendix (A).

2.4.1 Data1

For at vurdere betydningen af variablene i modellen i (2.92), kan vi estimere koefficienterne for hver af variablene (`time`, `group` og `gender`) og deres vekselvirkning (`time`: `group`) i modellen. Estimater for disse koefficienter, kan ses i tabel (2.3).

Variabel	Coef.	Std. Error	t value	$Pr(>\mid t \mid)$	
`intercept`	4.6597	0.9046	5.151	3.51e-06	***
`time2`	-0.0295	0.6302	-0.047	0.96269	
`time3`	3.5654	0.6329	5.634	3.67e-08	***
`time4`	4.1767	0.6724	6.212	1.50e-09	***
`time5`	3.8621	0.6336	6.096	2.92e-09	***
`group2`	5.1827	1.0616	4.882	2.38e-06	***
`gender1`	3.9660	0.9084	4.366	3.03e-05	***
`time2:group2`	2.4845	0.8650	2.872	0.00433	**
`time3:group2`	2.0027	0.8697	2.303	0.02188	*
`time4:group2`	3.8427	0.9105	4.221	3.12e-05	***
`time5:group2`	4.3106	0.8678	4.967	1.07e-06	***

Tabel 2.3: Koefficienterne for faste effekter i $data_1$, hvor signifikanskoder: 0 '***', 0.001 '**', 0.01 '*'.

Faste effekter:

For at finde de gennemsnitelige forventede ændringer i testsresultater fra en tidsmåling til den efterfølgende tidsmåling og de mulige forskelle i disse ændringer mellem lavt- og højt-præsterende elever, anvender vi tabel (2.3). Udfra denne tabel kan vi aflæse en skæring $\beta_{low} = 4.66$ svarende til gennemsnittet af testresultater for lavt-præsterende elever ved base-line ($time_1$), og en groupe effekt $\beta^G_{g(i)} = \beta_{high} - \beta_{low} = 5.18$ svarende til den gennemsnitlige forskel i testresultater mellem lavt- og højt-præsterende elever. Altså testresultaterne for højt-præsterende elever er i gennemsnit 5.18 point højere end for lavt-præsterende elever. Yderligere kan vi udfra tabel (2.3) aflæse en signifikant kønseffekt $\beta^{Gen}_{gen(i)} = 3.97$ svarende til den forventede forskel i testsresultater mellem drenge og piger. Altså testsresultaterne for drengene (`gender` $= 1$) er i gennemsnit 3.97 point højere end for pigerne.

Vi kan udfra tabel (2.3) finde λ, som er den gennemsnitelig tidseffekt af testresultater. Altså λ_{12} er den forventede gennemsnitlige ændring af testresultater fra $time_1$ til $time_2$, λ_{23} er den forventede gennemsnitlige ændring af testresultater fra $time_2$ til $time_3$, λ_{34} er den forventede gennemsnitlige ændring af testresultater fra $time_3$ til $time_4$) og λ_{45} er den forventede gennemsnitlige ændring af testresultater fra $time_4$ til $time_5$). Vi beregner λ'erne ved at benytte tabel (2.3), hvor resultaterne kan ses i følgende tabel.

λ	Coef.	Pr(>|t|)
λ_{12}	-0.03	0.96
λ_{23}	3.57-(-0.03)=3.60	<0.0001
λ_{34}	4.18-3.57=0.61	0.38
λ_{45}	3.86-4.18=-0.32	0.65

Tabel 2.4: De gennemsnitelige forventede ændringer i testresultater fra en tidsmåling til den efterfølgende tidsmåling for lavt- og højt-præsenterende elever i $data_1$. pr(>|t|) er p-værdierne for λ'erne baseret på t-test, hvor nulhypotesen er $H_0 : \lambda = 0$.

Vekselvirkning:

Vi er interesserede i at finde vekselvirkningskoefficienter, der kan fortolkes som forskellene i effekterne af variablene i modellen, afhængig af værdierne af de andre variable.
Vekselvirkningen δ mellem *time* og *group* i modellen defineret i (2.92) svarende til forskellen mellem ændringer for lavt- og højt-præsterende elever, kan vi aflæse fra tabel (2.3). Altså er $\delta_{12} = 2.48$ forskellen mellem ændringer for lavt- og højt-præsterende elever i perioden fra baseline til $time_2$, δ_{23} er forskellen mellem ændringer for lavt- og højt-præsterende elever i perioden fra $time_2$ til $time_3$, δ_{34} er forskellen mellem ændringer for lavt- og højt-præsterende elever i perioden fra $time_3$ til $time_4$ og δ_{45} er forskellen mellem ændringer for lavt- og højt-præsterende elever i den sidste periode (fra $time_4$ og $time_5$). Værdierne for δ kan vi få ud fra tabel (2.3) og kan ses i følgende tabel,

δ	Coef.	Pr(>|t|)
δ_{12}	2.48	0.004
δ_{23}	2.00-2.48=-0.48	0.59
δ_{34}	3.84-2.00=1.84	0.05
δ_{45}	4.31-3.84=0.47	0.62

Tabel 2.5: Forskellen i ændringer mellem lavt- og højt-præsterende elever i $data_1$. pr(>|t|) er p-værdierne for λ'erne baseret på t-test, hvor nulhypotesen er $H_0 : \delta = 0$.

Vi definerer et nyt parameter $\eta = \lambda + \delta$, som betegner ændringerne mellem fortløbende gennemsnet af testsresultater for lavt- og højt-præsterende grupper. Værdierne af η kan vi beregne ved at sætte værdierne af λ og δ i taballerne (2.4) og (2.5) sammen, og de ses i følgende tabel,

η	Coef.	Pr(>|t|)
η_{12}	-0.03+2.48=2.45	<0.0001
η_{23}	3.60-0.48=3.11	<0.0001
η_{34}	0.61+1.84=2.45	<0.0001
η_{45}	-0.32+0.47=0.15	0.81

Tabel 2.6: Ændringer mellem de gennemsnitelige fortløbende testresultater for lavt- og højt-præsterende elever i $data_1$. pr(>|t|) er p-værdierne for λ'erne baseret på t-test, hvor nulhypotesen er $H_0 : \eta = 0$.

Stokastiske effekter

En vigtig del af lineære mixede modeller er at estimere varianskomponenter, der angiver, hvor meget variation, der er i dataerne på tværs af grupper eller gentagne målinger. Disse varianskomponenter kan fortolkes som den mængde variation, der ikke kan forklares af de faste effekter i modellen. Resultaterne for stokastiske effekter baseret på modellen i (2.92), kan ses i den følgende tabel,

Grouppe	effekt	Varians	Std. Dev.
id	intercept	18.399	4.289
$clas_{id}$	intercept	1.179	1.086
Residual		8.754	2.959

Tabel 2.7: Resultatet fra R for estimationen for varianskomponenter i $data_1$, hvor `Number of obs: 456, groups: id, 107` ; $clas_id$, 11.

Ud fra tabel (2.7) kan vi aflæse estimatet for variansen imellem eleverne (indenfor-id varians) $\sigma^2 = 8.75$, estimatet for variansen mellem eleverne $\tau_1^2 = 18.40$ og estimatet for variationen mellem klasserne $\tau_2^2 = 1.18$. Bemærk her at total varians er $V = \tau_1^2 + \tau_2^2 + \sigma^2 = 28.33$, og dermed er den største andel (65%) af den samlede varians mellem eleverne. Variationen mellem klasserne er derimod relativt lille (4% af den samlede varians).

Hypotesetests

For at teste nulhypotesen om, at der ikke er nogen forskelle mellem ændringer i testsresultater for lavt- og højt-præsterende elever, altså

$$H_0: \quad \delta_{12} = \delta_{23} = \delta_{34} = \delta_{45} = 0 \tag{2.93}$$

