Blanda modellar i R

Jorunn Slagstad

Masteroppgåve i statistikk
Finansteori og forsikringsmatematikk

Matematisk institutt
Universitetet i Bergen

21. november 2006

Gratitude is merely the secret hope of further favors.

Francois de La Rochefoucauld (1613–1680)

Takk

Eg ynskjer først og fremst å takke rettleiaren min, Hans Julius Skaug, for god hjelp.

Eg vil så takke «svigerforeldra» mine, Connie og Kåre Espeland, for å ha teke meg og sambuaren min i hus då vi oppdaga at det å kjøpe leiligheit i Bergen ikkje let seg gjere i løpet av nokre korte seinsommarsveker i august.

Eg vil òg takke sambuaren min, Håkon Espeland, for å ha stått på med å pusse opp vår nykjøpte leiligheit, medan eg har jobba med masteroppgåva mi på Kroepeliens hus.

Vidare vil eg takke medstudentar for hyggelege lunsjpausar, gode diskusjonar og for å ha hjelpt meg med å telje ned dagane til innleveringsdato. (Er det ikkje i dag du skal levere, Jorunn?) Særleg vil eg takke Karl Ove Hufthammer for uvurderleg hjelp med LaTeX, og for tolmod med korrekturlesing.

Bergen, 21. november 2006
Jorunn Slagstad

If you don't find it in the index, look very carefully through the entire catalogue.

Sears, Roebuck, and Co. (Consumer's Guide 1897)

Innhald

In the begining there was nothing, and it exploded.

Terry Pratchett [om «big bang»-teorien]

Innleiing

Masteroppgåva mi handlar om miksa, eller som eg vil kalle dei, *blanda* modellar. Slike modellar er formulerte for å kunne forklare korrelasjonen av repeterte målingar på same objekt over tid. Eg vil sjå på både lineære blanda og generasliserte lineære blanda modellar.

For analysar av lineære blanda modellar har eg brukt programpakken R, for det meste versjon 2.4.0 (R Development Core Team 2006), og pakken `nlme` utvikla av Pinheiro og Bates (2000). Eg har teke utgangspunkt i datasettet `Orthodont` som ligg lagra i R i denne pakken. Datasettet inneheld 108 observasjonar av ein ortopedisk avstand måla av tannteknikarar på born i alderen åtte til tolv år. Avstanden er måla mellom to punkt i kjeven til borna, kalla «the pituitary gland» (hypofysen) og «the pterygomaxillary fissure», som er lett synleg på røngtenbileter. For kvart born er det gjort fire målingar, alle teke med to års mellomrom frå og med bornet er åtte år til og med det er fjorten år. Dette datasettet er blitt studert av Potthoff og Roy (1964), i samband med deira vekstkurvemodell, kjent på engelsk som *the growth curve model*. I si tilpassing av ein vekstkurvemodell til `Orthodont`, var Potthoff og Roy (1964) mellom anna interesserte i å finne ut om vekstkurvene skulle representerast ved eit andregradspolynom, eller om ein lineær modell kunne nyttast. Dei ynskte òg å undersøkje om jenter og gutar trong kvar sin modell for å forklare auka i avstanden, eller om vekstkurva kunne modellerast uavhengig av kjønn. Datasettet har seinare blitt trukke fram mellom anna av Pinheiro og Bates (2000) som døme på observasjonar der ein lineær blanda modell høver, og mine analysar bygger på antakinga om at vekstkurvene til borna er lineære.

I analysar av generaliserte lineære blanda modellar har eg sett på eit datasett av talet på krav i poliser i ulike yrkesgrupper over ein sjuårs-periode. Datasettet, som kan studerast i si heilheit i Klugman (1992, side 197), inneheld 931 observasjonar av antall krav i 133 ulike risikogrupper. Antonio og Beirlant (2006) har nytta datasettet til å studere ein GLMM med Poisson fordeling for talet på krav. Ein blanda modell for skadeforsikringsdata, som frekvens av krav, er svært aktuell sidan den let ein utnytte at det finnes likheitar mellom risikogrupper, samt ta høgd for at kvar

risikogruppe har sine eigne risikokarakteristikkar. Utfordringa i tilpassinga av ein GLMM er at dei mange moglegheitene for fordelinga til responsen og link-funksjon vil gjere at ein kan stå ovanfor eit problem med å *spesifisere* modellen. I tillegg vil ein i tilpassinga av diskrete fordelingar kunne få problem med i tilfeller der responsen er null.

Det var i Pinheiro og Bates (2000) si bok eg først støtte på definisjonen av ein lineær blanda modell. Seinare har eg nytta resultat frå Fitzmaurice *et al.* (2004) til å forstå kovariansstrukturen til blanda modellar. Sistnemnte bok har også vore til stor hjelp i forståinga av den generasliserte lineære blanda modellen.

There is no abstract art. You must always start with something. Afterward you can remove all traces of reality.

Pablo Picasso (1881 - 1973)

1

Ordinære lineære regresjonsmodellar

Eg vil starte oppgåva mi med å forklare to fundamentale metodar i statistikk, lineær regresjon og variansanalyse. Dei enklaste versjonane av desse er *enkel lineær regresjon*, og *einvegs*-variansanalyse.

Motivasjonen min i dette kapittelet er å sjå at ein ordinær regresjonsmodell ikkje vil passe så godt til eit datasett som inneheld repeterte målingar på same individ over tid. Datasettet `Orthodont` er eit slikt datasett. Det består av fire målingar av ein bestemt avstand på kvart av 27 individ over ein tidsperiode. Eg vil i dette kapittelet studere tilpassinga av ein enkel lineær regresjonsmodell, og ein ANCOVA-modell til `Orthodont`, og sjå på antakingane som blir gjort når ein tilpassar ordinære regresjonsmodellar til data av repeterte målingar.

For å sjekke tilpassinga av observasjonane i `Orthodont` til desse modellane, vil eg utføre analysar i statistikkprogrammet R. I samanheng med dette vil eg sjå på observatorar som seier noko om kor god ein modell er, såkalla «goodness of fit statistics». Desse gjev ein forsmak på emne modellsamanlikning som eg vil komme sterkare tilbake til i kapittel 3.

1.1 Ein lineær regresjonsmodell

Modellar der ein ser på samanhengen mellom ein målbar respons, til dømes høgd, og element som kan påverke eller forklare verdien til denne responsen, kalla forklaringsvariablar, har navnet *regresjonsmodellar*. Desse er blant dei mest brukte statistiske verktøya. Ein *lineær regresjonsmodell* er ein modell der responsen og *parameterane* til forklaringsvariablane er bunde saman i eit lineært forhold.

Ei standard nemning for respons- og forklaringsvariablar er y_i, $i = 1, 2, \ldots, n$, for observasjonar av ein stokastisk variabel Y, og x_{ij}, $i = 1, 2 \ldots, n$, for observasjonar av ein forklaringsvariabel X_j, der $j = 1, \ldots, p$ er indikator for kva for variabel ein observerer. Dersom $p = 1$, det vil seie at ein antar at responsen avhenger kun av éin forklaringsvariabel, har vi ein *enkel lineær regresjonsmodell*. I dette kapittelet vil vi anta at den stokastiske responsvariabelen Y kjem frå ei normalfordeling. Følgjande definisjon oppsummerar elementa i ein lineær regresjonsmodell:

Definisjon 1.1.1: Ein lineær regresjonsmodell

Ein *lineær regresjonsmodell* er ein modell på forma

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

der $Y_1, Y_2, \ldots, Y_n$ er uavhengige stokastiske variablar frå ei normalfordeling $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, og $X_{i1}, \ldots, X_{ip}$, der $i = 1, \ldots, n$ er forklaringsvariablar, medan $\beta_0, \beta_1, \ldots, \beta_p$ er $p + 1$ ukjente konstantar som ein kallar *parametrar*. Ein slik modell kan uttrykkast på vektorform som

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

der

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbb{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{og} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Det er ikkje alltid at forklaringsvariablane i ein regresjonsmodell er stokastiske. Av og til er dei kontrollerte i forsøket. Tildømes dersom ein ser på høgda til born ved bestemte aldrar, vil forklaringsvariabelen alder vere bestemt på førehand,

og er ikkje stokastisk. I modellane eg skal sjå på for datasettet `Orthodont`, vil forklaringsvariablane ikkje vere stokastiske.

Dersom ein har n observasjonar av ein variabel $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ der ein ynskjer å undersøkje relasjonen variabelen har til n observasjonar av p forklaringsvariablar, $X_{i1}, X_{i2}, \ldots, X_{ip}$, kan ein formulere ein lineær regresjonsmodell for observasjonane som

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n.$$

Denne kan på vektorform uttrykkast som

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.2}$$

der

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{og} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Det er denne modellen eg vil referere til når eg nemner ein ordinær lineær regresjonsmodell i samband med eit datasett.

Forventninga til responsen y_i er lik

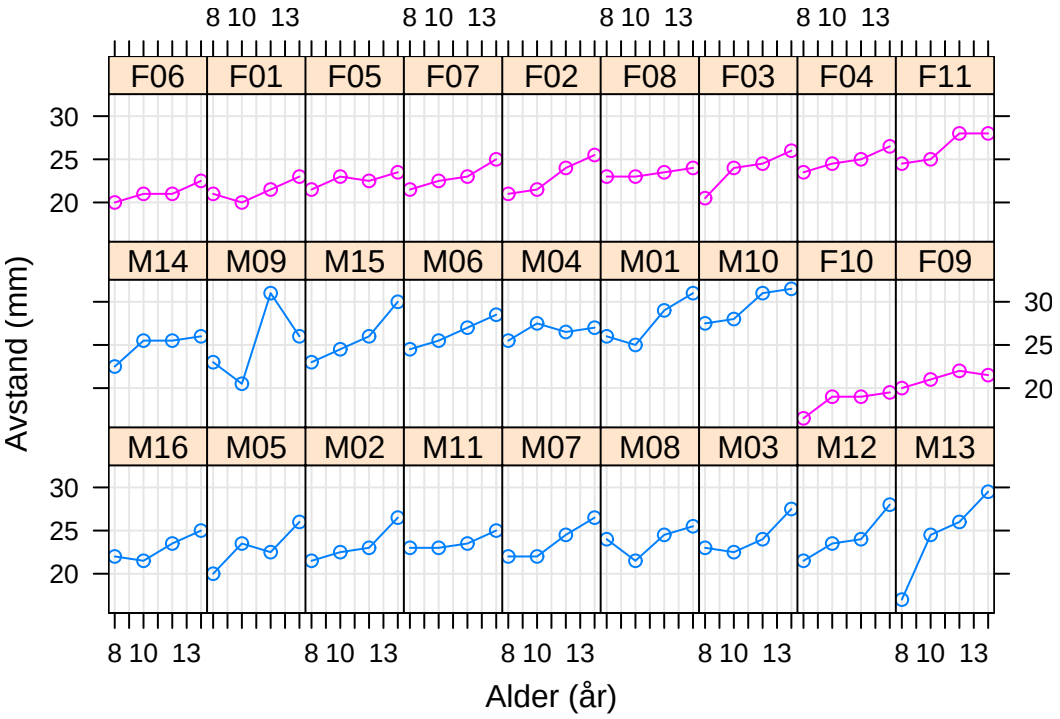
$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mu_i,$$

og variansen til responsen y_i er $\text{Var}(y_i) = \text{Var}(\epsilon_i) = \sigma_i^2$.

Parameteren ϵ_i kallast gjerne støyledd eller residual, og har forventning 0. Denne resresenterer tilfeldig variasjon mellom målingar, og ein antar at den er normalfordelt. Dersom ein antar at alle observasjonane i datasettet ein ser på er utsett for same variasjon, det vil seie at dei alle er likestilte observasjonar av variabelen Y, antar ein at støyledda er normalfordelte med same varians, altså $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Modellparametrane $\beta_0, \beta_1, \ldots, \beta_p$ er dei ukjente storleikane i likninga (1.2), og desse vil ein estimere. Eit krav for å kunne estimere parametrane er at talet på observasjonar, n, er større enn talet på forklaringsvariablar p, altså at $n > p$. Dersom dette ikkje er oppfylt vil ein få ein *overparameterisert* modell. I ein overparameterisert modell vil ein ha fleire parametrar enn likningar og dermed ha uendeleg antall løysingar for dei ukjente parametrane.

Ein har gjerne to hovudformål med å tilpasse ein regresjonsmodell på forma (1.2) til eit datasett av observasjonar. Det eine er å finne samanhengen mellom respons og forklaringsvariablar, for slik å finne ut kva for variablar som har påverknad på responsen. Det andre er å finne ein modell som ein kan nytte til å *predikere* framtidige verdiar av responsen basert på estimerte parameterverdiar og observasjonar av forklaringsvariablane. I begge tilfeller ynskjer ein å finne modellen som passar best til datasettet. Eg vil i denne oppgåva konsentrere meg det første formålet, å finne samanhengen mellom respons og forklaringsvariablar ved å finne modellen som passar best til eit bestemt datasett. Eg skal i dette kapittelet undersøkje nokre mulige lineære regresjonsmodellar for datasettet `Orthodont`.



Figur 1.1: Plott av vekstkurver i datasettet `Orthodont`. Desse viser avstanden mellom hypofysen og «the pterygomaxillary fissure» i millimeter.

1.1.1 Ein enkel lineær regresjonsmodell for Orthodont dataa

Datasettet `Orthodont` inneheld observasjonar av ein ortopedisk avstand hos 11 jenter og 16 gutar målt då borna var åtte, ti, tolv og fjorten år. I statistikkprogrammet R ligg `Orthodont` i pakken `nlme` (Pinheiro og Bates 2000, side 30). Responsvariabelen av interesse i dette datasettet er avstanden mellom to bestemte stadar i kjeven

kalla «pituitary gland» og «pterygomaxillary fissure», og mulige forklaringsvariablar er alder og kjønn. Av plottet, figur 1.1 på førre side, ser vi at den målte avstanden stort sett aukar med alder for individa.

Kommandoen `sapply` i R viser oss strukturen til variablane i datasettet. Datasettet består av fire variablar: avstand og alder som begge er numeriske variablar, individ som er ein ordna vektor, og kjønn som er ein faktor.

```
> sapply(Orthodont,data.class)

 distance       age   Subject       Sex
"numeric" "numeric" "ordered"  "factor"
```

Ein noko naiv modell vil vere å anta at for alle individa aukar avstanden lineært med alder etter same modell. Dette vil svare til ein enkel lineær regresjonsmodell på forma

$$y_i = \beta_0 + \text{alder}_i \cdot \beta + \epsilon_i \quad i = 1, 2, \ldots, 108.$$

Her antar vi at responsvariabelen er normalfordelt som $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, der $\mu_i = E(y_i) = \beta_0 + \text{alder}_i \cdot \beta$. Dersom denne modellen passar, vil det seie at ein kun treng å vite alderen til eit individ for å predikere avstanden Y til individet med rimeleg måleuvisse. Indeksen i står her for observasjonsnummeret i datasettet. Denne modellen har kun tre ukjente modellparametrar, konstantleddet β_0, stigningstalet β og residualvariansen σ^2.

Potthoff og Roy (1964) viste tidleg i sine analysar av vekstkurvemodellar at kjønn vil vere ein påverkande faktor for responsen når ein studerar vekst hos menneske. Dei definerte ein *multivariat* modell for observasjonane i `Orthodont`. Eg vil først vil eg definere ein enkel lineær regresjonsmodell for observasjonane i `Orthodont` på vektorform der indikatoren i no representerer individ i, slik at $i = 1, \ldots, 27$. Eg vil no uttrykke alle observasjonane til individ i ved ein vektor $\boldsymbol{y}_i$ på form

$$\boldsymbol{y_i} = \begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \end{bmatrix}, \quad i = 1, \ldots, 27.$$

Sidan datasettet er *balansert*, det vil seie at vi har like mange observasjonar av kvart individ, vil denne vektoren ha same dimensjon for alle individa.

Ein enkel lineær regresjonsmodell for observasjonane i `Orthodont` kan no formulerat på vektorform som

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad i = 1, \ldots, 27, \tag{1.3}$$

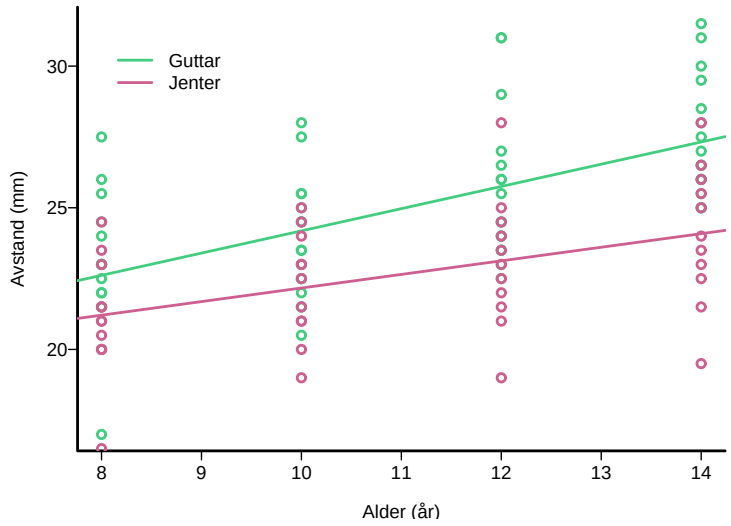
der

$$\boldsymbol{X} = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \quad \text{og} \quad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \epsilon_{3i} \\ \epsilon_{4i} \end{bmatrix}.$$

I modell (1.3) antar vi at alle observasjonar er uavhengige slik at responsvektoren $\boldsymbol{y}_i$ er *multivariat* normalfordelt på form

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2\boldsymbol{I}) \quad \text{der} \quad \boldsymbol{\mu}_i = E(\boldsymbol{y}_i) = \boldsymbol{X}\boldsymbol{\beta}.$$

og $\text{Var}(\boldsymbol{y}_i) = \sigma^2\boldsymbol{I}$ ei matrise med σ^2 langs diagonalen og null ellers.



Figur 1.2: Plott av avstand mot alder for jenter og gutar kvar for seg i settet Orthodont. Punkta viser observasjonane for dei respektive kjønna.

For å undersøkje ein eventuell forskjell mellom kjønn, plotta eg avstand mot alder for gutar og jenter separat. Dette resulterte i figur 1.2. Felles for alle individa, uavhengig av kjønn, er at avstanden ved alder 14 år er større enn ved alder 8 år. Altså tyder alt på at alder er ein forklaringsvariabel for responsen avstand. Men figuren viser at det er ein kjønnsforskjell. Jentene har generelt mindre avstandar enn gutane. Regresjonslinja for jentene har både ein lågare start- og sluttverdi enn regresjonslinja til gutane.

Når vi i neste steg skal inkludere kjønn for individ i i modellen, er det viktig å forstå kva det vil seie at ein variabel er ein faktor. For å forklare dette vil eg definere ein ANOVA modell.

1.2 Variansanalyse

Terminologien «analysis of variance», med forkortinga ANOVA, blir nytta for modellar der responsvariabelen er kontinuerlig og forklaringsvariablane kategoriske eller kvalitative. Slike forklaringsvariablar kallast faktorar, og desse har nivå eller undergrupper som ein ynskjer å finne ut om påverkar responsen. Kjønn er ein faktor som har to nivå eller undergrupper, jente og gut. Andre faktorar kan ha mange nivå. I ein ANOVA modell består designmatrisa $\boldsymbol{X}_i$ derfor gjerne av 1 og 0 for å representere nivåa til faktoren.

Den enklaste ANOVA-modellen er ein *éin-faktor*-modell, på engelsk kalla «one-way ANOVA». Eg gje eit døme på ein generell éin-faktor-modell.

Eksempel 1.2.1: Ein generell éin-faktor-modell

Dersom ein ynskjer å studere om ein enkelt faktor α, med $j = 1, \ldots, J$ nivå, påverknar ein respons y_i, der ein har $i = 1, \ldots, n$ observasjonar av responsen på kvart nivå av faktoren, kan ein formulere ein éin-faktor-modell på form

$$\begin{aligned} &y_{ij} = \mu + \alpha_j + \epsilon_{ij}, \quad j = 1, \ldots, J \quad i = 1, \ldots, n \quad \text{og} \\ &\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2). \end{aligned}$$

Denne kan uttrykkast på vektorform som

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n, \tag{1.4}$$

der

$$\boldsymbol{y}_i = \begin{bmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Ji} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\alpha}_i = \begin{bmatrix} \mu \\ \alpha_{1i} \\ \vdots \\ \alpha_{Ji} \end{bmatrix} \quad \text{og} \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \vdots \\ \epsilon_{Ji} \end{bmatrix}. \tag{1.5}$$

Når kvart nivå av faktoren, $j = 1, \ldots, J$, har like mange observasjonar, n, som i dette eksempelet, seier vi at vi har ein *balansert* ANOVA-modell.

For å unngå overparametrisering i ANOVA-modellar må ein innføre nokre restriksjonar. Ofte set ein $\mu = 0$. Ein kan også sette $\mu = \alpha_1$. Det tilsvarar ein *hjørne restriksjon*, på enkelst kalla «corner-point restriction». Ein anna restriksjon er *nullsum*, på engelsk kalla «sum-to-zero». Den går ut på å setje $\alpha_1 + \alpha_2 + \cdots + \alpha_J = 0$ (Dobson 2002, side 98).

1.2.1 Ny modell for Orthodont dataa

Eg vil no svare på spørsmålet om korleis ein kan inkludere ein faktor i ein regresjonsmodell. Eg ynskjer å inkludere faktoren kjønn i den enkle regresjonsmodellen (1.3) på side 8.

Vi ser først på ei utviding av modell (1.3) som inkluderar kjønn, og som inneheld fire parametrar. I tillegg til dei to parametrane i modell (1.3) kjem éin parameter for kjønn og éin parameter for samspelsleddet mellom alder og kjønn. Variabelen kjønn er ein faktor, og denne definerar eg i modellen med ein forklaringsvariabel, 1_i som er lik -1 eller 1. Parameteren betyr altså at vi har større verdi av responsen for eitt av kjønna. Samspelsleddet står for muligheita for at eitt av kjønna i ein viss alder påverkar responsen meir enn gjennomsnittleg.