, konstruerer vi et F-test i R således,

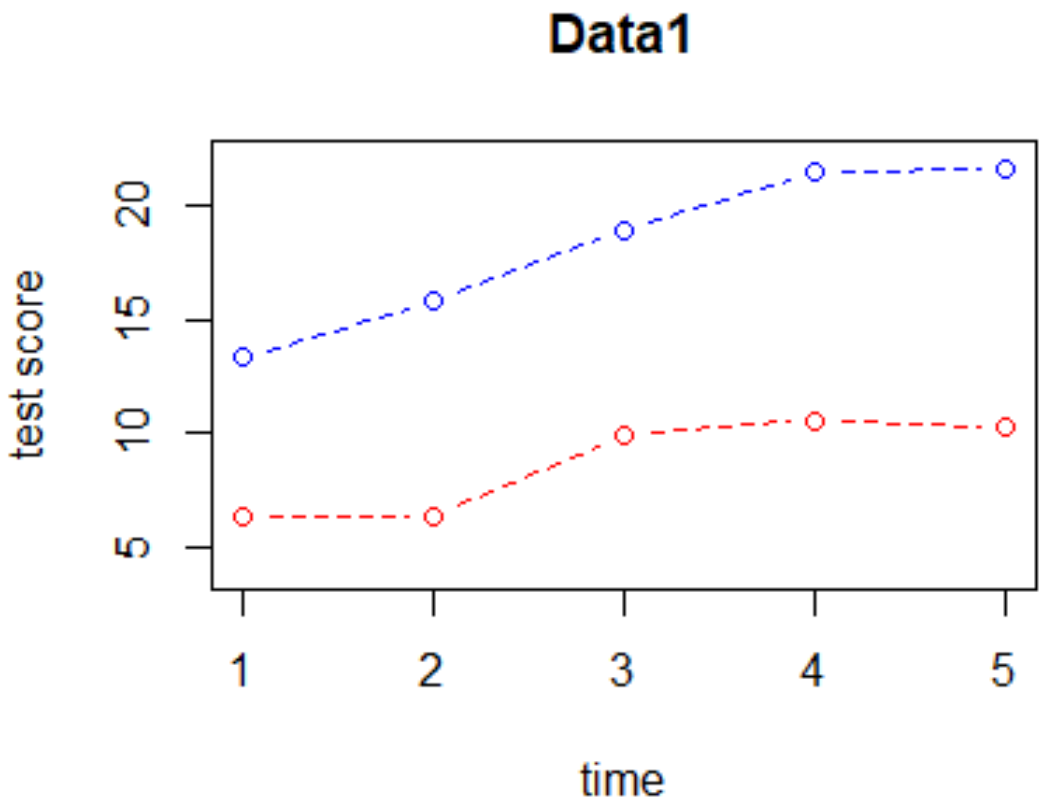
```
> drop1(lme1,test="F")
Single term deletions using Satterthwaite's method:

Model:
test ~ factor(time) * factor(group) + factor(gender) + (1 |id) + (1 |clas_id)
                          Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
factor(gender)            166.84 166.842     1 102.82 19.0595 3.026e-05 ***
factor(time):factor(group) 264.25  66.062     4 344.69  7.5467 7.743e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Resultaterne for denne test er markant signifikant med p-værdi $p < 0.0001$. Vi vil derfor forkaste nulhypotesen for, at der ikke er nogen forskelle mellem ændringer i testresultater for lavt- og højt-præsterende elever. Det vil sige, at udviklingen for brøklæring er forskellig for lavt- og højt-præseterende elever.

Bemærk, at forskellen mellem testresultater for lavt- og højt-præsterende elever primært sker i den første periode (mellem baseline og $\texttt{time}_2$) δ_{12}, idet p-værdien for δ_{12} ($p = 0.004$) er den enste signifikant p-værdi blandt p-værdierne for δ_{12} til δ_{45} (i tabel (2.5)).
Udviklingen i testresultater for lavt-præsterende elever i den første periode $\lambda_{12} = -0.03$ er statistiklig usignifikant forskellig fra 0 med p-værdi $p = 0.96$, mens udviklingen i testresultater for højt-præsterende elever i den første periode er estimeret til at være $\eta_{12} = 2.45$ og signifikant forskeligt fra 0 med p-værdi $p < 0.0001$ (fra tabel (2.6)).

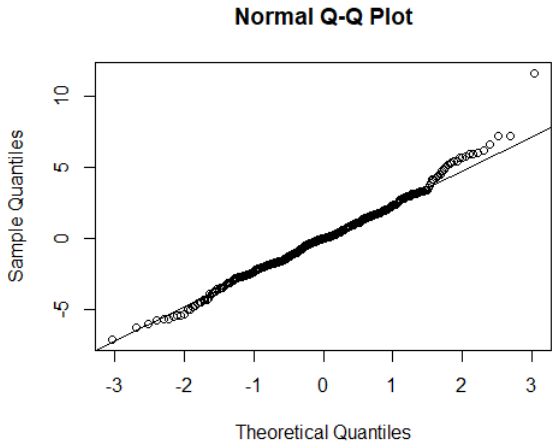
Betragter vi en stigende udvilkling i testresultater for lavt-præsterende elever, er ændringen i den anden testperiode $\lambda_{23} = 3.59$ den enste ændring der er signifikant forskeligt fra 0 ($p < 0.0001$) mellem ændringerne λ_{12} til λ_{45}.
For højt-præsterende elever er der observeret en siginfikant stigende udvikling i testresultater i første tre, hvor disse udviklinger er etimerede til at være $\eta_{12} = 2.45$, $\eta_{23} = 3.11$ og $\eta_{34} = 2.45$ med p-værdier $p < 0.0001$ for alle tre perioder. Men udviklingen over den sidste periode for højt-præsterende elever er estimeret til at være η_{45} og er usignifikant forskelligt fra 0. Udviklingen i testresultater for både lavt- og højt-præsterende elever er illustrerede i følgende plot.



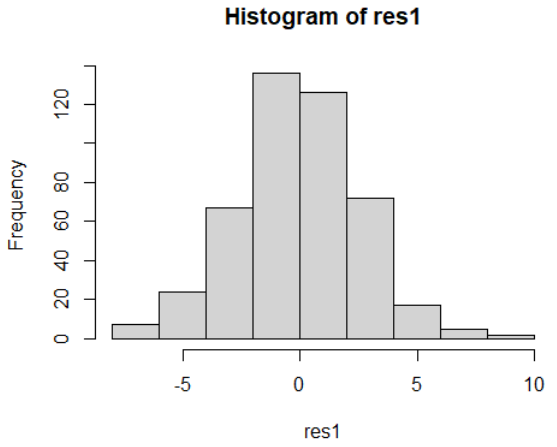
Figur 2.2: Multipel lineær regression model for $data_1$, hvor den rød linje viser udviklingen af testresultater for lavt-præsterende elever og den blå linje viser udviklingen af testresultater for højt-præsterende elever over de 5 tidsmålinger.

Model validering

Sluttelig er det vigtigt at undersøge, om fordelingsantagelserne for residualerne og stokastiske effekter er opfyldt. Til dette formål laver vi en grafisk undersøgelse, hvor vi først laver et QQ-normal plot og et histogram af residualerne, som ses i henholdsvis figuer (2.3) og (2.9).

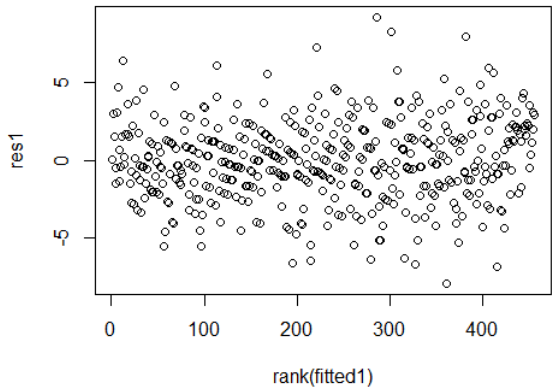


Figur 2.3: QQ-normal plot for residualerne i $data_1$.



Figur 2.4: Histogram for residualer i $data_1$.

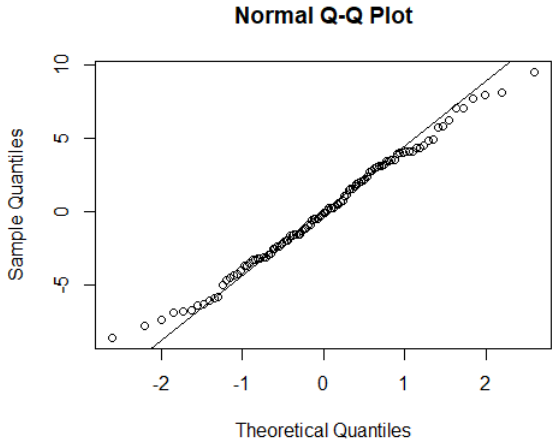
Plottene i (2.3) og (2.4) indikerer, at residualerne følger en normalfordeling med middelværdi 0.
For at undersøge outliers i residualerne, plotter vi residualerne mod de fittede værdier, som kan ses i figur (2.5).



Figur 2.5: Plot af residualerne mod de fittede værdier i $data_1$.

Det ovenstående plot indikerer, at der er outliers i residualer.

For at undersøge fordelingen for stokastiske effekter, finder vi kun BLUP-estimatet for stokastiske effekter, og vi tjekker derefter normalitetsantagelsen ved at lave et QQ-normal plot, som kan ses i figur (2.6).



Figur 2.6: QQ-normal plot for stokastiske effekter i $data_1$.

Det ovenstående plot indikerer, at stokastiske effekter følger en normalfordeling med middelværdi 0 .

2.4.2 Data2

Vi vil som før nævnt bruge den samme strategi, vi har brugte til at analysere $data_1$ til at analysere $data_2$. Vi vil indledningsvis tilpasse en linær mixed model defineret i (2.92) i R.

Estimaterne for de gennemsnitlige effekter for de forskellige variable for denne model kan ses i tabel (2.8).

Variabel	Coef.	Std Error	t value	$Pr(>\|t\|)$	
`Intercept`	5.8790	0.8407	6.993	2.05e-08	***
`time2`	-0.2711	0.5983	-0.453	0.650796	
`time3`	0.6067	0.6155	0.986	0.325014	
`time4`	3.9016	0.6235	6.257	1.26e-09	***
`time5`	3.8520	0.6022	6.397	5.65e-10	***
`group2`	4.2806	1.0504	4.075	7.42e-05	***
`gender1`	0.7459	0.9430	0.791	0.431083	
`time2:group2`	3.9142	0.8649	4.525	8.53e-06	***
`time3:group2`	4.1010	0.8803	4.659	4.68e-06	***
`time4:group2`	3.2930	0.9139	3.603	0.000364	***
`time5:group2`	4.6553	0.8675	5.366	1.55e-07	***

Tabel 2.8: Koefficienterne for faste effekter i $data_2$, hvor signifikanskoder: 0 '***', 0.001 '**', 0.01 '*'.

Faste effekter:

Udfra tabel (2.9) kan vi aflæse en skæring på $\beta_0 = \beta_{low} = 5.89$ svarende til gennemsnittet af testresultater for lavt-præsterende elever ved baseline, og en gruppe effekt $\beta^G_{g(i)} = \beta_{high} - \beta_{low} = 4.28$ svarende til den gennemsnitelige forskel i testresultater mellem lavt- og højt-præsterende elever. Det vil sige, højt-præsterende elever har i gennemsnit 4.28 points højere end det er for lavt-præsterende elever. derudover kan vi aflæse en kønseffekt på $\beta^{Gen}_{gen} = 0.75$, dog er denne effekt statistik usignifikant med p-værdi $p = 0.43$. Dette indikerer, at der ikke forventes nogen forskel mellem testresultater for piger og drenge i $data_2$.

Udraf tabel (2.9) kan vi også finde λ, som betegner (ligesom for $data_1$) den gennemsnitelige tidseffekt i testresultater. Værdierne af λ kan ses i følgende tabel

λ	Coef.	Pr(>|t|)
λ_{12}	-0.27	0.65
λ_{23}	0.61-(-0.27)=0.88	0.16
λ_{34}	3.90-0.61=3.29	<0.0001
λ_{45}	3.85-3.90=-0.05	0.94

Tabel 2.9: De gennemsnitelige forventede ændringer i testresultater fra en tidsmåling til den efterfølgende tidsmåling i $data_2$. Pr(>|t|) er p-værdierne for λ'erne baseret på t-test, hvor nulhypotesen er $H_0 : \lambda = 0$.

Vekselvirkning

Vi finder nu vekselvirkningskoefficienter δ mellem variablene `time` og `group` i modellen (2.92). Denne vekselvirkning svarer til forskellen mellem ændringer i testresultater for lavt- og højt-præsterende elever. Altså ligesom for $data_1$ lader vi δ_{12} at være forskellen i ændringerne mellem lavt- og højt-præsterende elever i perioden fra baseline til $\texttt{time}_2$, δ_{23} forskellen i ændringer mellem lavt- og højt-præsterende elever i perioden fra $\texttt{time}_2$ til $\texttt{time}_3$, δ_{34} forskellen

i ændringer mellem lavt- og højt-præsterende elever i perioden fra $\texttt{time}_3$ til $\texttt{time}_4$, og δ_{45} forskellen i ændringer mellem lavt- og højt-præsterende elever i perioden fra $\texttt{time}_4$ til $\texttt{time}_5$. Værdierne for δ for $data_2$ kan vi få udfra tabel (2.8) og kan ses i følgende tabel,

δ	Coef.	Pr(>|t|)
δ_{12}	3.91	<0.0001
δ_{23}	4.10-3.91=0.19	0.83
δ_{34}	3.29-4.10=-0.81	0.39
δ_{45}	4.66-3.29=1.37	0.14

Tabel 2.10: Forskellen i ændringerne mellem lavt- og højt-præsenterende elever i $data_2$. Pr(>|t|) er p-værdierne for δ'erne baseret på t-test, hvor nulhypotesen er $H_0 : \delta = 0$.

Vi definerer ligsom for $data_1$ en ny parameter η, som betegner forskellen mellem fortløbende gennemsnit af testresultaterne for lavt- og højt-præsterende elever. Værdierne for η kan ses i følgende tabel,

η_{12}	η_{12}	Pr(>|t|)
η_{12}	-0.27+3.91=3.64	<0.0001
η_{23}	0.88+0.19=1.07	0.10
η_{34}	3.29-0.81=2.48	<0.0001
η_{45}	-0.05+1.37=1.32	0.05

Tabel 2.11: Ændringerne mellem fortløbende gennemsnit af testresultater for lavt- og højt-præsenterende elever i $data_2$. Pr(>|t|) er p-værdierne for η'erne baseret på t-test, hvor nulhypotesen er $H_0 : \eta = 0$.

Stokastiske effekter

Resultaterne for stokastiske effekter baseret på modellen (2.92) for $data_2$, kan ses i følgende tabel,

`Grouppe`	`effekt`	`Varians`	`Std. Dev.`
`id`	`intercept`	15.594	3.949
$clas_{id}$	`intercept`	1.198	1.095
`Residual`		8.141	2.853

Tabel 2.12: Resultaterne fra R for estimationen for varianskomponenter i $data_2$, hvor `Number of obs:` 418 `, groups: id,` 92 ; $clas_id$, 10.

Bemærk her, at estimatet for variansen indenfor elever (indenfor-id varians) er $\sigma^2 = 8.14$. Derudover er estimatet for variansen mellem eleverne $\tau_1^2 = 15.60$ og estimatet for variansen mellem klasserne $\tau_2^2 = 1.20$. Den totale varians er derfor $V = 8.14 + 1.20 + 15.60 = 24.94$. Altså er den største andel (62%) af den samlede varians mellem eleverne. Dog er variansen mellem klasserne relativt lille (4.8%). Bemærk, at estimaterne for varianser i $data_2$ næsten stemmer overens med estimaterne for varianser i $data_1$.

Hypotesetests

Ligsom for $data_1$ konstruerer vi et F-test for $data_2$ for at undersøge nulhypotesen: at der ikke er nogen forskelle mellem ændringer i testresultater for lavt- og højt-præsenterende elever. Vi tester således nulhypotesen

$$H_0: \quad \delta_{12} = \delta_{23} = \delta_{34} = \delta_{45} = 0 \tag{2.94}$$

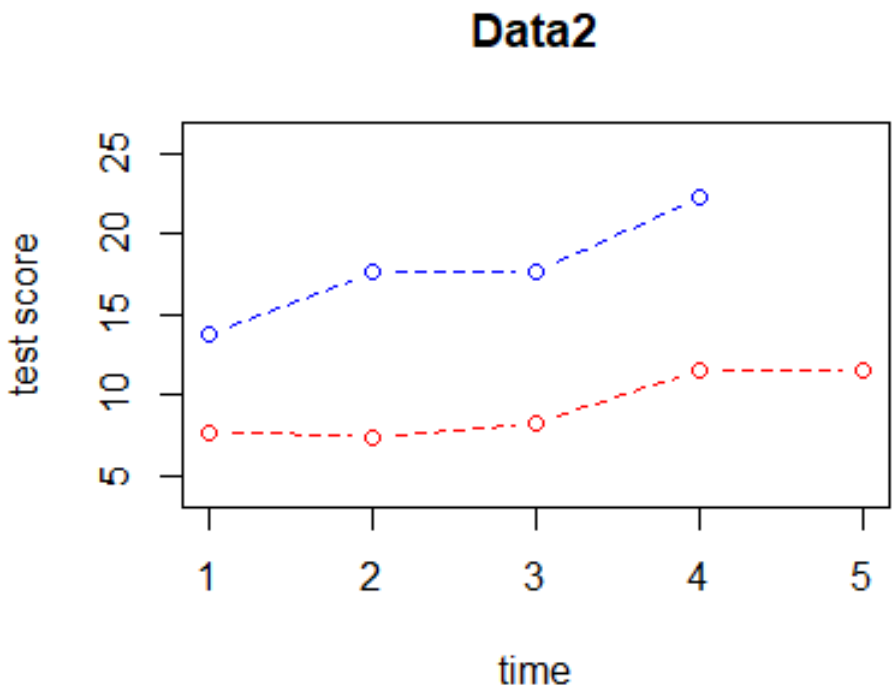
Resultater for F-testen er ligsom for $data_1$ signifikant med p-værdi $p < .0001$, og dermed forkaster vi nulhypotesen.