Modellen er på forma

$$y_i = \beta_0 + \beta_1 \cdot \text{alder}_i + \beta_2 \cdot \text{kjønn}_i + \beta_3 \cdot (\text{kjønn i viss alder})_i + \epsilon_i,$$

der $i = 1, \ldots, 27$, og feil-ledda, ϵ_i, er normalfordelte som før. I denne modellen treng vi å vite både alder og kjønn til individet for å estimere responsen avstand. Pinheiro og Bates har ikkje formulert ein modell for begge kjønn i si analyse, og dermed har dei ikkje trengt å definere matrisa $\boldsymbol{X}_i$ når kjønn inkluderast. Eg vil definere denne matrisa ved ein indikatorvariabel $1_{\text{kjønn}} = 1_i$, som har verdiar

$$1_i = \begin{cases} 1, & \text{dersom individet er ein gut} \\ -1, & \text{dersom individet er ei jente} \end{cases}$$

Dermed vil regresjonsmodellen

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad i = 1, \ldots, 27 \tag{1.6}$$

bestå av vektorane $\boldsymbol{y}_i$ og $\boldsymbol{\epsilon}$ som er definerte som før, mens

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 8 & 1_i & 8 \cdot 1_i \\ 1 & 10 & 1_i & 10 \cdot 1_i \\ 1 & 12 & 1_i & 12 \cdot 1_i \\ 1 & 14 & 1_i & 14 \cdot 1_i \end{bmatrix}, \quad \text{og} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

Gutane vil ha $\boldsymbol{X}_i$-matriser på forma

$$\boldsymbol{X}_1 = \boldsymbol{X}_1 = \cdots = \boldsymbol{X}_{16} = \begin{bmatrix} 1 & 8 & 1 & 8 \\ 1 & 10 & 1 & 10 \\ 1 & 12 & 1 & 12 \\ 1 & 14 & 1 & 14 \end{bmatrix},$$

medan jentene vil ha $\boldsymbol{X}_i$-matriser på forma

$$\boldsymbol{X}_{17} = \boldsymbol{X}_{18} = \cdots = \boldsymbol{X}_{27} = \begin{bmatrix} 1 & 8 & -1 & -8 \\ 1 & 10 & -1 & -10 \\ 1 & 12 & -1 & -12 \\ 1 & 14 & -1 & -14 \end{bmatrix}.$$

Vi vil i modellen (1.6) på førre side anta at støyledda er uavhengige og normalfordelte på forma $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

Pinheiro og Bates (2000) har kalla ein liknande modell for settet `Orthodont` for ein ANCOVA-modell. Uttrykket ANCOVA er kort for «analysis of covariance». Ein ANCOVA-modell er ein modell der responsen avhenger av forklaringsvariablar som sjølv er avhengige mellom forsøk. Ein ANCOVA-modell bind saman ein kontinuerlig respons, som avstand i dette tilfellet, til både ein klassifiseringsfaktor, som kjønn, og ein kontinuerlig kovariat, som alder i vårt eksempel (Pinheiro og Bates 2000, side 30).

1.3 Sjekk av modellar

Etter å ha spesifisert modellen ynskjer ein å finne ut om den passar dataa våre godt.

Terminologien «goodness of fit» går igjen når ein i lærebøker les om godheita til ein modell som er tilpassa eit bestemt datasett. For å kunne tolke verdiane av desse må ein vite korleis dei oppstår og kva dei måler.

1.3.1 Modelltilpassing

I store datasett er det gjerne mange element ein ynskjer å undersøkje om påverkar responsen. Ein kan ha multiple regresjonsmodellar, som den i likning (1.1) på side 4. Ein kan ha to-faktor-modellar, på engelsk kalla «two-way anova», der ein måler effekten av to faktorar simultant, og modellar med endå fleire faktorar. Vidare kan ein også ha ein multivariat multippel regresjonsmodell, der ein måler effekten av forklaringsvariablar på fleire responsvariablar samstundes, og ein kan ha MANOVA- og MANCOVA-modellar der ein måler påverknaden av nivåa ein eller fleire faktorar har på responsar frå ulike populasjonar, (Johnson og Wichern 2002). Men ein ynskjer å finne ein enklast mulig modell som forklarar responsen best mulig.

Jo fleire parametrar vi har med i modellen vår, jo meir nøyaktig vil høgresida i modellen forklare variasjonenen i responsen. For å sette ting på spissen vil ein modell som forklarar 100 % av variasjonen i datasettet, innehalde ein parameter for kvar av observasjonane våre. Men ein slik modell vil vere ubrukelig sidan den ikkje fortel oss noko nytt om forholdet mellom respons og forklaringsvariablar.

Modellen med flest mulig forklaringsvariablar, som tek med alle element vi kan forestille oss at vil bidra til ei endring i responsen, kallast ein *metta* modell. Dersom ein variabel verkeleg påverkar responsen, altså skiljer seg frå tilfeldig støy, ϵ_i, seier vi at variabelen eller faktoren er statistisk *signifikant*. Dersom ein, ved å fjerne ein variabel frå den metta modellen, får ein modell som forklarer omtrent like mykje av variansen i responsen, kan vi konkludere med at den ekskluderte variabelen ikkje påverkar responsen, og dermed *ikkje* er statistisk signifikant. Vidare kan vi fjerne endå ein forklaringsvariabel, og undersøkje ei mulig endring på nytt. For å avgjere om endringa i responsen er signifikant ved eksklusjon eller inklusjon av ein variabel kan vi formulere ein hypotesetest.

1.3.2 Hypotesetesting

Ein kan formulere ei hypotese som testar tilpassinga til to modellar, ein generell med p variablar, og ein spesifikk med $q < p$ variablar der nokre av dei p variablane er sett lik null, som

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0 \quad \text{mot} \quad H_1 : \text{minst éin ulik null.} \tag{1.7}$$

Vi benyttar så ein observator kalla Fisher-observator, eller berre F-observator, som ikkje avheng av den ukjente variansen σ^2, til responsvariabelen. Denne er definert som (Dobson 2002, side 82)

$$F = \frac{D_0 - D_1}{p - q} / \frac{D_1}{n - q}, \tag{1.8}$$

der D_0 og D_1 er kalla *deviansen* til høvevis modellen under H_0 med q parametrar og modellen under H_1 med p parametrar, og der n er talet på observasjonar. Deviansen er definert som

$$D_j = \frac{1}{\sigma^2}(\boldsymbol{y} - \hat{\boldsymbol{y}}^{(j)})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}^{(j)}), \quad j = 0{,}1. \tag{1.9}$$

der $\boldsymbol{y}$ er vektoren med observasjonane våre, og $\hat{\boldsymbol{y}}^{(j)}$ er predikerte modellverdiar definert som

$$\hat{\boldsymbol{y}}^{(j)} = \boldsymbol{X}^{(j)}\hat{\boldsymbol{\beta}}^{(j)} \quad j = 0, 1$$

der j viser om vi ser på predikerte verdiar til modellen under H_0 eller under H_1. Parametervektoren $\boldsymbol{\beta}$, består under H_0 av $\boldsymbol{\beta}^{(0)} = [\beta_1, \ldots, \beta_q]$, og under H_1 av $\boldsymbol{\beta}^{(1)} = [\beta_1, \ldots, \beta_p]$. Matrisa $\boldsymbol{X}^{(j)}$ inneheld dei respektive forklaringsvariablane.

Testobservatoren F i likning (1.8) har ei sentral $F_{p-q,n-q}$ fordeling dersom H_0 er korrekt. Dersom H_0 ikkje er korrekt vil observatoren ha ei ikkje-sentral F-fordeling. Dermed vil store observasjonsverdiar $\hat{F}$, av F, gi grunnlag for å forkaste H_0. Store $\hat{F}$-verdiar gir små p-verdiar, $\text{p} = P(F > \hat{F} \mid H_0 \text{ er sann})$.

At den observerte verdien $\hat{F}$ vil bli liten når H_0 er sann og stor når H_0 er usann, kan vi òg sjå av likningane (1.8) og (1.9). Dersom modellen under H_0 passar dataa godt, vil den tilhøyrande deviansen vere liten, og teljaren i likning (1.8) også vere liten, sidan D_1 alltid vil vere mindre enn D_0. Dette er fordi ein modell med fleire parametrar aldri vil forklare responsen dårligare enn ein modell der nokre av parametrane er sett lik null. Dermed vil ein ikkje forkaste modellen under H_0 når $\hat{F} < F_{n-p,p-q}(\alpha)$. Dersom ein har avhengige forklaringsvariablar kan ein få ulike konklusjonar alt ettersom kva for forklaringsvarablar ein har i den generelle modellen. Dette er fordi ein ved å fjerne ein avhengig variabel, vil oppleve at dei gjenværande variablane også blir endra.

Observatoren F er éin av fleire observatorar som kallast «goodness of fit statistics», på norsk oversett til godheitsobservatorar, ettersom dei vurderar kor godt ein modell passar til eit datasett. Ein annan godheitsobservator er R^2. Observatoren R^2 er definert som andelen av den totale variansen modellen vår forklarar (Dobson

2002, side 94). I tilfellet der vi har éin parameter for kvar observasjon, slik at all variasjon i observasjonene blir forklart, vil R^2 vere lik 1.

1.4 Analysar i R

Vi går tilbake til individa i `Orthodont`, og den enkle regresjonsmodellen i likning (1.3) på side 8. I R kan vi definere ein enkel lineær regresjonsmodell for `Orthodont` som

```
> fit00 <- lm(distance~age,data=Orthodont)
```

Med kommandoen `summary(fit00)` vil vi så kunne studere verdiar av dei ukjente parametrane β_0, β og σ^2, og verdi av observatoren F. Vi vil samanlikne desse verdiane med dei vi får for modellen i likning (1.6) på side 10. Modellen (1.6) kan vi definere i R som

```
> fit0 <- lm(distance~age*Sex,data=Orthodont)
```

Multiplikasjonsteiknet `*` mellom alder og kjønn i koden, er ein forkorta skrivemåte for `distance = age + Sex + age:Sex`, der det siste leddet står for samspelet mellom alder og kjønn. Denne modellen har ukjente parametrar $\beta_0, \beta_1, \beta_2, \beta_3$ og σ^2. I tabell 1.1 står verdiane R gav meg for desse modellane.

	Modell (1.3)		Modell (1.6)	
	Estimat	St.avvik	Estimat	St.avvik
β_0	16,761	1,226	16,3406	1,4162
β_1	0,6602	0,1092	0,7844	0,1262
β_2			1,0321	2,2188
β_3			−0,3048	0,1977
σ^2	6,436		5,094	
p_1	1		3	
$N - p_1$	107		105	
$\hat{F}(p_1, N - p_1)$	36,56		25,39	
R^2	0,2495		0,4061	

Tabell 1.1: Estimerte verdiar av parametrar og observatorar i den enkle regresjonsmodellen og ANCOVA modellen for datasettet `Orthodont`.

I tabellen er N lik talet på observasjonar, og $N - p$ talet på observasjonar minus talet på parametrar. Vi ser frå tabellen at modellen som inkluderar faktoren kjønn har ein høgare verdi av R^2, og forklarar altså meir av variasjonen i dataa enn modellen med kun alder som forklaringsvariabel. Modellen som inkluderar

kjønn har også ein mindre residualvarians. Men, som utskrifta i A.1 viser, er parametrane for kjønn og interaksjonsleddet mellom kjønn og alder ikkje statistisk signifikante. Særleg parameteren til kjønn, β_2, har fått ein høg p-verdi. I tabellen ser vi at estmatet av denne parameteren også har eit stort standardavvik. Ein systematisk sjekk viser at ved eksklusjon av variabelen kjønn blir parameteren til interaksjonsleddet statistisk signifikant.