```
    > drop1(lme2,test="F")
Single term deletions using Satterthwaite's method:

Model:
test ~ factor(time) * factor(group) + factor(gender) + (1 |id) + (1 |clas_id)
                          Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)
factor(gender)             5.093   5.093     1  87.96  0.6257    0.4311
factor(time):factor(group) 298.472  74.618    4 319.50  9.1658 5.07e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Vi forventer derfor, at ændringer i testsresultaterne for lavt- og højt-præsenterende elever varierer over tid. Altså er udviklingen i brøklæring er forskellig for lavt- og højt-præsterende elever.

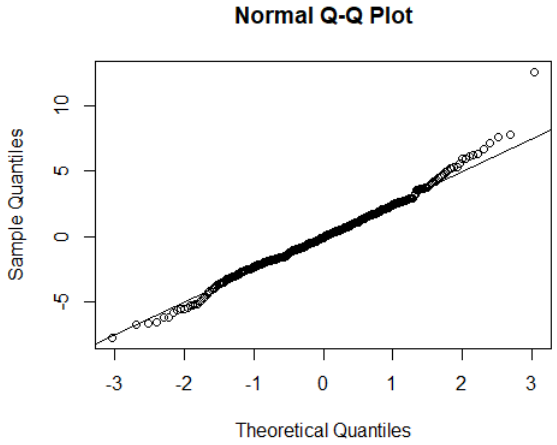
Bemærk, at forskellen i testresultater for lavt- og højt-præsterende elever er sket i den første periode (baseline og $\texttt{time}_2$) δ_{12} med en p-værdi $p < 0.0001$ (fra tabel(2.10)). Forskellen i testresultater for lavt- og højt-præsterende elever i de andre perioder δ_{23}, δ_{34} og δ_{45} er statistisk usignifikante forskellig fra 0.
Udviklingen i testresultater for lavt-præsterende elever er kun signifikant i den tredje testsperiode $\lambda_{34} = 3.29$, med p-værdi $p < 0.0001$.
For højt-præsterende elever er udviklingen i testresultater estimerede til at være positiv over alle testperioder, og dermed sker en stigende udvikling i testresultater for højt-præsterende elever. Endvidere er der signifikant forskellig fra 0 stigning i udviklingen i den første og trdje perioder $\eta_{12} =$ og $\eta_{34} =$ med p-værdierne $p < 0.0001$ (fra tabel 2.11). Udviklingen i testresultater for både lavt- og højt-præsterende elever i $data_2$ er illustreret i følgende plot,



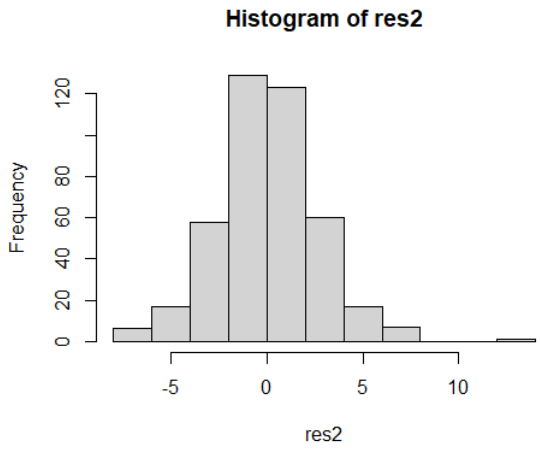
Figur 2.7: Multipel lineær regression model for $data_2$, hvor den rød linje viser udviklingen af testresultater for lavt-præsterende elever og den blå linje viser udviklingen af testresultater for højt-præsterende elever over de 5 tidsmålinger.

Model validering:

For at undersøge fordelingsantagelserne for residualerne, laver vi et QQ-normal plot og et histogram af residualerne, som kan ses i figurerne nedenunder.



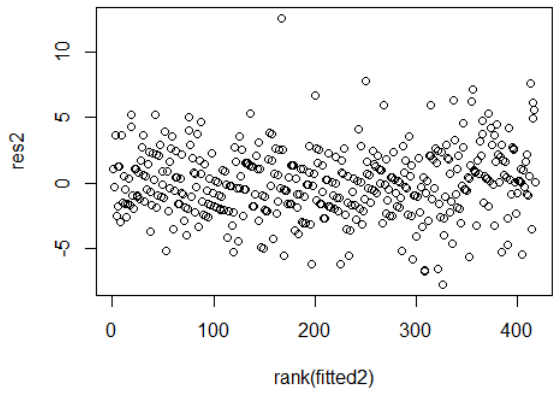
Figur 2.8: QQ-normal plot for residualer i $data_2$.



Figur 2.9: Histogram for residualer i $data_2$.

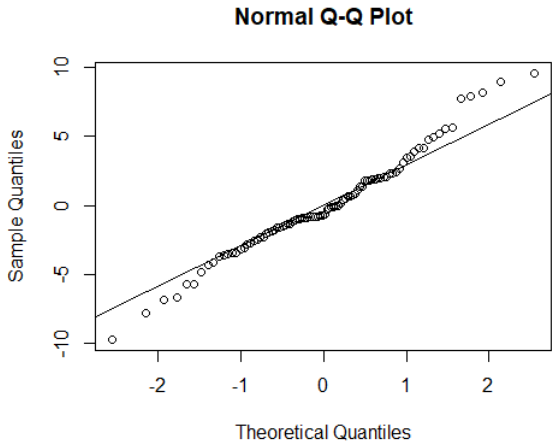
Både QQ-normal plot og histogrammet i figurerne (2.8) og (2.9) indikerer, at residualerne er normalfordelte med en middelværdi 0.

Nu plotter vi residualerne mod de fittede værdier for at undersøge outliers i residualer, dette ses i følgende figur.



Figur 2.10: Plot af residualerne mod de fittede værdier i $data_2$.

Det ovenstående plot indikerer, at der er outliers i residualer i $datast_2$ ligesom for $datast_1$. For at undersøge om fordelingsantagelse for stokastiske effekter er opfyldt, laver QQ-normal plot af disse effekter. Dette kan ses i følgende figur.



Figur 2.11: QQ-normal plot for stokastiske effekter i $data_2$.

Konklusion og diskussion:

Vi kan konkludere, at udviklingen af brøkregningsfærdigheder for lavt- og højt-præsterende elever varierer signifikant over tid. Testsresultaterne for højt-præsterende elever i brøkregningsfærdigheder er forbedret i alle perioderne, undtagen den sidste (ligningsinstruktionsperioden), hvor ændringen er meget lille for $data_1$; dog er en større ændring observeret for $data_2$. Det gælder omvendt for lavt-præsenterende elever. De har kun oplevet en forbedring i brøkinstruktionsperioden.

3 | Latent Curve Models (LCM)

I dette kapital vil vi introducere andre modeller end den lineære mixed model til at undersøge udviklingen af en variabel over tid eller alder i longitudinelle data. Disse modeller kaldes for *Latent Curve Models (LCM)*. LCM beskriver typisk, hvordan en målt variabel ændrer sig over tid og kan give overblik over, hvordan denne variabel udvikler sig og påvirkes af forskellige faktorer. En typisk anvendelse af LCM er at beskrive vækst, højde og vægt af børn over tid.

LCM indeholder stokastiske skæringer og stokastiske hældninger, som tillader, at hvert individ i populationen kan have sin egen sti over tid (en population med n individer indeholder n stier), hvor denne sti beskriver udviklingen af individet i populationen over tid. Disse stier kan være lineære funktioner af tid, men de kan også have nogle andre former, såsom kvadratiske eller kubiske funktioner af tid [2, s. 1].

Vi vil introducere LCM fra en *Structural equation model (SEM)* perspektiv og sammenligne disse modeller med lineære mixed modeller. Vi vil begynde med at definere de forskellige niveauer af LCM. For at undersøge validiteten af LCM vil vi præsentere nogle mål, der kan anvendes til at evaluere og validere modellen. Disse mål vil give os værdifuld indsigt i, hvor godt den tilpassede model passer til vores data og hvor pålidelige vores estimater er.