1.4.1 Hypotestesting i R

Med resultata i tabell 1.1 på førre side vil vi teste ei hypotese på forma

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{mot} \quad H_1 : \text{minst éin ulik null}$$

for å finne ut kva for modell som passar best til dataa. Til dette bruker vi funksjonen `anova`.

```
> anova(lm1,lm2)

Analysis of Variance Table

Model 1: distance ~ age
Model 2: distance ~ age * Sex

  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    106 682.34
2    104 529.76   2    152.58 14.977 1.923e-06 ***

Signif. codes: '***' 0.001
```

Her blir modellen (1.3), der kun alder er forklaringsvariabel, forkasta på eit $\alpha = 0{,}001$ nivå. Dermed kan vi foreløpig konkludere med at både alder og kjønn påverkar responsen avstand. Sjølv om analysa i R viste at leddet `Sex` ikkje er statistisk signifikant, vil eg foreløpig halde på modellen i likning (1.6), som inkluderer parameteren β_2 for kjønnet til individet. Men eg ynskjer å finne ein modell som forklarar endå meir av variasjonen i datasettet enn det modell (1.6) gjer, og som dermed har ein mindre residualvarians.

Vi ser av figur 1.1 på side 6 at det ser ut til å vere stor variasjon i start- og sluttverdi til responsen avstand for dei ulike individa, både mellom og innad kjønn. Sidan vi enno ikkje har teke omnsyn til faktoren individ kan dette vere forklarande for den låge R^2-verdien til modell (1.6). Korleis kjem denne faktoren inn i bilete? Vi må no forlate den generelle regresjonsmodellen og gå over til ein blanda modell.

A theory has only one alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.

Manfred Eigen

2

Lineære blanda modellar

Dette kapittelet tek for seg teorien bak blanda modellar, på engelsk kalla «mixed effects models» eller «two-stage models», som er modellar med *faste* og *variable parametrar*. Den blanda modellen er hovudsakeleg aktuell i tilfelle der ein har målingar av fleire individ eller objekt over tid, såkalla *longitudinelle* observasjonar, og der ein har grupperte målingar, til dømes målingar på tvers av geografiske område, også kalla klyngedata, på engelsk «clustered data». Både longitudinelle observasjonar og klyngedata blir gjerne nemnt som *repeterte* målingar.

Modellane vi hittil har sett på er modellar med kun faste parametrar. I desse var det kun støyledda som var kilde til variasjon. Eg vil i dette kapittelet introdusere den lineære blanda modellen, og tilpasse to blanda modellar til settet `Orthodont`.

Eg har henta teorien min hovudsakeleg frå Pinheiro og Bates (2000), Dobson (2002), Fitzmaurice *et al.* (2004) og Molenberghs og Verbeke (2004).

2.1 Ein lineær blanda modell

Blanda modellar er regresjonsmodellar som består av to delar, éin fast og éin variabel. Uttrykket *blanda* speglar på at modellen har ei blanding av faste og

variable komponentar. Eg vil sjå på den *lineære blanda modellen*. I den lineære blanda modellen opptrer både dei faste og dei variable komponentane lineært i modellen. Når ikkje anna er spesifisert, antar vi at observasjonane våre er normalfordelte.

Den *faste* delen av modellen består av parametrar som vi antar gjeld for ein heil populasjonen. Desse har terminologien «fixed effects» på engelsk, og på norsk kan ein kalle dei *faste parametrar*. Dei faste parametrane skal fortelje oss noko generelt om høve mellom responsen og variablane i forsøket uttrykt ved parametrar som gjeld for alle objekta vi har observert. Modellane vi hittil har sett på har alle hatt kun faste parametrar.

Den *variable* delen består av parametrar knytte til eksperimentelle einingar i forsøket vårt, og vert kalla «random effects» på engelsk. Ei eksperimentell eining kan være eit måleinstrument, eit område, ei tidsavgrensing, og andre element som kan føre til at vi får større varasjon av ein respons mellom to einingar enn innad eininga. Parametrane kallast *variable*, på engelsk «random», fordi dei ikkje har same verdi for ulike individ eller grupper i modellen. Parametrane føl ei eiga sannsynsfordeling, med forventning null, og vert lagt til ei delmengde av dei faste parametrane i modellen, til dømes til konstantleddet i modellen. Det er desse variable parametrane som skil blanda modellar frå vanlege regresjonsmodellar. Vi er interesserte i å estimere variansen til desse parametrane, sidan variansen vil fortelje oss kor mykje parametrane varierar mellom einingane.

Ein generell lineær blanda modell kan skrivast på forma

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}). \end{aligned} \tag{2.1}$$

Her representerer vektoren $\boldsymbol{y}_i$ observasjonane på den i-te eininga, matrisa $\boldsymbol{X}_i$ er designmatrisa med av observasjonar på $p \geq 1$ forklaringsvariablar i forsøket, og vektoren $\boldsymbol{\beta}$ inneheld dei tilhøyrande $p + 1$ faste parametrane. Matrisa $\boldsymbol{Z}_i$ representerer ei delmengd av designmatrisa $\boldsymbol{X}_i$, og består av variablar som vi vil legge til variable parametre. Denne har dimensjon $q \times n$ der $q \leq p$. Vektoren $\boldsymbol{b}_i$ består av dei tilhørande parametrane $b_{i0}, b_{i1}, \ldots, b_{iq}$. Vi antar at desse parametrane er multivariat normalfordelte med forventningsvektor $\boldsymbol{0}$ og kovariansmatrise $\boldsymbol{\Psi}$, og at dei er uavhengige av støyledda $\boldsymbol{\epsilon}_i$.

Til forskjell frå elementa i residualvektoren $\boldsymbol{\epsilon}_i$, er elementa i vektoren med dei variable parametrane $b_{i0}, b_{i1}, \ldots, b_{iq}$ til same objekt i, korrelerte. Kovariansmatrisa $\boldsymbol{\Psi}$ er dermed ikkje ei diagonalmatrise slik kovariansmatrisa til støyledda er, men

ei symmetrisk matrise. Vi skal også, i følge Pinheiro og Bates, anta at matrisa er *positivt definitt*. Ei positivt definitt matrise har kun positive eigenverdiar (Walpole *et al.* 1998, side 000). Derimot er variable parametrar for ulike objekt ikkje korrelerte. Som nemnt er det kovariansmatrisa $\boldsymbol{\Psi}$, og ikkje vektoren $\boldsymbol{b}_i$, ein typisk er interessert i å estimere.

2.2 To synsvinklar

Eit anna navn for ein blanda modell er *tostegs*-modell, eller på engelsk *two-stage model*. Dette navnet speglar på at ein kan sjå på modellen i to steg. I det første steget har modellen ei populasjonsretta tolkning. Vi ser då kun på dei faste parametrane. I det andre ser vi på modellen som individsretta. Vi inkluderar då dei variable parametrane. Fitzmaurice *et al.* (2004) påpeikar at sjølv om tostegsformuleringa er hjelpsam i forklaringa og forståinga av ein blanda modell, så medfører formuleringa nokre unødvendige restriksjonar. Derfor vil eg ikkje omtale ein blanda modell som ein tostegs-modell. Forklaringa derimot, på tolkninga av parametrane, vil eg ta med meg vidare.

Molenberghs og Verbeke (2004) hevdar at ein kan innta to ulike synsvinklar på fordelinga til responsvariabelen i ein blanda modell, alt ettersom vi ser på den *marginale* eller den *hierarkiske* fordelinga. Sjølv om fordelingane er like, vil ein kunne gjere ulike antakelsar alt ettersom ein ser på fordelinga til responsen frå ein marginal eller ein hierarkisk synsvinkel. Molenberghs og Verbeke gir med dette to navn på ein blanda modell, ein marginal blanda modell dersom ein ser på den marginale fordelinga til responsen, og ein betinga eller hierarkisk blanda modell dersom ein ser på fordelinga til responsen frå eit hierarkisk, eller betinga, synspunkt. Som i tostegs-modellen vil vi i den marginale modellen ha ei *populasjonsretta* tolkning, medan vi i den hierarkiske modellen vil ha ei *individsretta* tolkning.

Den marginale fordelinga til responsen er gitt som

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \tag{2.2}$$

for observasjonar $i = 1, \ldots, n$, der $E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ og $\text{Var}(\boldsymbol{y}_i) = \boldsymbol{V}_i = \sigma^2\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T$. Når vi ser på fordelinga til responsen på denne måten, representerar forventningsvektoren $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ den forventa responsvektoren til populasjonen. Dette gjev oss eit estimat av dei faste parametrane i modellen.

Den hierarkiske eller betinga fordelinga til responsen, definert som

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \sigma^2\boldsymbol{I}) \\ \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \end{aligned} \tag{2.3}$$

vil derimot ha ein forventningsvektor som gir forventa respons for eit spesifikt objekt i tillegg til den generelle forventninga, sidan $E(\boldsymbol{y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$. Marginalt vil dette gje same fordeling som den i likning (2.2) på førre side, sidan

$$\begin{aligned} E(\boldsymbol{y}_i) &= EE(\boldsymbol{y}_i|\boldsymbol{b}_i), \quad \text{og} \\ \operatorname{Var}(\boldsymbol{y}_i) &= E\operatorname{Var}(\boldsymbol{y}_i|\boldsymbol{b}_i) + \operatorname{Var}E(\boldsymbol{y}_i|\boldsymbol{b}_i) \\ &= E\left[\sigma^2\boldsymbol{I}\right] + \operatorname{Var}\left[\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i\right] \\ &= \boldsymbol{0} + \operatorname{Var}\boldsymbol{\epsilon}_i + \boldsymbol{Z}_i\operatorname{Var}\boldsymbol{b}_i\boldsymbol{Z}_i^T \\ &= \sigma^2\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T = \boldsymbol{V}_i \end{aligned}$$

ved kjente teorem (Casella og Berger 2002, side 164 og 167).

2.2.1 Kovariansstruktur

Eg vil gå litt nærare inn på kovariansstrukturen i dei to modellane. I tillegg til at den marginale har ei populasjonsretta, og den betinga ei individsretta tolkning, har dei to modellane eit ulikt syn på varianselementa i kovariansmatrisene sine. Som det kjem fram av likning (2.2) og (2.3), gjer ein i den betinga modellen antakingar om kovariansstrukturen på to nivå. I den marginale modellen vil kovariansmatrisa $\boldsymbol{V}_i$ bestå av både residualvarians og individ-spesifikk varians. Den har forma (Fitzmaurice *et al.* 2004, side 198)

$$\begin{aligned} \boldsymbol{V}_i &= \operatorname{Cov}(\boldsymbol{y}_i) = \operatorname{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) + \operatorname{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\operatorname{Cov}(\boldsymbol{b}_i)\boldsymbol{Z}_i^T + \operatorname{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}. \end{aligned}$$

Matrisa $\boldsymbol{V}_i$ er ikkje-diagonal sidan målingar på same individ er korrelerte, det vil seie $\operatorname{Cov}(y_{ij}, y_{ik}) \neq 0$ for to målingar j og k på eit individ i. I følge Molenberghs og Verbeke (2004) vil ein anta at matrisa $\boldsymbol{V}_i$ er positivt definitt.