Til dette formål betragter vi en population bestående af n individer, hvor hvert enkelt individ er blevet målt T gange. Da defineres LCM som

$$y_{it} = \alpha_i + \lambda_t \beta_i + \epsilon_{it} \quad i = 1, \ldots, n \quad \text{og} \quad t = 1, \ldots, T \quad [2, s.126] \tag{3.1}$$

hvor y_{it} er værdien af variablen (stivariablen) y for det i'te individ ved tiden t og α_i og β_i er stokastisk skæring og hældning for den sti, der beskriver udviklingen af det i'te individ over tid. Variablen λ_t betegner en tidstendensvariabel, som tager værdier $\lambda_1 = 0$ og $\lambda_2 = 1$ for $t = 1, 2$, mens værdien af λ_t for $t \geq 3$ indikerer, om stierne er lineære eller ikke-lineære [2, s. 127]. For eksempel for en lineær LCM er $\lambda_t = t - 1$ for alle t. Og ϵ_{it} er et fejlled.
Bemærk, at Ligning (3.1) kaldes for niveau 1 ligningen for LCM eller *ubetinget LCM*. Den er ubetinget i den forstand, at den ikke tager højde for kovariater i modellen. Men vi kan ændre den ubetingede LCM til *en betinget LCM* ved at definere niveau 2 ligninger for stokastiske skæringer og hældninger. Så hvis vi antager, at der er K observerede kovariater (køn, social status, vægt, osv.) i data $x_1, \ldots, x_K$, er niveau 2 ligninger for LCM givet ved,

$$\alpha_i = \mu_\alpha + \gamma_{\alpha_1} x_{1i} + \cdots + \gamma_{\alpha_K} x_{Ki} + \zeta_{\alpha i} \tag{3.2}$$

$$\beta_i = \mu_\beta + \gamma_{\beta_1} x_{1i} + \cdots + \gamma_{\beta_K} x_{Ki} + \zeta_{\beta i} \tag{3.3}$$

hvor μ_α og μ_β er skæringer for ligninger, der prædikterer de stokastiske skæringer (3.2) og hældninger (3.3) på tværs af alle individer. Helt konkret er μ_α og μ_β henholdsvis middelværdi

af skæringer og middelværdi af hældninger, når kovariaterne i modellen er 0. $\gamma_{\alpha_1}, \ldots, \gamma_{\alpha_K}$ er kovariaters koefficienter i (3.2) og $\gamma_{\beta_1}, \ldots, \gamma_{\beta_K}$ er kovariaters koefficienter i den stokastiske hældningsligning (3.3). Kovariaterne $x_{1i}, \ldots, x_{Ki}$ er generelt tidsuafhængige. Det vil sige, at kovariaterne kan tage forskellige værdier for forskellige individer, men de ændrer sig ikke over tid. Og ζ_{α_i} og ζ_{β_i} er fejlled [2, s. 127].

LCM er bygget på nogle antagelser. Blandt andet antager vi,

1. at middelværdien af fejlene er 0, altså

$$\mathbb{E}[\epsilon_{it}] = 0 \quad \text{for} \quad i = 1, \ldots, n \quad \text{og} \quad t = 1, \ldots, T \tag{3.4}$$
$$\mathbb{E}[\zeta_{\alpha_i}] = 0 \quad \text{for} \quad i = 1, \ldots, n \tag{3.5}$$
$$\mathbb{E}[\zeta_{\beta_i}] = 0 \quad \text{for} \quad i = 1, \ldots, n \tag{3.6}$$

2. og at fejlene indbyrdes er ukorrelerede og at de er ukorrelerede med kovariaterne x_{ki}, altså er den enste kovarians der ikke 0, er $Cov(\zeta_{\alpha_i}, \zeta_{\beta_i})$.

[2, s. 127].

Vi kan kombinere niveau 1 ligningen og skærings- og hældningsligninger i en enkelt ligning ved at indsætte (3.2) og (3.3) i (3.1). På den måde får vi en *kombinationsligning* for LCM således,

$$y_{it} = (\mu_\alpha + \lambda_t \mu_\beta) + (\gamma_{\alpha_1} + \lambda_t \gamma_{\beta_1}) x_{1i} + \cdots + (\gamma_{\alpha_K} + \lambda_t \gamma_{\beta_K}) x_{Ki} + (\zeta_{\alpha i} + \lambda_t \zeta_{\beta i} + \epsilon_{it}) \tag{3.7}$$

hvor de første tre parenteser (skæring og koefficienter) i (3.7) repræsenterer de faste komponenter af LCM og den sidste parentes (fejlene) i (3.7) repræsenterer de stokastiske komponenter i LCM [2, s. 128]. Bemærk, at stien y_{it} er en funktion af en sammensat skæring, sammensatte koefficienter for kovariater $x_{1i}, \ldots, x_{Ki}$, der ændrer sig med λ_t.

Bemærk, at vi kalder den betingede LCM for LCM i dette projekt.

Matrixformel for LCM

Vi kan omskrive LCM til en matrixform, som vi vil benytte senere i estimationsafsnittet. Vi omskriver først niveau 1 ligning defineret i (3.1) for LCM til en matrixform således,

$$y_i = \Lambda \eta_i + \epsilon_i \tag{3.8}$$

hvor,

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & T-1 & T-2 & \cdots & 1 \end{bmatrix} \tag{3.9}$$

$$, \quad \eta_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \quad \text{og} \quad \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT} \end{bmatrix} \tag{3.10}$$

Bemærk, at y_t i ligning (3.8) er i en form, der tillader T tidspunkter, altså en $T \times 1$ vektor, Λ er en $T \times T$ matrix, η_i er en vektor af skærings- og hældningsværdier for det i'te individ og ϵ_i er en vektor af residualer for det i'te individ og $\epsilon_i \sim N(0, \Theta)$ [2, s. 133]. Θ er en kovarians matrix for fejlene ϵ_i således

$$\Theta = \begin{bmatrix} \theta_{\epsilon_1} & 0 & \cdots & 0 \\ 0 & \theta_{\epsilon_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{\epsilon_T} \end{bmatrix} \tag{3.11}$$

hvor $\theta_{\epsilon_t} = \text{Var}(\epsilon_{it})$ for $t = 1, \ldots, T$ og $i = 1, \ldots, n$.
Tilsvarende omskrives niveau 2 for LCM for de latente kurveparametre (α_i og β_i) på en matrixform således,

$$\eta_i = \mu_\eta + \Gamma X_i + \zeta_i \tag{3.12}$$

hvor

$$\eta_i = \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \quad , \quad \mu_\eta = \begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix} \quad , \quad \Gamma = \begin{bmatrix} \gamma_{\alpha_1} & \gamma_{\alpha_2} & \cdots & \gamma_{\alpha_K} \\ \gamma_{\beta_1} & \gamma_{\beta_2} & \cdots & \gamma_{\beta_K} \end{bmatrix} \tag{3.13}$$

$$, \quad x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{bmatrix} \quad \text{og} \quad \zeta_i = \begin{bmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{bmatrix} \tag{3.14}$$

hvor μ_{α_η} er en vektor af middelværdier for skæringer og hældninger, Γ er en $2 \times K$ matrix af koefficienter af kovariaternes effekter, x_i er $K \times 1$ vektor af kovariaterene for det i'te individ og $\zeta_i = (\zeta_{\alpha_i}, \zeta_{\beta_i})^T$ er en vektor af stokastiske effekter for det i'te individ og $\zeta_i \sim N(0, \Psi)$, hvor

$$\Psi = \begin{pmatrix} \psi_{\alpha\alpha} & \psi_{\alpha\beta} \\ \psi_{\beta\alpha} & \psi_{\beta\beta} \end{pmatrix}. \tag{3.15}$$

hvor $\psi\alpha\alpha = \text{Var}(\alpha_i)$, $\psi_{\beta\beta} = \text{Var}(\beta_i)$ og $\psi_{\alpha\beta} = Cov(\alpha, \beta)$.
Kombinerer vi ligningerne (3.8) og (3.12), får vi *kombinationsligning* for LCM således,

$$y_i = \Lambda(\mu_\eta + \Gamma x_i) + \Lambda\zeta_i + \epsilon_i \tag{3.16}$$

Bemærk her, at den sidste ligning er en generel form, der tillader et hvert T antal af gentagne målinger og K kovariater for latente kurveparametre.

Forholdet mellem LCM og LMM

Vi betragter ligningen (3.16). Den første del af denne ligning $\Lambda(\mu_\eta + \Gamma x_i)$ beskriver faste komponenter (faste effekter af latente variable η og faste effekter af observerede variable x) i LCM. Disse komponenter svarer til faste effekter $X\beta$ i lineær mixed model defineret i (2.1)

således $\Lambda(\mu_\eta + \Gamma x_i) = \tilde{X}\beta$, hvor $\Lambda x_i = \tilde{X}$ og $\beta = \mu_\eta + \Gamma$.
Derudover er den anden del af ligningen (3.16) $\Lambda\zeta_i$, beskriver stokastiske komponenter i LCM, og svarer til stokastiske effekter i lineær mixed model ZU således $\Lambda = Z$ og $\zeta_i = U_i$. Det vil sige, at vi kan skrive LCM som en lineær mixed model således

$$y = \tilde{X}\beta + ZU + \epsilon_i \tag{3.17}$$

hvor y er responsvariablen, $\tilde{X}$ og Z er kendte matricer, og ϵ_i er et fejlled. Bemærk, at den eneste forskel mellem LCM og LMM er, at i LCM antager vi eksplicit en latent variabel, mens i LMM antager vi en tilfældig effekt, som kan fortolkes som en latent variabel. Bemærk også, at selvom LMM og LCM er ækvivalente, er der stadig forskelle i de to modeller, og valget af model afhænger af formålet med analysen og egenskaberne for data [5, s. 1401].

Identifikation:

I dette afsnit vil vi undersøge, hvordan kan en LCM identificeres, herunder hvilke betingelser, der skal være opfyldt for at opnå entydig identifikation. Hvad der menes med identifikation er, at der er mindst lige så mange observerede variable, som der er parametre i modellen [2, s. 130]. For en LCM for en population med n individer og T gentagne målinger uden nogen kovariater, er de observerede variable middelværdien, variansen og kovariansen af y_t for hvert $t = 1, \ldots, T$. Det giver i alt $\frac{1}{2}T(T+3)$ kendte middelværdier, varianser og kovarianser i modellen. Betragter vi nu en LCM med K kovariater, er de observerede variable middelværdier, varianser og kovarianser af y_t og x_k ($\mathbb{E}[y_{it}]$, $\mathbb{E}[x_{ki}]$, $Cov(y_{it}, y_{i,t-s})$ for $s > 0$ og $t - s \geq 0$, $Cov(x_{ki}, x_{k-m,i})$ og $Cov(y_{it}, x_{ki})$). altså er der i alt $\frac{1}{2}(T + K)(T + K + 3)$ observerede variable [2, s. 130]. En LCM er identificeret hvis to følgend betingelder er opfyldt,

1. Parametrene i den ubetingede LCN (niveau 1 af LCM) har unike værdier i form af middelværdi, varians eller kovarians af de observerede variable, når der er mindst 3 tidsmålinger ($T = 3$). FX for $T = 2$ er der $\frac{1}{2}T(T + 3) = 5$ oberverede variable og $T + 5 = 7$ parametre, og dermed er modellen underidentificeret.
2. alle kovariater i modellen, er observerede (Der er ikke nogen latente kovariater) [2, s. 129].

Hvis antallet af de observerede variable er større end antallet af modelparametre, siges det at modellen er *overidentificeret*. Tilsvarende gælder det, at hvis antallet af parametre er mindre end antallet af de observerede variable i modellen, siges det at modellen er *underidentificeret*. Derudover for at kunne være i stand til at identificere alle modellers parametre, betragter vi flere antagelser. Vi antager blandt andet, at alle individer har den samme varians ved den samme tidsperiode, på trods af at variansen kan variere over tid, altså $Var(\epsilon_{it}) = Var(\epsilon_t)$.

Vi vil i følgende eksempel præsentere et specielt tilfælde af LCM. En lineær LCM med tre gentagne målinger $t = 1, \ldots, 3$ og to kovariater x_{1i} og x_{2i}.

Eksempel 3.1 Lineær LCM med tre tidsmålinger og to kovariater:

For at introducere en lineær LCM med tre gentagne målinger $T = 3$ og to kovariater x_i1 og x_2i, benytter vi Ligning (3.8).

$$y_{it} = \Lambda\eta_i + \epsilon_{it} \tag{3.18}$$

hvor

$$\eta_i = \mu_\eta + \Gamma x_i + \zeta_i \tag{3.19}$$

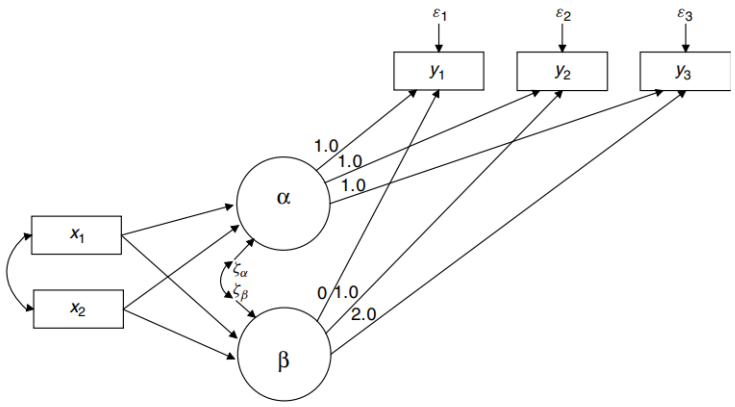
Altså

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{bmatrix} \tag{3.20}$$

hvor

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix} + \begin{bmatrix} \gamma_{\alpha_1} & \gamma_{\alpha_2} \\ \gamma_{\beta_1} & \gamma_{\beta_2} \end{bmatrix} \begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} + \begin{bmatrix} \zeta_{\alpha_i} \\ \zeta_{\beta_i} \end{bmatrix} \tag{3.21}$$

Denne model er illustreret med et path diagram på Figur (3.1).



Figur 3.1: En lineær LCM med tre gentagne målinger og to kovariater [2, s. 128].

Vi vil nu vise, at en lineær LCM med tre gentagne målinger og to kovariater er identificeret. Vi vil først liste de observerede variable og parametrene i modellen. De observerede variable er, $\mu_{y_{i1}}, \mu_{y_{i2}}, \mu_{y_{i3}}, \mu_{x_{1i}}, \mu_{x_{2i}}, Cov(y_{i1}, y_{i2}), Cov(y_{i1}, y_{i3}), Cov(y_{i2}, y_{i3}), Cov(x_{1i}, x_{2i})$, Det lader til 20 variable. Tilgengæld er der 12 parametre, som vi vil undersøge om de er identificerede. De er μ_α, μ_β, γ_{α_1}, γ_{α_2}, γ_{β_1}, γ_{β_2}, $\psi_{\alpha\alpha}$, $\psi_{\beta\beta}$, $\psi_{\alpha\beta}$ og $T = 3$ parametre af λ_t. Vi betragter derfor ligningen (3.20) og tager den forventede værdi på begge side af denne ligning således,

$$\mathbb{E}[y_{it}] = \begin{bmatrix} \mu_{y_{i1}} \\ \mu_{y_{i2}} \\ \mu_{y_{i3}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix} = \begin{bmatrix} \mu_\alpha \\ \mu_\alpha + \mu_\beta \\ \mu_\alpha + 2\mu_\beta \end{bmatrix} \tag{3.22}$$

Så middelværdierne af skæringer og hældninger er,

$$\mu_{\alpha_i} = \mu_{y_{i1}} \tag{3.23}$$

$$\mu_{\beta_i} = \mu_{y_{i2}} - \mu_{y_{i1}} \tag{3.24}$$

Og dermed er middelværiderne af skæringer og hældninger identificerede. Tilsvarende er varianser og kovarianser af observerede variable y_{it},

$$\Sigma = \Lambda \Psi \Lambda^T + \Theta_\epsilon \tag{3.25}$$

$$\Sigma = \begin{bmatrix} \text{VAR}\,(y_{i1}) & \text{COV}\,(y_{i1}, y_{i2}) & \text{COV}\,(y_{i1}, y_{i3}) \\ \text{COV}\,(y_{i2}, y_{i1}) & \text{VAR}\,(y_{i2}) & \text{COV}\,(y_{i2}, y_{i3}) \\ \text{COV}\,(y_{i3}, y_{i1}) & \text{COV}\,(y_{i3}, y_{i2}) & \text{VAR}\,(y_{i3}) \end{bmatrix} \tag{3.26}$$

$$= \begin{bmatrix} \psi_{\alpha\alpha} + \text{VAR}\,(\epsilon_{i1}) & \psi_{\alpha\alpha} + \psi_{\alpha\beta} & \psi_{\alpha\alpha} + 2\psi_{\alpha\beta} \\ \psi_{\alpha\alpha} + \psi_{\alpha\beta} & \psi_{\alpha\alpha} + \psi_{\beta\beta} + 2\psi_{\alpha\beta} & \psi_{\alpha\alpha} + 2\psi_{\beta\beta} + 3\psi_{\alpha\beta} \\ & + \text{VAR}\,(\epsilon_{i2}) & \psi_{\alpha\alpha} + 4\psi_{\beta\beta} + 4\psi_{\alpha\beta} \\ \psi_{\alpha\alpha} + 2\psi_{\alpha\beta} & \psi_{\alpha\alpha} + 2\psi_{\beta\beta} + 3\psi_{\alpha\beta} & + \text{VAR}\,(\epsilon_{i3}) \end{bmatrix} \tag{3.27}$$

og dermed er

$$\psi_{\alpha\beta} = Cov(y_{i1}, y_{i3}) - Cov(y_{i1}, y_{i2}) \tag{3.28}$$
$$\psi_{\alpha\alpha} = 2Cov(y_{i1}, y_{i3}) - Cov(y_{i1}, y_{i2}) \tag{3.29}$$
$$\psi\beta\beta = \big(Cov(y_{i2}, y_{i3}) - Cov(y_{i1}, y_{i4})\big)\,/2 \tag{3.30}$$

Yderligere er varianser for fejlene kan fåes ved brug af

$$\Theta_\epsilon = \Lambda \Psi \Lambda^T + \Sigma \tag{3.31}$$

givet at variablene $\psi_{\alpha\alpha}$, $\psi_{\alpha\beta}$ og $\psi_{\beta\beta}$ er identificerede. Og dermed er varianser af fejlene

$$\text{Var}(\epsilon_{i1}) = \text{Var}(y_{i1}) - \psi_{\alpha\alpha} \tag{3.32}$$
$$\text{Var}(\epsilon_{i2}) = \text{Var}(y_{i2} - \psi_{\alpha\alpha} - \psi_{\beta\beta} - 2\psi_{\alpha\beta} \tag{3.33}$$
$$\text{Var}(\epsilon_{i3}) = \text{Var}(y_{i3}) - \psi_{\alpha\alpha} - 4\psi_{\beta\beta} - 4\psi_{\alpha\beta} \tag{3.34}$$

Dette betyder, at en lineær LCM med tre gentagne målinger og to kovariater er identificeret.

◀

3.0.1 Estimation

Et hovedformål med latent variabel analyse er at benytte de gentagne målinger for observationer til at estimere uobserverede stier, der giver anledning til de gentagne målinger. Vi kan estimere parametrene i LCM ved brug af en lineær mixed model eller ved brug af en traditionel strukturel ligningsmodel (SEM) [2, s. 36]. Vi vil derfor i dette afsnit begynde med at definere *implicit middelværdi struktur* og *kovarians struktur* for LCM.

Implicit middelværdi struktur

For at finde ML-estimat for LCM i form af SEM, kan vi benytte implicit middelværdistruktur (implied mean structure). Implicit middelværdistruktur for en LCM kan bestemmes ved at tage den forventede værdi på begge side af ligning (3.16). Altså

$$\mu_y(\theta) = \Lambda(\mu_\eta + \Gamma \mu_x) \tag{3.35}$$

hvor $\mu_y(\theta)$ er middelværdivektoren i modellen for alle y, og θ er ukendte modelparametre, som skal estimeres. Ud fra ligningen (3.35) kan vi bestemme implicit middelværdistruktur for hvert y_{it}. Derudover er implicit middelværdistruktur for kovariater x, blot middelværdier af x_k, idet x_k er eksogene variable [2, s. 134]. Ved at kombinere middelværdien af y og x, får vi implicit middelværdistruktur for alle observerede variable.

$$\mu = \mu(\theta) \Leftrightarrow \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} = \begin{bmatrix} \Lambda(\mu_\eta + \Gamma\mu_x) \\ \mu_x \end{bmatrix} \tag{3.36}$$

Implicit kovariansstruktur

På samme måde som for implicit middelværdistruktur, kan vi udlede en kovarians strukturel ligning, som beskriver kovarianser og varianser af de observerede variable som en funktion af parametrene θ, altså $\Sigma = \Sigma(\theta)$, hvor Σ er populations kovariansmatrix af observationerne y og kovariater x. En vigtig antagelse for at bestemme implicit kovarians er, at fejlvarianser varierer over tid, altså $\mathrm{Var}(\epsilon_{it}) = \mathrm{Var}(\epsilon_t)$.
Trækker vi ligningen (3.35) fra ligningen (3.16), får vi forskellen mellem y_i og dens middelværdi μ_y, altså

$$\begin{aligned} y_i - \mu_y &= \big(\Lambda(\mu_\eta + \Gamma x_i) + \Lambda\zeta_i + \epsilon_i\big) - \big(\Lambda(\mu_\eta + \Gamma\mu_x)\big) && (3.37)\\ &= \Lambda\Gamma(x_i - \mu_x) + \Lambda\zeta_i + \epsilon_i && (3.38)\\ &= \Lambda\left(\Gamma(x_i - \mu_x) + \zeta_i\right) + \epsilon_i && (3.39) \end{aligned}$$

Da x_k er observerede variable, kan vi ikke simplificere forskellen $x_i - \mu_x$ mere.
Benytter vi definitionen af en kovariansmatrix (den forventede værdi af afvigelsesvariable ganget med afvigelsesvariablen transformeret), får vi at,

$$\begin{aligned} \Sigma(\theta) &= \mathbb{E}\left(\begin{bmatrix} y_i - \mu_y \\ X_i - \mu_x \end{bmatrix}\begin{bmatrix} y_i - \mu_y \\ X_i - \mu_x \end{bmatrix}^T\right) && (3.40)\\ &= \begin{bmatrix} \Lambda(\Gamma\Sigma_{xx}\Gamma^T + \Psi)\Lambda^T + \Sigma_{\epsilon\epsilon} & \Lambda\Gamma\Sigma_{xx} \\ \Sigma_{xx}\Gamma^T\Lambda^T & \Sigma_{xx} \end{bmatrix} && (3.41) \end{aligned}$$

hvor Σ_{xx} er kovariansmatrix af kovariater x. Bemærk, at elementet øverst til venstre af $\Sigma(\theta)$ viser en nedbrydning af kovariansmatricen for gentagne målinger af y_i i forhold til modellens strukturelle parametre. Elementerne øverst til højre og nederst til venstre af $\Sigma(\theta)$ giver kovariansmatricer af x_k og y_i, som funktioner af modellers parametre, og eeimentet nederst til højre giver kovariansmatricer for kovariater x_k.