For kovariansmatrisene i den betinga fordelinga derimot ser vi på

$$\begin{aligned} \operatorname{Cov}(\boldsymbol{y}_i|\boldsymbol{b}_i) &= \operatorname{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2\boldsymbol{I}, \quad \text{og} \\ \operatorname{Var}(\boldsymbol{b}_i) &= \boldsymbol{\Psi}. \end{aligned}$$

I denne modellen antar ein, i følge Molenberghs og Verbeke, at både matrisa $\sigma^2 \boldsymbol{I}$, og matrisa $\boldsymbol{\Psi}$ er vere positivt definitte. Dette er ei strengare antaking enn i den marginale modellen. Ei følge er at to ulike betinga modellar kan gje same marginale modell. Ein annan konsekvens av dei ulike restriksjonane er at inferens for elementa i kovariansmatrisene blir ulike.

Eg vil no gje nokre dømer på lineære blanda modellar. Eg vil studere litt nærare kva det betyr for kovarianselementa i desse modellane at ein ser på respektivt ei marginal og ei betinga fordeling for responsen.

2.3 Ein random intercept-modell

Ein blanda modell med éin random effect er ein modell der vi antar at dei eksperimentelle einingane påverkar kun éin parameter i modellen vår. Den enklaste typen lineær blanda modell får vi når vi antar at kun konstantleddet varierar mellom dei eksperimentelle einingane. Ein slik modell kallast ofte for ein *random intercept-modell*.

Går vi tilbake til `Orthodont` kan no sjå på individ som ein eksperimentell faktor med 27 einingar. Istadanfor å anta at personar med same kjønn og alder veks i følge ein modellen med same konstantledd kan vi no anta at startverdien for avstanden varierar mellom individ. Dersom vi tek omsyn til at denne startverdien er ulik, men vil undersøkje om individa veks like raskt kan vi formulere ein random intercept-modell for individa i `Orthodont` som

$$\begin{aligned} &\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i + \boldsymbol{\epsilon}_i, \\ &\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \quad b_i \sim \mathcal{N}(0, \sigma_b^2), \quad i = 1, \ldots, 27 \end{aligned} \tag{2.4}$$

der $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_3)^T$ er av same storleik som i modell (1.6) på side 10, parameteren b_i er eit tillegg i konstantverdi for individ i, og matrisene $\boldsymbol{X}_i$ og $\boldsymbol{Z}_i$ har elementa

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 8 & 1_i & 8 \cdot 1_i \\ 1 & 10 & 1_i & 10 \cdot 1_i \\ 1 & 12 & 1_i & 12 \cdot 1_i \\ 1 & 14 & 1_i & 14 \cdot 1_i \end{bmatrix}, \quad \text{og} \quad \boldsymbol{Z}_i = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

I modellen i likning (2.4) har vi no definert ei ny kjelde til variasjon, nemleg det stokastiske konstantleddet b_i, som på engelsk gjev opphav til navnet «random

intercept-modell». Eg vil nytte det engelske navnet på denne modellen, sidan den norske oversetjinga, «ein stokastisk konstantledd-modell», er tungvindt.

Som nemnt er det ikkje sjølve parameteren b_i ein som regel er interessert i, men variansen til parameteren, $\text{Var}\, b_i = \sigma_b^2$. Dersom det er stor variasjon i startverdi mellom individa, vil dette gjenspeglar seg i ein stor varians, σ_b^2, for det stokastiske konstantleddet. Dersom det er liten variasjon i startverdi, slik at parameteren b_i er overflødig i modellen, vil vi forvente ein liten verdi av variansen σ_b^2. Dei ukjende parametrane vi ynskjer å estimere i modellen kan oppsummerast i ein vektor θ med seks element, $\theta = (\beta_0, \ldots, \beta_3, \sigma, \sigma_b)$. Dersom σ_b^2 er tilnærma lik null vil eg behalde modellen frå kapittel 1 som ikkje har eit stokastisk konstantledd.

Ei minke i residualvarians vil gje oss ein indikasjon på om den blanda modellen er betre enn den ordinære forgjengaren. Før eg analyserer modellen i R vil eg sjå litt nærare på kovariansstrukturen i ein random intercept-modell.

2.3.1 Kovarians i random intercept-modell

Når ein legg til eit stokastisk stigningstal, endrar ein synspunktet på kovariansen mellom observasjonane. Ein vil då anta at observasjonar på same individ i er korrelerte, slik at

$$\text{Cov}(y_{ij}, y_{ik}) = \text{Cov}(b_i + e_{ij}, b_i + \epsilon_{ik}) = \text{Cov}(b_i, b_i) = \sigma_b^2.$$

For observasjonar på ulike individ, i og k, derimot har ein ingen korrelasjon. Variansen til ein vilkårlig observasjon på individ i vil dermed vere modellert som

$$\text{Var}(y_{ij}) = \sigma_b^2 + \sigma^2, \quad j = 1, \ldots, 4, \quad i = 1, \ldots, 27.$$

Dermed vil kovariansmatrisa under den marginale random intercept-modellen ha utsjånad (Fitzmaurice *et al.* 2004, side 192)

$$\text{Cov}(\boldsymbol{y}_i) = \boldsymbol{V}_i^{(1)} = \begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 \end{bmatrix}, \tag{2.5}$$

der σ_b^2 er variansen til random effects parameterane, og σ^2 er residualvariansen. Eit krav for denne matrisa er at den må vere *positivt definitt*.

Korrelasjonen til to observasjonar y_{ij} og y_{ik} er definert som

$$\rho_i = \text{Corr}(y_{ij}, y_{ik}) = \frac{\text{Cov}(y_{ij}, y_{ik})}{\sqrt{\text{Var}(y_{ij})}\sqrt{\text{Var}(y_{ik})}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Negativ korrelasjon for observasjonar på eit individ i tyder at $\rho_i < 0$, som medfører at $\sigma_b^2 < 0$. Dette vil ikkje stride mot antakinga om at kovariansmatrisa (2.5) på førre side er positivt definitt. I den marginale modellen kan vi dermed anta at varianskomponenten σ_b^2 har eit definisjonsområde definert ved $\sigma_b^2 \in (-\infty, \infty)$. Det kan sjå rart ut at kvadratet av noko kan bli negativt, men tolkninga er rimeleg (Molenberghs og Verbeke 2004).

I den betinga modellen antar vi at kovariansmatrisa $\boldsymbol{\Psi}$, beståande utelukkande av elementa σ_b^2, er positivt definitt. Det vil medføre at vi antar at $\sigma_b^2 > 0$, og dermed at negativ korrelasjon ikkje er mulig. I den betinga modellen antar vi derfor at varianselementet σ_b^2 har er definert i området $\sigma_b^2 \in (0, \infty)$, der 0 ligg på randa av definisjonsområdet.

Ein hypotesetest på varianskomponentane i modellen, vil derfor ha ulike formuleringar alt ettersom kva for synsvinkel ein brukar. Eg vil komme nærare tilbake til dette i eksempel 3.3.3 på side 44. Først vil eg definere endå ein mulig modell for individa i `Orthodont`.

2.4 Blanda modellar med fleire variable parametre

Blanda modellar med fleire variable parametrar er modellar der vi legg til variabele parametrar på meir enn èin av dei faste parametrane i modellen. Vi kan til dømes ha modellar med stokastiske samspelsledd eller der nokre stigningstal i modellen er stokastiske. Vi skal no sjå korleis vi kan tilpasse ein modell med to variable parametrar for `Orthodont`.

For individa i `Orthodont` kan vi, som nemnt over, ta høgde for at vekstkurvene til individa følger ulike stigningstal ved undersøkje ein modell der vi også legg til eit stokastisk stigningstal b_{1i} til stigningstalet β_1 i modell (2.4).

Ein modell med to variable parametrar for individa i `Orthodont` kan definerast som

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}) \quad i = 1, \ldots, 27. \end{aligned} \tag{2.6}$$

Endringane i denne modellen frå modell (2.4) på side 20 ligg i matrisa $\boldsymbol{Z}$, vektoren $\boldsymbol{b}_i$ og i matrisa $\boldsymbol{\Psi}$.

Matrisa $\boldsymbol{Z}$ er no 4×2. Den er framleis uavhengig av kjønn, og består no av eittala i modell (2.4) på side 20 og vektoren `age`. Altså

$$\boldsymbol{Z} = \begin{bmatrix} 1 & 8 \\ 1 & 10 \\ 1 & 12 \\ 1 & 14 \end{bmatrix}, \quad \text{mens} \quad \boldsymbol{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}.$$

Den tilhøyrande 2×2- kovariansmatrisa til vektoren $\boldsymbol{b}_i$ har elementa

$$\Psi = \begin{bmatrix} \psi_1 & \psi_{12} \\ \psi_{21} & \psi_2 \end{bmatrix} \tag{2.7}$$

og er symmetrisk slik at $\psi_{12} = \psi_{21}$. Støyledda, $\boldsymbol{\epsilon}_i$, følger same fordeling som før. Dei ukjende parametrane i denne modellen er, uttrykt ved ein vektor θ med åtte element, $\theta = (\beta_0, \ldots, \beta_3, \sigma, \psi_1, \psi_{12}, \psi_2)$.

2.4.1 Kovarians i modell med to variable parametrar

I modell (2.6) på førre side er kovariansstrukturen endå meir kompleks. Vi antar at observasjonane har ei marginal kovariansmatrise på form

$$\text{Cov}(\boldsymbol{y}_i) = \boldsymbol{V}_i^{(2)} = \boldsymbol{Z}_i \boldsymbol{\Psi} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}$$

der matrisa $\boldsymbol{Z}_i$ no har dimensjon 4×2, og kovariansmatrisa til variable parametrar-vektoren, $\boldsymbol{\Psi}$, er uttrykt i likning (2.7).

I denne modellen er det varianselementet ψ_2 til det stokastiske stigningstalet som skil modellen frå å vere ein random intercept-modell. Dersom $\text{Var}\, b_{i1} = \psi_2 = 0$, vil også $\text{Cov}(b_{i0}, b_{bi1}) = \psi_{12} = 0$, og kun varianselementet $\psi_1 = \sigma_b^2$ står att.

Definisjonsområdet til elementet ψ_2 vil, som elementet σ_b^2, i ein marginal modell kunne ha negative verdiar. Det vil seie at det har definisjonsområdet $\psi_2 \in (-\infty, \infty)$ i ein marginal modell. I betinga modell derimot, vil ein anta at $\psi_2 > 0$, slik at elementet har definisjonsområde $\psi_2 \in (0, \infty)$, der 0 ligg på randa. Ein hypotesetest for elementet ψ_2 vil dermed også ha ulike formuleringar alt ettersom ein antar ein marginal eller hierarkisk modell.

Pinheiro og Bates (2000, side 58) antar at matrisa Ψ i ein blanda modell er positivt definitt. Dette svarar til ein betinga synsvinkel på varianselementa i modellen. Eg vil som dei gjere same antakinga i mine modellar for `Orthodont`, og ser dermed på modellane som betinga modellar.