ML-estimat

Da en LCM er ækvivalent til en lineær mixed model, kan vi benytte log-likelihoodfunktionen i (2.29) til at finde log-likelihoodfunktionen for hvert individ i populationen for en LCM som

$$\ell_i(\theta) = -\frac{1}{2}\log \mid \Sigma_i(\theta) \mid -\frac{1}{2}\left(y_i - \mu_i(\theta)\right)^T\left(\Sigma_i(\theta)\right)^{-1}\left(y_i - \mu_i(\theta)\right) \quad [2, s.136] \tag{3.42}$$

for $i = 1, \ldots, n$, z_i er en vektor af observerede variable for det i'te individ, og C_i er en konstant der ikke er uafhængig af θ [2, s. 136]. Log likelihoodsfunktionen for alle individer er summen

af log likelihoodsfunktioner for alle individer i populationen, Altså

$$\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta) \tag{3.43}$$

Vi finder et $\hat{\theta}$ som maksimerer $\ell(\theta)$. $\hat{\theta}$ opfylder almindelige egenskaber for ML-estimater (konsistent, asymptotisk normal, asymptotisk unbiased og asymptotisk efficient) [2, s. 136]. Endvidere er asymptotisk kovarians matrix for $\hat{\theta}$ lig med den inverse af Fisher information matricen.

3.0.2 Modelvalidering

Der er adskillige overordnede metoder (overall fit) til at undersøge, hvor godt modellen passer på data. Her præsenterer vi et par af disse metoder, som er nyttige i modelvurdering for LCM. En vigtig teststatistik for modelvurdering for LCM er χ^2-teststatistik defineret som,

$$T_{LM} = (n-1)\ell \tag{3.44}$$

for den tilpassede ML-funktion for LCM. Nulhypotesen, der skal undersøges er,

$$H_0 : \mu = \mu(\theta) \quad \text{og} \quad \Sigma = \Sigma(\theta) \tag{3.45}$$

Hvis nulhypotesen (3.45) og fordelingsantagelser, der begrunder ML-estimat holder, er den asymptotiske fordeling af teststatistikken central χ^2-fordeling med frihedsgrad, $df = \frac{1}{2}(T + K)(T + K + 3) - u$, hvor u er antallet af de estimerede parametre i modellen. I en fuldstændig identificeret model (antallet af ukendte parametre er lig med antallet af kendte parametre), vil teststatistikken ikke være brugbar, idet de tilpassede funktioner og teststatistikken er 0. Vi kan derfor kun bruge denne teststatistik, når modellen er overidentificeret med positive frihedsgrader, hvor signifikantheden tyder på, at modelspecifikation ikke nødvendigvis gengiver middelværdier eller kovariansmatricen for de observerede variabler [2, s. 44].
χ^2-teststatistik har nogle kritiske egenskaber, som får os til ikke at tro udelukkende på disse tests for overordnet modelvurdering. En af disse egenskaber er, at de observerede variabler kan stamme fra multivariate fordelinger, der udviser overskydende kurtosis, hvilket kan føre til, at teststatistikken er for høj eller for lav.

Der er også nogle andre mål som er brugt til overordnet modelvalidering. Det drejer sig om *baisline indekser*, som i modsætning til χ^2-teststatistik ikke er afhængige af stikprøvestørrelse n.

Baseline Fit indekser

Baseline indekser sammenligner to modeller. En baselinemodel (nul-model) og hypotesemodel (tilpasset model).
Vi vil nu introducere nogle baseline indekser.

Tucker Lewis

Den mest kendt baseline indeks brugt i overordnet modelvalidering for LCM er den såkaldt *TuckerLewis indeks (TLI)*. Lader vi T_b repræsentere teststatistik for basislinemodellen, T_h

for hypotesemodellen, df_b for frihedsgrader for basismodellen og df_h for frihedsgrader af hypotesemodellen, er Tucker-Lewis-indekset (TLI) givet ved,

$$\text{TLI} = \frac{T_b/\text{df}_b - T_h/\text{df}_h}{T_b/\text{df}_b - 1} \tag{3.46}$$

TLI varierer generelt mellem 0 og 1. En værdi på 1 er en idealtilpasning. Værdierne mindre end 1 er kritiske for models tilstrækkelighed og værdierne der er større end 1 tyder på muligheden for model overtilpasning eller at have en model med for mange parametre, hvor nogle af parametrene udnytter tilfældige variationer i data [2, s. 46].

Incremental fit indeks

En anden baseline fit indeks er *Incremental fit indeks (IFI)*, som er givet ved,

$$\text{IFI} = \frac{T_b - T_h}{T_b - \text{df}_h} \tag{3.47}$$

IFI tager værdier mellem 0 og 1, hvor 1 er en idealtilpasning. Værdierne der er meget større end 1, indikerer overfitting og værdierne under 0.9 er uacceptable [2, s. 46].

Root-mean-sqaure error of approximation (RMSEA)

En statisik måling, som er anvendt til at vurdere overordnet tilpasning af model på data, er *Root mean-square error of approximation (RMSEA)*. RMSEA er bygget på en ikke centarl χ^2 fordeling, som er en asymptotisk fordeling for teststatistikken, når H_0 er ugyldig, og graden af fejlspecifikation ikke er for signifikant [2, s. 47]. RMSEA er givet ved,

$$\text{RMSEA} = \sqrt{\frac{T_h - \text{df}_h}{(N-1)\text{df}_h}} \tag{3.48}$$

Bemærk, at tælleren for RMSEA under kvadratrodstegnet $T_h - df_h$ er et asymptotisk unbiased estimat for en ikke central parameter for den ikke-centrale χ^2 fordeling er underliggende T_h. Derudover er divisionen med $n-1$ en justering for at tage højde for prøvestørrelseseffekten på den ikke central parameter, og df_h giver en straf for at opbruge modelfrihedsgrader. RMSEA har en nedre grænse på 0 men ikke en øvre grænse. En lille RMSEA-værdi tyder på, at modellen passer godt på data (jo tættere på nul den er, jo bedre passer modellen til data). En fordel ved RMSEA er værdierne af lave og høje grænser for konfidensintervaller. Der foreslås retningslinjer, sådan at værdier på mindre end 0,05 indikerer en vel tilpasset model. Værdierne er større end 0,10 indikerer en dårlig pasform, og værdierne derimellem er moderate [2, s. 47].

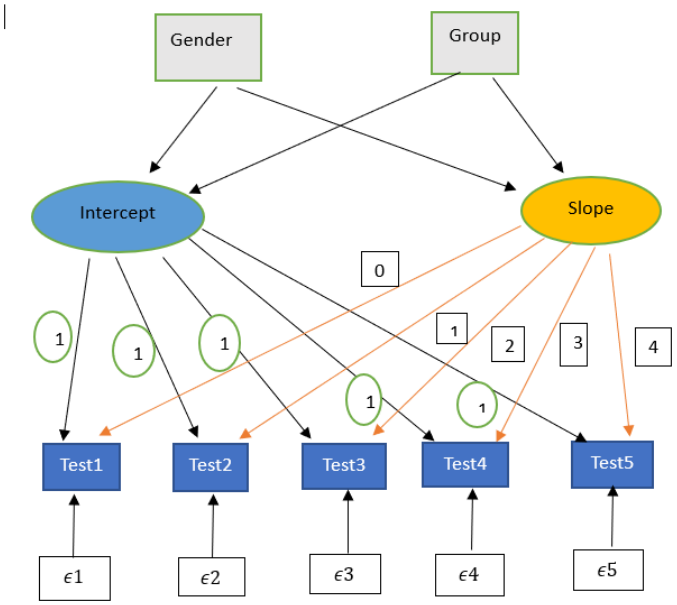
3.1 Dataanalyse

I dette afsnit vil vi afprøve de beskrevne metoder (LCM) på et praktisk datasæt. Vi vil benytte de samme to datasæt, som vi brugte i afsnit (2.4). Vi vil også benytte den samme analysestrategi; det vil sige vi anvender analysesresultaterne for $dataset_2$ til at bekræfte de observerede mønstre i $dataset_1$. For at kunne benytte data til at konstruere en LCM, kræver det at data er i såkaldt *wide format*. Vi laver derfor $dataset_1$ og $dataset_2$ om til $widedata_1$ og $widedata_2$. Variablerne for widedata er illustreret i følgende tabel.

Variabel	Forklaring
id	elevs id
$test1$	testresultater ved baseline
$test2$	testresultater ved den 2. tidsmåling
$test3$	testresultater ved den 3. tidsmåling
$test4$	testresultater ved den 4. tidsmåling
$test5$	testresultater ved den 5. tidsmåling
gender	køn (1 dreng, 0 pige)
group	(1: højt-præsterende elever, 0: lavt-præsterende elever)

Tabel 3.1: Variablene i widedata.

Vi laver indledningsvis et pathdiagram, der illustrerer en lineær LCM med 5 gentagne målinger ($test_1, test_2, test_3, test_4$ og $test_5$) og to kovariater `gender` og `group` på vores data. Diagrammet kan ses i følgende figur,



Figur 3.2: Et path diagram for LCM i (3.52).

Vi tilpasser nu en lineær LCM, defineret i (3.1), (3.2) og (3.3), på tværs af 5 gentagne målinger for testscore for elever i widedata. Til dette formål modellerer vi variablen `test` med en skæring og en lineær hældning. Skæringen repræsenterer det forventede testresultat for elever ved baseline ($\texttt{time}_1$) og har loadings på 1 for alle $t = 1, \ldots, 5$, mens hældningsfaktoren har en loadings på $0, \ldots, 4$ for $t = 1, \ldots, 5$. Altså,

$$\texttt{test}_{it} = \texttt{intercep}_i + \texttt{time}_t \cdot \texttt{slope}_i + \epsilon_{it} \tag{3.49}$$

hvor

$$\texttt{intercept}_i = \mu_{\texttt{intercept}} + \gamma_{01} \cdot \texttt{gender} + \gamma_{02} \cdot \texttt{group} + \zeta_{\texttt{intercept}_i} \tag{3.50}$$

og

$$\texttt{slope}_i = \mu_{\texttt{slope}} + \gamma_{11} \cdot \texttt{gender} + \gamma_{12} \cdot \texttt{group} + \zeta_{\texttt{slope}_i} \tag{3.51}$$

hvor $\texttt{time}_t = \lambda_t$ og tager værdier $0, 1, 2, 3, 4$ og $t = 1, \ldots, 5$, idet det er en lineær LCM. Vi indsætter nu Ligninger (3.50) og (3.51) i Ligning (3.49), så vi får en kombinationsligning for LCM

$$\texttt{test}_{it} = (\mu_{\texttt{intercept}} + \texttt{time}_t \cdot \mu_{\texttt{slope}}) + (\gamma_{01} + \texttt{time}_t \cdot \gamma_{11}) \cdot \texttt{gender} \tag{3.52}$$

$$+ (\gamma_{02} + + \gamma_{12} \cdot \texttt{time}_t) \cdot \texttt{group} + (\zeta_{\texttt{intercept}_i} + \texttt{time}_t \cdot \zeta_{\texttt{slope}_i} + \epsilon_{it}) \tag{3.53}$$

Vi benytter R til at beregne parametrene i modellen. Til dette formål benytter vi en R-pakke kaldt *lavaan*, som er brugt til at modellere latente variable. Navnet *Lavaan* kommer af *latent* (la) *variable* (va) *analysis* (an). Syntaksen for *lavaan* ligner R-regression model stil. Den tillader arbitrære navner for skæringen og hældningen i modellen. Endvidere angiver relationssymbolet $=\sim$, at den venstre side er latente variable og den højre side er observerede variable. Vi bruger desuden funktionen `growth` til at tilpasse LCM. Dette er illuatreret i følgende kod.

```
install.packages("lavaan")
library(lavaan)
#latente variable
test1 <- "
  intercept=~  1*test_1  +  1*test_2 + 1*test_3  +  1*test_4 + 1*test_5
  slope =~  0*test_1  +  1*test_2 + 2*test_3  +  3*test_4 + 4*test_5
#regressions
intercept ~ gender + group
slope ~ gender + group
"
testmodel1 = growth (test1, data= widedata1)
summary(testmodel1)
```

Udvalgte R-kod for denne analyse kan ses i Apendix (A).