2.5 Samanlikning av modellverdiar

Frå kapittel 2.2.1 har vi at variansen til responsvektoren består av både residualvarians og varians til variable parametrar. Ei kraftig minke i residualvarians vil tyde på at inklusjon av parameteren b_i er vellykka. Linja under viser formulering av random intercept-modell for `Orthodont` når metoden *maksimum likelihood* nyttast. Nærmare forklaring kan sjåast i tillegg A.

```
> fit1 <- lme(distance~age*Sex,random=~1|Subject,method="ML")
```

Dersom vi har ein betrakteleg mindre residualvarians σ^2 i den blanda modellen, vil vi anta at denne passar `Orthodont` betre. Tabell 2.1 viser resultat av analyser i R, samt verdiane frå den ordinære regresjonsmodellen i likning (1.6) på side 10. For dei faste parametrane har eg også oppgitt p-verdi.

	Modell (1.6)		Modell (2.4)	
	Estimat	p-verdi	Estimat	p-verdi
β_0	16,341	0	16,341	0
β_1	0,784	0	0,784	0
β_2	1,032	0,64	1,032	0,51
β_3	−0,305	0,13	−0,305	0,01
σ	2,257		1,369	
σ_b	0		1,741	
$\lvert\theta\rvert$	5		6	
logLik	−239,12		−214,32	
AIC	488,24		440,64	

Tabell 2.1: Parameterestimat med tilhøyrande p-verdi, standardavvik for residual- og individvarians, verdi av log-likelihood funksjon, og AIC-verdi for den ordinære regresjonsmodellen og random intercept-modellen for `Orthodont`-dataa.

Av tabellen ser vi at estimata av dei faste parametrane er omlag uendra etter inklusjon av eit stokastisk konstantledd. Men p-verdiane til estimata er derimot ulike. Parameteren β_3 for interaksjonsleddet har i random intercept-modellen

fått ein signifikant p-verdi. Vidare ser vi at residualvariansen er betrakteleg redusert i random intercept-modellen, og at estimatet av standardavviket til random intercept-parameteren har ein større verdi enn standardavviket til residualvariansen. Rada merka $|\theta|$ viser antall parametrar i dei respektive modellane.

Verdien av log-likelihood har auka ein god del ved inklusjon av b_i. Det er naturleg at ein modell med fleire parametrar vil ha betre likelihood av same grunnar som nemnt for «goodness-of-fit» observatorane i kapittel 1. Til slutt ser vi at random intercept-modellen har ein mindre AIC-verdi enn den ordinære modellen. For AIC-observatoren gjeld at jo mindre verdi jo betre. Av to modellar tilpassa same datasett, vil vi foretrekke modellen med lågast AIC-verdi. Eg vil omtale definisjonen av AIC nærare i kapittel 3.

Eg vil no undersøkje om ein modell med både stokastisk konstantledd og stokastisk stigningstal, uttrykt i likning (2.6) på side 22, gjev forbetring i residualvariansen, σ^2. Eg vil også undersøkje endringar i parameterestimat og observatorar. Modellen kan skrivast inn i R ved kun å oppdatere den forrige, `fit1`, til å ha ein varianskomponent også på stigningstalet i modellen.

```
> fit2 <- update(fit1, random=~age|Subject)
```

Resultat frå analyse i R er oppsummert i tabell 2.2.

	Estimat	p-verdi
β_0	16,341	0
β_1	0,784	0
β_2	1,032	0,52
β_3	−0,305	0,02
σ	1,310	
$\lvert\boldsymbol{b}_i\rvert$	2	
σ_b^2	2,135	
ψ_2	0,154	
ψ_{12}	−0.603	
$\lvert\theta\rvert$	8	
logLik	−213.903	
AIC	443,8	

Tabell 2.2: Verdiar av parametrar og observatorar, samt p-verdi for dei faste parametrane, for den blanda modellane med to variable parametrar tilpassa settet `Orthodont`.

Vi ser av tabellen at standardavviket til residualvariansen har fått eit noko mindre estimat i den utvida modellen enn i random intercept-modellen, og standardavviket av variansen til random intercept-leddet har fått eit større estimat. Estimatet av variansen til det stokastiske stigningstalet, b_{i1}, er lite. Estimata av dei faste parametrane med tilhøyrande p-verdiar er omlag uendra. Rada merka $|\boldsymbol{b}_i|$ gjev talet på variable parametrar i dei ulike modellane.

Vidare ser vi at verdien log-likelihood er omlag uendra, og at AIC-verdien ikkje er lågare. I seinare kapittel skal vi sjå om dette betyr at eit stokastisk stigningstal er overflødig for individa i `Orthodont`.

Tabell 2.3 oppsummerar verdiane vi har fått i analysene av dei tre siste modellane eg har definert for `Orthodont`. Alle desse modellane har same faste parametrar,

	Residualvarians σ^2	logLik	AIC
modell	2,000	239,1	488,2
modell	1,369	−214,319	440,6
modell	1,310	−213,903	443,8

Tabell 2.3: Verdiar av parametrar og observatorar for ordinær regresjonsmodell, random intercept-modell og modell med to variable parametrar tilpassa datasettet `Orthodont`.

altså er $\boldsymbol{X}_i$-matrisa den same for alle modellane i tabell 2.3. Av denne tabellen ser vi at residualvariansen minkar som følge av inklusjon av variable parametrar. Verdien av log-likelihood aukar (blir mindre negativ). Men AIC-verdien har fått eit minimum hos random intercept-modellen.

For å avgjere kva for modell som passar best for vekstkurvene til individa i `Orthodont` vil vi no utføre ein hypotesetest liknande den i likning (1.7) på side 12. Forskjellen no er at vi ikkje skal nytte F-observatoren i likning (1.8), men likelihood ratio-observatoren. Før eg går i gang med dette vil eg forklare litt om modellsamanlikning med AIC og hypotesetesting med likelihood ratio-observatoren.

There is no quality in this world that is not what it is merely by contrast. Nothing exists in itself.

Herman Melville

3 Modellsamanlikning

Eg vil no forklare litt generelt om samanlikning av modellar, med vekt på samanlikning av blanda modellar. Av definisjonen til ein lineær blanda modell i likning (2.1) på side 17, ser vi at overgangen frå ein ordinær lineær regresjonsmodell ikkje er rett fram ved å legge til ein parameter i modellen. I ein blanda modell antar vi også ein annan kovariansstruktur enn i ein ordinær modell, og dette gjer at inferens for blanda modellar ikkje følger den generelle teorien i alle punkt. Dermed er det vansklegare å finne den beste modellen for eit datasett, og det særleg dersom to modellar ser ut til å passe dataa like godt.

I dette kapittelet vil eg hovudsakleg sjå på residual, maksimum likelihood estimat og likelihood ratio observator for blanda modellar. Eg er særleg interessert i å avgjere kva for ein av modellane med stokastisk konstantledd og med både stokastisk konstantledd og stigningstal, som passar best for individa i `Orthodont`.

3.1 Analyse av residual

Når ein ynskjer å predikere ein respons ved uavhengige variablar gjer ein automatisk ei antaking om fordelinga til responsen.

For dei lineære regresjonsmodellane vi har sett på har vi antatt at responsen vår kjem frå ei normalfordeling. Vidare vil vi også anta at observasjonane vi har av responsen er uavhengige. Regresjonsmodellen på forma (1.2) på side 5 tek utgangspunkt i at vi kan gjere desse antakingane om observasjonane våre. For individa i `Orthodont` er antakinga om uavhengigheit kun rimeleg for observasjonar på ulike individ. Ein lineær regresjonsmodell, som modell (1.6) på side 10, vil dermed ikkje gje eit realistisk estimat av variansen til dataa. Dette avviket vil komme til syne dersom vi ser på *residuala* til modellen.

3.1.1 Residual

Ein måte å sjå kor godt ein modell passar til eit datasett er å studere residuala til modellen.

Residual er differansen mellom observerte og estimerte verdiar. Dersom ein har ein regresjonsmodell, som den på forma (1.1) på side 4, der vi har funne eit estimat av modellparametrane

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T,$$

vil residuala vere definerte som

$$y_i - \hat{\beta}_0 - \hat{\beta}_{1i} x_{1i} - \cdots - \hat{\beta}_{pi} x_{pi} = y_i - \hat{y}_i = \hat{\epsilon}_i.$$

Notasjonen $\hat{\epsilon}_i$ er ikkje konsekvent brukt for residual, ofte nyttast gjerne e_i eller r_i. Har ein n observasjonar av responsen, og n observasjonar av dei uavhengige forklaringsvariablane, vil ein ha n residual, $\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n$.

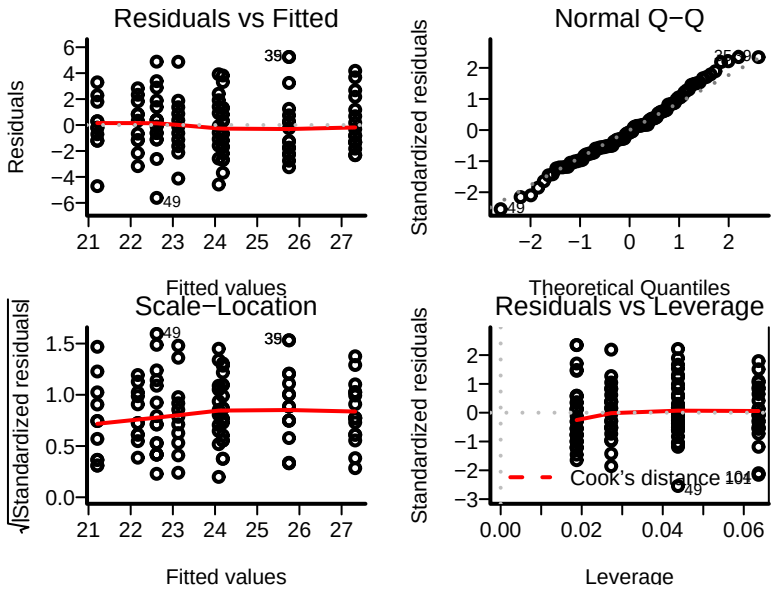
Dersom modellen vi har tilpassa er ein god realisasjon av dataa vil residuala $\hat{\epsilon}_i$ vere eit estimat av støyledda ϵ_i, og desse har vi antatt er uavhengige og normalfordelte med forventning 0 og varians σ^2 (Walpole *et al.* 1998, side 377).

Eit plott av residuala vil kunne avsløre om residuala $\hat{\epsilon}_i$ er gode realisasjonar av ϵ_i eller ikkje. Vi ynskjer at plottet skal vise tilfeldig spredte residual.

Eit plott av standariserte residual, kalla QQ-plott, er også informativt. Her blir dei standariserte residuala

$$\frac{\hat{\epsilon}_i}{\mathrm{Var}(\hat{\epsilon}_i)}, \quad i = 1, \ldots, n, \tag{3.1}$$

plotta mot standard normalfordelinga, $\mathcal{N}(0,1)$. I dette plottet vil dei standariserte residuala ligge tett langs linja dersom modellen vår er god.



Figur 3.1: Diagnostikkplott for modell (1.6) på side 10. Figuren viser frå øvst til venstre (og med klokka) eit plott av residual, eit QQ-plott, eit standardisert residualplott og eit plott «leverage».

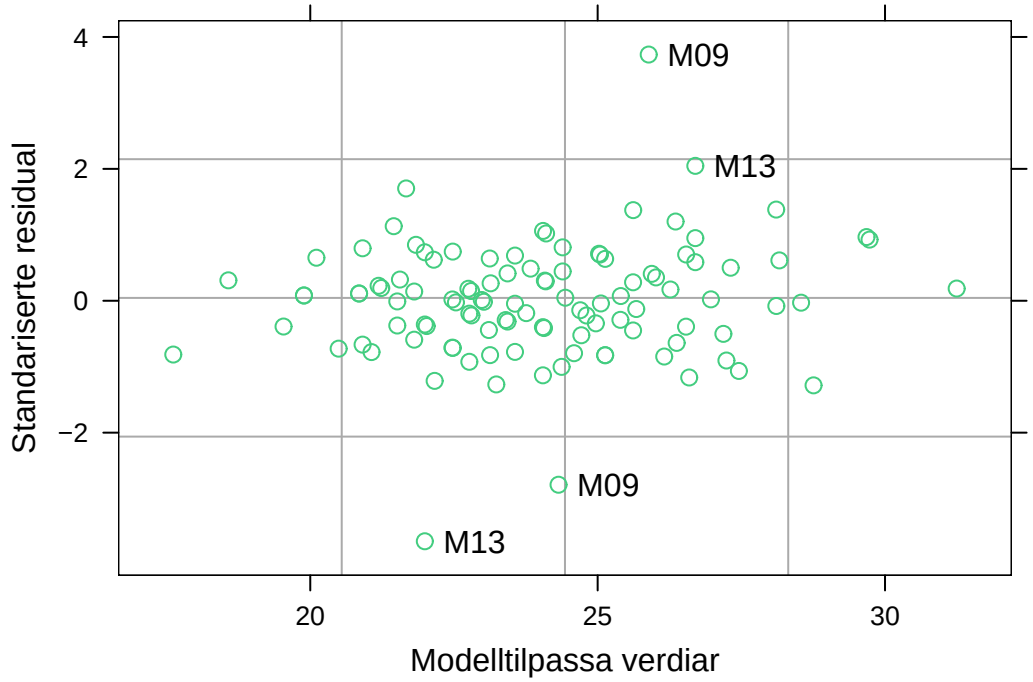
Både i residualplott og i QQ-plott vil vi kunne sjå eventuelle observasjonar som blir dårleg tilpassa av modellen. I R blir desse som regel identifiserte med nummer. Slike observasjonar kallast gjerne på norsk for *uteliggarar*, frå engelsk «outliers», og er observasjonar med uvanleg store residualverdiar. Ein *inflytelsesrik* observasjon er ein observasjon som har stor effekt på estimata av parameterane i modellen (Dobson 2002, side 89). Ein uteliggar treng ikkje å vere ein innflytelsesrik observasjon, og ein inflytelsesrik observasjon treng ikkje vere ein uteliggar. For å finne ut om ein uteliggar er ein inflytelsesrik observasjon, kan ein tilpasse modellen til dataa utan den uteliggande observasjonen.

Eit plott av residuala til den ordinære regresjonsmodellen for `Orthodont` i likning (1.6) på side 10 vises i figur 3.1. I residualplottet i figuren øvst til venstre blir observasjon 39 og 49 merka av som uteliggarar. Vi ser i figuren øvst til høgre, som er eit QQ-plott, at residuala ligg ganske godt langs linja, men vi kan sjå at det er grupperingar av residuala. Grupperingane kjem dessverre ikkje så godt fram i figur 3.1. Dersom ein lagar eit reint QQ-plott, slik eg har gjort i figur A.1 på side 96, kjem grupperingane betre til syne. Det er ikkje urimeleg at vi observerer slike grupperingar av residuala, sidan denne modellen ikkje tek høgde for korrelasjon mellom observasjonar på same individ. I QQ-plottet er observasjonane 35, 39 og 49 merka av som uteliggarar. Desse tilhøyrer respektivt individ `M09`, `M10` og `M13`.

Om vi går tilbake til figur 1.1 på side 6 ser vi at desse skil seg ut ved å ha høvevis ei ikkje-lineær kurve, svært høge verdiar av responsen, og ei veldig bratt kurve. I figuren nederst til høgre vises residual mot «leverage» for observasjonane. Eit «leverage» for ein observasjon, indikerar kor langt vekke observasjonen er frå dei $n - 1$ andre observasjonane. I figuren blir observasjon 49, 104 og 107 avmerka som observasjonar med stor «leverage». Dei to siste observasjonane høyrer respektivt til individ `F10` og individ `F11`. Observasjon 49 høyrer til individ `M13`, og han har den brattaste vekstkurva. Individa `F10` og `F11` er dei to vekstkurvene blandt jentene som utmerkar seg med å ha høvevis dei minste verdiane og dei største verdiane av responsvariabelen.

Ved å tilpasse ein blanda modell vil vi anta at vi i denne gjer rede for avhengigheit mellom observasjonar på same individ. Eg vil no sjå om plott av residuala til dei blanda modellane i likning (2.4) og (2.6) på side 22 viser uavhengige residual.

3.1.2 Residual til random intercept-modell



Figur 3.2: Plott av residual til modellen med éin random effect.

For modellen med éin variabel parameter, i likning (2.4), vil vi anta at observasjonane kan modellerast marginalt som

$$\hat{\boldsymbol{y}}_i = \boldsymbol{X}_i \hat{\boldsymbol{\beta}} \quad i = 1, \ldots, 27 \tag{3.2}$$

der $\hat{\boldsymbol{\beta}}$ er parameterestimata av den faste delen av modellen. Så langt er modellen lik den ordinære regresjonsmodellen (1.6) på side 10. For variansen til observasjonane derimot, antar vi at responsvektoren har ei kovariansmatrise på forma $\boldsymbol{V}_i^{(1)}$ i likning (2.5). Her er observasjonar på same individ korrelerte.

Figur 3.2 på førre side viser eit plott av residuala til random intercept-modellen. Kode som viser korleis eg laga desse plotta står i tillegg A.3. Figuren viser at residuala er tilsynelatande ukorrelerte. Vi ser at fire uteliggarar er avmerkte i plottet. To av dei tilhøyrer individ `M09`, og dei to andre tilhøyrer individ `M13`. Frå figur 1.1 på side 6, der vi kan sjå vekstkurvane til alle individa, ser vi at det er nettopp kurvene til individ `M09` og individ `M13` som skiljer seg frå resten ved å vere høvevis tilsynelatande ikkje-lineær og svært bratt.

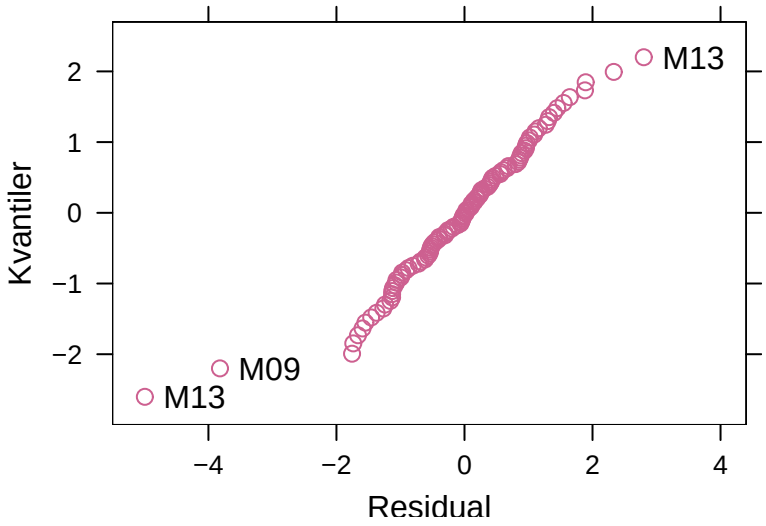
Av Pinheiro og Bates blir det påpeikt at observasjonar som har negativ differanse, altså der avstanden minkar med alder, truleg er feile eller er utsett for uvanleg stor målefeil. Pinheiro og Bates foreslår at ein kan gjere ei analyse av `Orthodont` utan desse observasjonane. Vi vil då ha eit ubalansert datasett, der individ som er registrert med negative differansar får desse observasjonane fjerna. Eg vil i denne oppgåva ikkje ta omsyn til dette.

For å sjekke om antakinga om normalfordeling av residuala til random intercept-modellen vil vi sjå på eit QQ-plott.

Vi ser frå QQ-plottet, figur 3.3 på neste side, at dei fleste residuala ligg godt på linja. For desse er antakinga om normalfordeling er rimeleg. Det ser ut til at grupperingane vi observerte i figur A.1 på side 96 er fjerna. Dei fire observasjonane som i figur 3.2 på førre side blei merkte som uteliggarar er dei same som vi her ser at ikkje ligg godt på linja. Dette betyr at dei standariserte residula stemmer godt med ei antaking om normalfordeling.

3.1.3 Residuala til modellen med to variable parametrar

Vi kan videre studere residuala når vi har tilpassa modellen med to variable parametrar til observasjonane i `Orthodont`.



Figur 3.3: Eit QQ-plott av standariserte residual til modell (2.4) på side 20 mot kvantiler frå standard normalfordelinga.

Vi vil anta at dersom den marginale modellen (2.6) passar godt til `Orthodont` så vil observasjonane vere henta frå ei fordeling på form

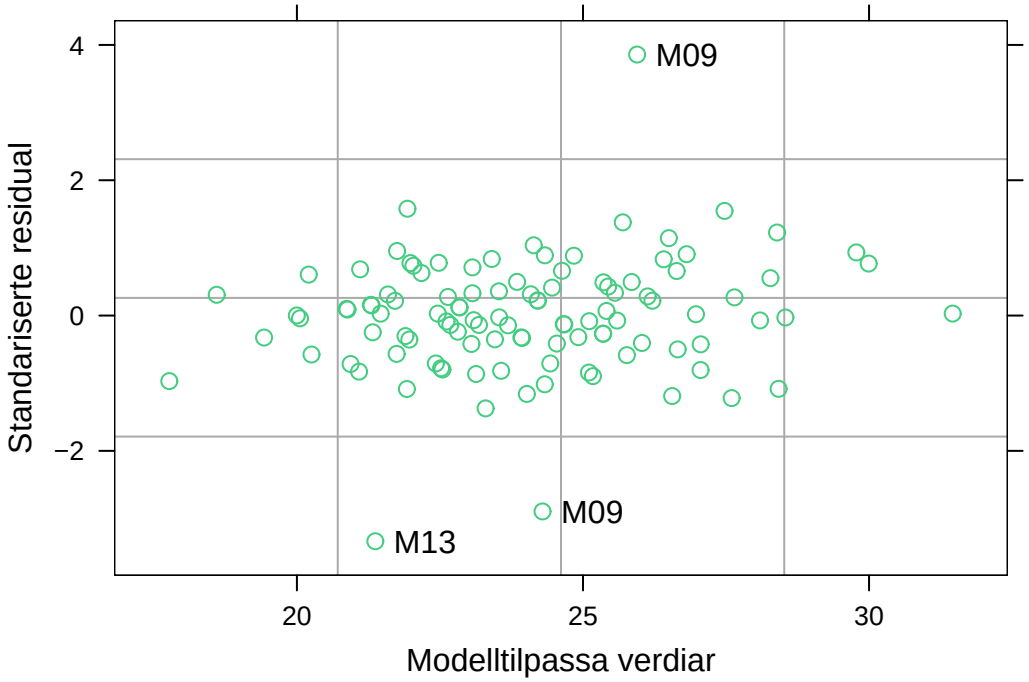
$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i^{(2)}), \quad i = 1, \ldots, 27. \tag{3.3}$$

Eit plott av residuala til modell med stokastisk konstantledd òg stigningstal er vist i figur 3.4 på neste side Vi ser at residula til denne modellen har ein mindre uteliggar. Individ `M13` har no kun ein observasjon som er dårlig tilpassa av modellen.