3.1.1 LCM-analysen

Vi vil i det følgende gennemgå resultaterne fra R for LCM-analysen for $widedata_1$ og $widedata_2$.

Middelværdier af skærings- og hældningsfaktorer for $widedata_1$

```
    Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .test_1            0.000
   .test_2            0.000
   .test_3            0.000
   .test_4            0.000
   .test_5            0.000
   .intercept         8.835    0.904    9.773    0.000
   .slope             1.317    0.265    4.966    0.000
```

Fra ovenstående resultat kan vi aflæse en statistisk signifikant middelværdi af skæringen på $\mu_{intercept} = 8.84$ med p-værdi $p = 0$. Dette svarer til at gennemsnittet af testresultaterne for eleverne ved den første tidsmåling (baseline) er 8.84.
Middelværdien for hældningen er estimeret til at være $\mu_{slope} = 1.32$ og er også statistisk

signifikant med p-værdi $p = 0$. Altså er der en positiv lineær stigning i elevernes testresultater fra en tidsmåling til den efterfølgende tidsmåling på 1.32.

Middelværdier af skærings- og hældningsfaktorer for $widedata_2$

```
   Intercepts:
                 Estimate  Std.Err  z-value  P(>|z|)
  .test_1           0.000
  .test_2           0.000
  .test_3           0.000
  .test_4           0.000
  .test_5           0.000
  .intercept        5.829    1.960    2.974    0.003
  .slope            0.817    0.520    1.570    0.116
```

Baseret på det ovenstående resultat kan vi konkludere, at der er en statistisk signifikant gennemsnitsværdi for skæringsfaktoren på 5.83 point i $widedata_2$, med en p-værdi på 0.00. Dette betyder, at gennemsnittet af elevernes testresultater ved den første tidsmåling (baseline) er 5.83. Middelværdien for hældningsfaktoren er estimeret til at være 1.82, men statistisk usignifikant, med en p-værdi $p = 0.116$. Dette indikerer, at der ikke over tid er sket en signifikant udvikling i elevernes testresultater.

Regressionsparametre for $widedata_1$

Vi har to kovariater i vores LCM, `gender` og `group`, hvor estimaterne for deres effekter kan ses i følgende resultat

```
   Regressions:
     Estimate  Std.Err  z-value  P(>|z|)
intercept ~
  gender            3.671    0.744    4.934    0.000
  group            -0.495    0.215   -2.300    0.021
slope ~
  gender            0.195    0.218    0.893    0.372
  group             0.166    0.063    2.638    0.008
```

Dette viser i hvor høj grad testresultaterne for drengene adskiller sig fra testresultaterne for pigerne, og om der også er forskel på testresultatet for lavt- og højt-præsterende elever. Vi kan aflæse et positiv estimat af skæringsfator for drenge (`gender` = 1) på 3.67. Denne effekt er statistisk signifikant med p-værdi $p = 0.0$. Dette indikerer, at drenge i gennemsnit har højere scoring i testresultatet end pigerne har i den første tidsmåling.
Vi kan også aflæse en statistisk signifikant negativ skæringsværdi for gruppen lavt-præsterende elever på -0.495 med p-værdi $p = 0.02$. Dette indikerer, at testresultater for lavt-præsterende elever i gennemsnit er lavere end det er for højt-præsterende i den første tidsmåling.
Da p-værdien af regressionskoefficienten for drenge af den lineære hældningsfaktor er 0.372, er testresultattet for drengene ikke en signifikant forudsigelse til at forklare den lineære vækst-trend af testresultater. Imidlertid er p-værdien af estimatet for lavt-præsterende elever til den lineære hældningsfaktor positiv og signifikant med p-værdien 0.008. Dette indebærer, at lavt-præsterende elever udviser en stejlere lineær vækststendens end højt-præsterende elever.

Regressionsparametre for $widedata_2$

Ligesom for $widedata_1$ har vi to kovariater i vores LCM, `gender` og `time`, hvor estimater af deres effekter kan ses i følgende

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  intercept ~
    gender            1.035    0.727    1.423    0.155
    group             0.371    0.446    0.832    0.405
  slope ~
    gender            0.094    0.193    0.487    0.626
    group             0.176    0.118    1.487    0.137
```

Vi kan aflæse en skæring for drenge (`gender` = 1) på 1.035, som er statistisk usignifikant med p-værdi $p = 0.155$. Endvidere kan der aflæses en lille statistisk usignifikant hældning for drenge på 0.094 med p-værdi $p = 0.626$. Dette indikerer, at elevers køn ikke har nogen effekt på testresultater i den første tidsmåling i $widedata_2$.
Vi kan også aflæse en statistisk usignifikant skæring for lavt-præsterende elevers gruppe på 0.371 med p-værdi $p = 0.405$. Dette indikerer, at lavt- og højt-præsterende elever ikke har forskellige testresultater i baseline tidsmåling. Endvidere kan der aflæses en statistisk usignifikant hældning for lavt-præsterende elevers gruppe på 0.176 med p-værdi $p = 0.137$. Dette indikerer, at både lavt- og højt-præsterende elever har den samme lineære hældning med tiden.

Varians og kovarians parametre for $widedata_1$

```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .test_1            4.874    1.183    4.120    0.000
   .test_2            7.568    1.210    6.254    0.000
   .test_3            8.113    1.329    6.105    0.000
   .test_4           11.453    1.883    6.081    0.000
   .test_5           11.296    2.328    4.853    0.000
   .intercept        12.561    2.188    5.742    0.000
   .slope             0.595    0.212    2.809    0.005
```

Variansen af skæring er estimeret til at være $\text{Var}(\zeta_{\alpha_i}) = 12.561$ og er statistisk signifikant med p-værdi $p = 0$. Denne varians repræsenterer de individuelle forskelle i testresultater for elever ved baseline tidsmåling. Dette indikerer, at der er en statistisk signifikant forskel mellem elever i deres indledende testresultater. Endevidere er estimater af varianserne af fejlleden i modellen i (3.49), $\text{Var}(\epsilon_{i1}) = 4.874$, $\text{Var}(\epsilon_{i2}) = 7.568$, $\text{Var}(\epsilon_{i3}) = 8.113$, $\text{Var}(\epsilon_{i4}) = 11.453$ og $\text{Var}(\epsilon_{i5}) = 11.296$, som er alle sammen statistisk signifikante med p-værdi $p = 0$.
Variansen af hældningen er estimeret til at være $\text{Var}(\zeta_{\beta_i}) = 0.595$ og er også statistisk signifikant med p-værdi $p = 0.005$. Denne varians repræsenterer de individuelle forskelle i den lineære vækstendens. Det vil sige, at variansen af hældningen viser, hvordan individuelle elever adskiller sig i deres forløb af deres skoleår på tværs af de fem testmålinger. Da variansen er statistisk signifikant indikerer det, at der eksisterer individuelle forskelle i den lineære vækst af testresultater.

Et andet vigtig estimat er kovariansen mellem skæringen og hældningen for $widedata_1$.Estimatet for denne kovarians kan aflæses fra det nedstående resultat.

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
.intercept ~~
```

```
   .slope             1.426    0.492    2.896    0.004
```

Bemærk, at vi har en positivt kovarians på $\psi_{\alpha\beta} = 1.426$. Dette indikerer, at højtt-præsterende elever, er mere tilbøjelige til at have en stejlere lineær væksttendens, end lavt-præsterende elever. Endvidere er estimatet for kovariansen mellem skæringen og hældningen statistisk signifikant med p-værdi $p = 0.004$, hvilket indikerer, at denne kovarians er signifikant forskellig fra 0.

Varians og kovarians parametre for $widedata_2$

Vi vil nu se på varianser af skæringen og hældningen og kovariansen mellem skæringen og hældningen for $widedata_2$.

```
  Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .test_1            6.576    1.193    5.512    0.000
   .test_2            5.712    0.889    6.425    0.000
   .test_3            6.596    0.985    6.694    0.000
   .test_4            7.086    1.129    6.276    0.000
   .test_5            5.897    1.354    4.355    0.000
   .intercept        13.622    2.161    6.302    0.000
   .slope             0.601    0.169    3.553    0.000
```

Fra ovenstående resultat kan vi aflæse en statistisk signifikant varians af skæringen $\text{Var}(\zeta_{\alpha_i}) = 13.622$ med p-værdi $p = 0$. Dette indikerer, at der er individuelle forskelle i testresultater for elever ved baseline tidsmåling. Endvidere er estimater af varianserne af fejlleden i modellen i (3.49), $\text{Var}(\epsilon_{i1}) = 6.576$, $\text{Var}(\epsilon_{i2}) = 5.712$, $\text{Var}(\epsilon_{i3}) = 6.596$, $\text{Var}(\epsilon_{i4}) = 7.086$ og $\text{Var}(\epsilon_{i5}) = 5.897$, som er alle sammen statistisk signifikante med p-værdi $p = 0$.
Variansen af hældningsfaktoren er estimeret til at være $\text{Var}(\zeta_{\beta_i}) = 0.601$ og statistisk signifikant med p-værdi $p = 0$. Dette indikerer altså, at individuelle testresulater for elever adskiller sig i løbet af deres skoleår på tværs af de fem testmålinger, og dermed eksisterer der over tid individuelle forskelle i den lineære vækst af testresultaterne.

```
  Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
.intercept ~~
   .slope             1.426    0.428    3.334    0.001
```

Vi kan fra det ovenstående resultat også aflæse en positiv kovarians mellem skæringen og hældningen på $\text{Var}(\psi_{\alpha\beta}) = 1.426$, hvilket indikerer, at lavt-præsterende elever er mere tilbøjelige til at have en stejlere lineær væksttendens end højt-præsterende elever. Derudover er estimatet for denne kovarians statistisk signifikant med p-værdi $p = 0.001$, hvilket indikerer, at denne kovarians er signifikant forskellig fra 0.

Modelvalidering for både $widedata_1$ **og** $widedata1$

For at undersøge validiteten af LCM, som vi har tilpasset på vores data, laver vi en χ^2 teststatistik og bergner TLI, ICI og RMSEA baseret på $widedata_1$ og $widedata_2$.
I følgende resultat kan ses χ^2 teststatistik for $widedata_1$.

```
    Model Test User Model:

  Test statistic                                68.667
```

```
  Degrees of freedom                                16
  P-value (Chi-square)                           0.000

   Model Test Baseline Model:

  Test statistic                               533.267
  Degrees of freedom                            20
  P-value                                        0.000
```

Da $\chi^2(16) = 68.667$ med p-værdi $p = 0.00$, indikerer det, at den tilpassede lineære LCM ikke passer helt til data. Vi undersøger derfor værdierne af CFI, TLI og RMSEA for $widedata_1$, som er givet i følgende resultat

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.897
  Tucker-Lewis Index (TLI)                       0.872

Root Mean Square Error of Approximation:

  RMSEA                                          0.168
  90 Percent confidence interval - lower         0.128
  90 Percent confidence interval - upper         0.210
  P-value H_0: RMSEA <= 0.050                    0.000
  P-value H_0: RMSEA >= 0.080                    1.000
```

Værdierne for både CFI og TLI er i henholdsvis 0.897 og 0.872, som begge er næsten lig med 0.9, hvilket indikerer en god model tilpasning. Yderligere er RMSEA 0.168 som er større end 0.01, hvilket indikerer en dårlig model tilpasning.

I følgende resultat kan der ses χ^2 teststatistik for $widedata_2$.

```
 Model Test User Model:

  Test statistic                                49.992
  Degrees of freedom                            16
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                               627.284
  Degrees of freedom                            20
  P-value                                        0.000
```

Ligesom for $widedata_1$ er $\chi^2(16) = 49.992$ teststatistik signifikant med p-værdi $p = 0.00$, hvilket indikerer en dårlig modeltilpasning.
Vi undersøger derfor værdierne af CFI, TLI og RMSEA for $widedata_2$.

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.944
  Tucker-Lewis Index (TLI)                       0.930

Root Mean Square Error of Approximation:
```

```
RMSEA                                           0.125
90 Percent confidence interval - lower          0.087
90 Percent confidence interval - upper          0.166
P-value H_0: RMSEA <= 0.050                     0.001
P-value H_0: RMSEA >= 0.080                     0.973
```

Fra ovenstående resultat kan vi aflæse CFI og TLI på henholdsvis 0.944 og 0.930 som begge er større end 0.9, hvilket indikerer en god modeltilpasning. Yderligere er RMSEA på 0.125, som er større end 0.1, hvilket indikerer en dårlig modeltilpasning.
Bemærk, at både χ^2 teststatistik og RMSEA indikerer en dårlig model tilpasning, mens CFI og TLI begge indikerer en god model tilpasning for $widedata_1$ og $widedata_2$. Dette betyder, at denne lineær LCM ikke så godt passe til vores data. Dette kan måske skyldes et dårligt valg af λ_t (lineæriteten), hvor vi har betragtet en lineær udvikling af testresultater over tid uden at tage hensyn til at der kan være forskellige udviklinger i de forskellige undervisnings perioder.

Konklussion

Udfra LCM analysen kan vi konkludere, at der er en lineær stigning af testresultater for både lavt- og højt-præsterende elever over tid. Altså varierer udviklingen af elevers færdigheder i brøkregning over tid. Endvidere er der sket en større udvikling af de højt-præsterende elevers færdigheder i brøkregning sammenlignet med, hvad der ses for de lavt-præsterende elever.
Yderligere er der en kønseffekt på 3.67 i $widedata_1$. Altså drengene har i gennemsnit 3.67 point flere end pigerne. Der er også blevet observeret en kønseffekt i $widedata_2$ på 1.06.

Smmenligning mellem LMM-analysen og LCM-analysen

I både LMM- og LCM-analysen har vi konkluderet, at elevernes færdigheder i brøkregning for lavt- og højt-præsterende elever varierer over tid.
I LMM-analysen har vi konkluderet, at der er sket en forbedring i elevernes færdigheder i brøkregning i alle undervisningsperioder undtagen den sidste periode for den højt-præsterende gruppe. Sammenlignet er der kun sket en forbedring i elevers færdigheder i brøkregning for den lavt-præsterende gruppe i en enkelt periode.
Dette stemmer overens med, hvad vi har konkluderet ud fra LCM-analysen. Altså at der er sket en større udvikling i elevers færdigheder i brøkregning for den højt-præsterende gruppe sammenlignet med den lavt-præsterende gruppe.
I begge analyser har vi observeret en kønseffekt på 3.97 og 3.67 for henholdsvis for LMM og LCM i $data_1$ og $widedata_2$. For både $data_2$ og $widedata_2$ har vi observeret en statistisk usignifikant kønseffekt for både LCM og LMM.
Bemærk, at LCM og LMM har forskellige middelværdistrukturer og forskellige kovarians strukturer, og derfor er de lidt svært at sammenligne. I LCM har hver gruppe den samme hældning over tid (i alle perioder), mens i LMM er der forskellige hældninger i de forskellige perioder.
LMM tillader en varians i hver tidsmåling, mens i LMM er der kun en fælles varians i alle tidsmålinger. Vi har observeret forskellige varianser imellem eleverne i LCM og LMM. Vi har altså observeret en varians imellem eleverne i LCM for $widedata_1$ på 12.561 og en varians imellem eleverne i LMM for $data_1$ på 18.399. Tilsvarende har vi observeret en varians imellem eleverne i LCM for $widedata_2$ på 13.622, mens variansen imellem eleverne i LMM for $data_2$ var 15.594.

3.1.2 Fortolkning og diskussion

Gennem de beskrevne analyser, LCM-analysen og LMM-analysen, har vi blandt andet undersøgt forskellen mellem lavt- og højt-præsterende elevers udvikling af deres færdigheder i brøkregning i løbet et skoleår. Ligeledes har vi kigget på, hvordan disse forskelle er relateret til undervisning i brøkregning og undervisning i andre matematiske emner, såsom geometri og algebra. Vi har også undersøgt om der er nogle kovariater som fx køn, der har påvirket disse udviklinger og forskelle.
I LMM-analysen har vi tilpasset en lineær mixed model, hvor vi har betragtet variablene `time`, `group` og `gender` som faste effekter, og variablene `id` og `class-id` som stokastiske effekter.
Vi har også undersøgt validiteten af LMM ved at lave en grafisk analyse, hvor vi undersøgt om fordelingsantagelserne for residualerne og stokastiske effekter er opfyldt. Til dette formål har vi først lavet et QQ-normal plot og et histogram af residualerne, hvilket indikerede, at residualerne følger en normalfordeling med middelværdi 0. Vi har derefter plottet residualerne mod de fittede værdier for at undersøge, om der er outliers i residualerne, hvilket indikerede, at der er nogen outliers.

I LCM-analysen har vi tilpasset en lineær LCM på data, hvor vi har betragtet en stokastisk skæring og en lineær hældning. Vi har også undersøgt effekten af to kovariater (`gender` og `group`) på modellen. Vi har yderligere undersøgt validiteten af den tilpassede LCM ved at benytte χ^2 test statistik og beregne CFI, TLI og RMSEA.
Vi kunne ligeledes have tilpasset en kvadratisk LCM, men dette udeladt.

Sluttelig har vi sammenlignet resultaterne fra LMM-analysen og LCM-analysen. Vi har blandet andet konkluderet, at elevers færdigheder i brøkregning for både lavt- og højt-præsterende elever varierer over tid, og at de højt-præsterende elever har opnået større forbedring i deres færdigheder i brøkregning end de lavt-præsterende elever. Der er også blevet observeret en kønseffekt, hvor drengene har opnået højere udvikling i deres færdigheder i brøkregning sammenlignet med pigerne.

4 | Bibliografi

[1] S. C. Andersen. Multilevel-modeller: en introduktion og et eksempel. *polition*, 23(39):294–316, 2007. URL: `https://pure.au.dk/ws/files/42052535/Andersen_Multilevel_modeller.pdf`.

[2] K. A. Bollen and P. J. Curran. *Latent Curve Models A Structural Equation Perspective*. John Wiley Sons, Inc., Hoboken, New Jersey, 1. edition, 2006.

[3] C. Li. Littles test of missing completely at random. *The Stata Journal*, 15(4):795809, 2013. URL: `https://journals.sagepub.com/doi/pdf/10.1177/1536867X1301300407`.

[4] H. Madsen and P. Thyregod. *Introduction to general and generalized linear models*. Texts in statistical science. CRC Press, 2010.

[5] D. McNeish and T. Matta. Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Psychonomic Society*, 17(4):13981414, 2017. URL: `https://link.springer.com/content/pdf/10.3758/s13428-017-0976-5.pdf`.

[6] B. T.West, K. B.Welch, and A. T.Galecki. *LINEAR MIXED MODELS*. A Practical Guide Using Statistical Software. Chapman Hall/CRC, 2007.

[7] R. Waagepetersen. Predictio [lecture notes], 2022. URL: `https://people.math.aau.dk/~rw/Undervisning/MM_BI/Handouts/lektion7.pdf` [cited 2022-4-21].

[8] R. Waagepetersen. Maximum likelihood estimation for linear mixed models [lecture notes], 2023. URL: `https://people.math.aau.dk/~rw/Undervisning/MM_BI/Handouts/lektion2.pdf`.

A | Appendiks

R-koden for at tilpasse en lineær mixed model for både $data_1$ og $data_2$ er

```
# en lineær mixed model for data1
lme1 <- lmer(test ~ factor(time) *factor(group)+factor(gender)
+(1|id) + (1|clas_id), data=data1)
summary(lme1)
#en lineær mixed model for data2
lme2 <- lmer(test ~ factor(time) *factor(group)+factor(gender)+
(1|id) + (1|clas_id), data=data2)
summary(lme2)
```

R-koden til at plotte modellerne $lme1$

```
moddata1 <- unique(select(dataa[dataa$intervention == 1,], id, gender, group))
meangirls1 <- nrow(moddata1[moddata1$gender == 1,])/nrow(moddata1)
fixef(lme1)
beta1mod <- fixef(lme1)[1]+meangirls1*fixef(lme1)[8]+
  c(0, fixef(lme1)[2:5])
beta2mod <- beta1mod+fixef(lme1)[6]+meangirls1*fixef(lme1)[8]+
  c(0, fixef(lme1)[9:12])
plot(NA,NA, xlim= c(1,5),ylim = c(4,22), ylab ="test score",xlab ="time",
main = "Data1")
points(beta1mod,type = "b", lty = "dashed")
points(beta2mod,type = "b", lty = "dashed")
```

R-koden for at plotte modellen $lme2$

```
moddata2 <- unique(select(dataa[dataa$intervention == 2,], id, gender, group))
meangirls1 <- nrow(moddata2[moddata2$gender == 1,])/nrow(moddata2)
fixef(lme2)
beta3mod <- fixef(lme2)[1]+meangirls1*fixef(lme2)[8]+
  c(0, fixef(lme2)[2:5])
beta4mod <- beta3mod+fixef(lme2)[6]+meangirls1*fixef(lme2)[8]+
  c(0, fixef(lme2)[9:12])
plot(NA,NA, xlim= c(1,5),ylim = c(4,26), ylab ="test score",xlab ="time",
main = "Data2")
points(beta3mod,type = "b", lty = "dashed")
points(beta4mod,type = "b", lty = "dashed")
```

R-koden, som er brugt til at plotte QQ-normal af residualerne for både $data_1$ og $data_2$, er

```
#residualer for data1
res1=residuals(lme1)
qqnorm(res1)
qqline(res1)
```

```
#residualer for data2
res2=residuals(lme2)
qqnorm(res2)
qqline(res2)
```

R-koden som er brugt til at lave long-format data om til wide-format data

```
install.packages("gule")
install.packages("dplyr")
library("dplyr")
library("tidyr")
library("glue")
nwidedata = dataa %>%
  select (id, test, time, gender, intervention, group) %>%
  pivot_wider( names_from = time, values_from= test) %>%
 rename_at(vars('0','1','2', '3', '4'),function(x) glue::glue('test_{x}'))
#mutate(NA= as.numeric(group)-0)
print(widedata)
View (widedata)
widedata1 <- subset(nwidedata, intervention != "2")
View(widedata1)

widedata2 <- subset(nwidedata, intervention != "1")
View(widedata2)
```

R-koden for at tilpasse en lineær LCM for widedata1

```
test1 = "
  intercept=~  1*test_1  +  1*test_2 + 1*test_3  +  1*test_4 + 1*test_5
  slope =~  0*test_1  +  1*test_2 + 2*test_3  +  3*test_4 + 4*test_5

#regressions
intercept ~ gender + group
slope ~ gender + group

"
library(lavaan)
testmode1 = growth (test1, data= widedata1)
summary(testmode1, fit.measures=TRUE)
```

R-koden for at tilpasse en lineær LCM for widedata2

```
test2 = "
  intercept=~  1*test_1 +  1*test_2 + 1*test_3  +  1*test_4 + 1*test_5
  slope =~  0*test_1  +  1*test_2 + 2*test_3  +  3*test_4 + 4*test_5
#regressions
intercept ~ gender + group
slope ~ gender + group

"
library ("lavaan")
testmode2 = growth (test2, data= widedata2)
summary(testmode2, fit.measures=TRUE)
```