3.1.4 Residual som verktøy i modellsamanlikning

Utifrå desse to residualplotta ser både modell (2.4) og (2.6) på side 22 ut til å vere gode modellar for individa i `Orthodont`. Vi vil forvente at jo fleire parametrar jo betre tilpassa modell, så det vi må undersøkje no er om modellen med flest parametrar, modellen med to variable parametrar, er tilstrekkeleg betre enn modellen med éin random effect.

Plott av residual er altså ikkje alltid nok til å avgjere kva for modell som er best, men vil gje oss eit bilete på om vi er på rett spor. I tillegg vil vi sjå kor god normalitetsantakinga er i det bestemte tilfellet, og vi vil kunne identifisere ekstreme observasjonar. Men når ein, som i vårt tilfelle, har to modellar som begge ser ut til å passe godt, må vi nytte meir spesifikke metodar.



Figur 3.4: Plott av residual til modellen (2.6) på side 22.

Vi ynskjer å finne ut om det stokastiske stigningstalet i modellen med to variable parametrar for `Orthodont` gjev ei tilstrekkeleg forbetring i å forklare variasjonen i observasjonane. Ved ordinære regresjonsmodellar kan dette avgjerast ved å samanlikne AIC verdiar og p-verdiar ved å benytte `lme`, `summary` og `anova` i R. Men når vi ser på to blanda modellar er det ikkje rett fram å tolke verdiane til desse observatorane. Dette kjem av at vi utfører testar på varianskomponentar, ikkje faste parametrar som i ein ordinær regresjonsmodell, og det medfører eit avvik frå den generelle teorien.

3.2 Maksimum likelihood estimering i modellsamanlikning

Sjølv om ikkje parameteresimata til ein blanda modell gjev nokon direkte informasjon om ein modell passar godt til datasettet vi ser på, spelar det ei viktig rolle kva for fordeling ein brukar i estimeringa når ein skal samanlikne to modellar tilpassa same datasett. Eg vil i dette delkapittelet sjå hovudsakleg på metodane *maksimum likelihood* og *restricted maximum likelihood* med utgangspunkt i utledningane til Pinheiro og Bates og arbeid av Laird og Ware (1982). Det er desse to metodane vi

kan nytte i R-funksjonen `lme`. Det er ved bruk av desse metodane at vi vil merke ulikheita mellom marginale og hierarkiske modellar.

Vi skal først sjå på metoden maksimum likelihood, forkorta ML, som er velkjend for dei fleste statistikarar. Å estimere parametrar ved maksimum likelihood vil seie å finne modellparametrane som gjer at *likelihood-funksjonen* blir maksimert. Følgande definisjon forklarar ein likelihood-funksjon.

Definisjon 3.2.1: Likelihood-funksjon

Likelihood-funksjonen for n uavhengige data, representert ved ein vektor $\boldsymbol{y}$, som ein antar er observasjonar frå ei fordeling f med ukjente parametrar representert ved ein vektor $\boldsymbol{\theta}$, er definert som

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta}) \tag{3.4}$$

der L står for likelihood-funksjonen, og $\boldsymbol{y} = (y_1, \ldots, y_n)$ observasjonane.

Medan fordelingsfunksjonen f ser på datapunkta $y_1, \ldots, y_n$ der parametrane i vektoren $\boldsymbol{\theta}$ antas kjente, ser likelihood-funksjonen L på parametrane i vektoren $\boldsymbol{\theta}$ når datapunkta er kjente.

I normalfordelinga er det to ukjente parametrar, μ og σ^2. Dersom vi antek at fordelinga f er normalfordelinga og at observasjonane har kjent varians σ^2, men ukjent snitt μ_i, vil likelihood-funksjonen for n observasjonar, $y_1, \ldots, y_n$, ha forma

$$L(\boldsymbol{\theta}, \boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2} \tag{3.5}$$

3.2.1 Maksimum likelihood estimat i normalfordelinga

Eit maksimum likelihood estimat av parametervektoren θ, er eit estimat som reknast ut ved å maksimere likelihood-funksjonen i likning (3.4) når observasjonane $y_1, y_2, \ldots, y_n$ blir haldne konstante. Definisjonen til Casella og Berger (2002, side 316) forklarar godt korleis dette estimatet kan bereknast.

Definisjon 3.2.2: Maksimum likelihood estimator

La $\hat{\boldsymbol{\theta}}(y_i)$ vere ein verdi av $\boldsymbol{\theta}$ der likelihood-funksjonen $L(\boldsymbol{\theta}, y_i)$ har sitt maksimum når y_i bli halden fast. Ein maksimum likelihood estimator av parametervektoren $\boldsymbol{\theta}$ basert på n observasjonar, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, vil då vere $\hat{\boldsymbol{\theta}}(\boldsymbol{y}_i)$.

Den enklaste situasjonen for maksimum likelihood estimering, i normaltilfellet, er når ein kan anta at alle observasjonane er uavhengige og frå same fordeling, og har lik forventning og varians, $y_i \sim \mathcal{N}(\mu, \sigma^2)$, muligens ikkje begge ukjente. Eg vil no vise eit døme på ML-estimering i normalfordelinga.

Eksempel 3.2.3: Maksimum likelihood estimat i normalfordelinga

Gitt n observasjonar frå ei $\mathcal{N}(\mu, 1)$ fordeling, ynskjer vi å finne ML-estimat av forventninga μ. Likelihood-funksjonen L vil då, med utgangspunkt likning (3.5) på førre side, vere lik

$$\begin{aligned} L(\mu, y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-1/2(y_i-\mu)^2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^{n}(y_i-\mu)^2} \end{aligned}$$

Vi ser her at likelihood-funksjonen for μ, i dette tilfelle, har sitt maksimum der eksponenten har sitt minimum. Maksimum til ein parameter i ein funksjon finnast ved å derivere med omnsyn på parameteren, sette uttrykket lik null, og finne uttrykket for parameteren. Ved å derivere uttrykket i eksponenten med omsyn på μ og sette dette lik null, sit vi att med

$$\begin{aligned} &\frac{d}{d\mu}\left(\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \\ &= \sum_{i=1}^{n}(y_i-\mu) = 0, \end{aligned}$$

som betyr at ML-estimatet til forvetninga μ er $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$, som er det same som det det empiriske gjennomsnittet av observasjonane. Deretter bør ein sjekke at den dobbeltderiverte er mindre enn null i det aktuelle området for å bekrefte at maksimum likelihood estimatet er eit globalt maksimum, (Casella og Berger 2002, side 317).

I eksempel 3.2.3 er maksimum likelihood estimatet og *minste kvadraters* estimatet av μ det same. Dette er særskildt for normalfordelinga. Minste kvadraters estimat finnes ved å minimere

$$\sum_{i=1}^{n} \left(y_i - E(y_i)\right)^2 .$$

Sidan $E(y_i) = \mu_i$ i normalfordelinga ser vi at dette blir det same som å løyse likninga i eksempel 3.2.3.

Formelen (3.5) på side 34 forklarar likelihood funksjonen for enkle, uavhengig normalfordelte observasjonar. Det vil seie at kvar observasjon er gjort på ein variabel på eit individ eller objekt om gangen. Dersom vi i kvar observasjon måler verdiar av fleire variablar, har vi multivariate observasjonar.

3.2.2 Multivariat likelihood

Før den blanda modellen var kjend, blei repeterte målingar ofte modellert som vanlege multivariate målingar (Laird og Ware 1982), med snitt $\boldsymbol{\mu}_i$ med indeks i for individ, og kovariansmatrise $\boldsymbol{\Sigma}$ med dimensjon $n_i \times n_i$ der n_i er lik talet på målingar per individ. Den multivariate modellen hadde sin svakhet i kovariansmatrisa $\boldsymbol{\Sigma}$, sidan den vil vere vanskeleg å estimere ved n_i stor, og sidan den ikkje tek høgde for den individuelle variasjonen.

Når vi velger ein blanda modell, der vi forbetrar kovariansmatrisa enten ved å velge enten ei marginal eller hierarkisk fordeling for observasjonane, kan vi like fullt nytte formuleringa til likelihood-funksjonen for den multivariate modellen for å formulere likelihood-funksjonen til dei longitudinale dataa.

Definisjon 3.2.4

Likelihood-funksjonen for n observasjonar frå ei multivariat normalfordeling, $\boldsymbol{y}_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), i = 1, \ldots, n$, er av Walpole *et al.* (1998, side 168) definert som

$$L(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2} \cdot |\boldsymbol{\Sigma}|^{1/2}} e^{-(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)/2} \\ \frac{1}{(2\pi)^{np/2} \cdot |\boldsymbol{\Sigma}|^{n/2}} e^{\sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i)/2} \tag{3.6}$$

der p er talet på variablar.

Walpole *et al.* har brukt eit resultat frå lineær algebra som forenklar eksponenten i (3.6), men eg vil ikkje gå nærmare inn på dette her.

3.2.3 Likelihood-funksjon for blanda modellar

Pinheiro og Bates foreslår ei noko komplisert utledning av maksimum likelihood funksjonen for blanda modellar, men påpeikar at den er lett å implementere.

Dei definerar ei matrise $\mathbf{\Delta}$ som dei kallar *relativ presisjonsfaktor* (Pinheiro og Bates 2000, side 59). Dette er ei vilkårleg matrise som tilfredsstiller

$$\frac{\mathbf{\Psi}^{-1}}{1/\sigma^2} = \mathbf{\Delta}^T\mathbf{\Delta} \tag{3.7}$$

Videre lar Pinheiro og Bates vektoren $\boldsymbol{\theta}$ innehalde den ubegrensa mengda av parametrar som bestemmer matrisa $\mathbf{\Delta}$. Eg vil i dette avsnittet, for å unngå å blande inn tidlegare definisjonar med notasjon $\boldsymbol{\theta}$, kalle denne vektoren $\boldsymbol{\vartheta}$. Altså inneheld vektoren $\boldsymbol{\vartheta}$ den ubegrensa mengda parametrar som bestemmer matrisa $\mathbf{\Delta}$. Av likning (3.7) framgår det at vektoren $\boldsymbol{\vartheta}$ vil innehalde varianselementa til kovariansmatrisa $\mathbf{\Psi}$.

Dei ukjende parametrane i ein likelihood-funksjon for ein blanda modell på forma (2.1) på side 17, gitt ein datavektor $\boldsymbol{y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ av longitudinale observasjonar, vil dermed vere varianselementa i vektoren $\boldsymbol{\vartheta}$, parametrane i vektoren $\boldsymbol{\beta}$, og residualvariansen σ^2. Likelihood-funksjonen ein generell blanda modell, gitt data, vil dermed vere på forma

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2 \mid \boldsymbol{y}) = \prod_{i=1}^{n} f(\boldsymbol{y_i} \mid \boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2). \tag{3.8}$$

Vidare benyttar Pinheiro og Bates uavhengigheita mellom random effects-ledda $\boldsymbol{b}_i$ og støyledda $\boldsymbol{\epsilon}_i$ til å uttrykke den betinga fordelinga $f(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2)$ som

$$f(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2) = \int f(\boldsymbol{y_i}|\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2) f(\boldsymbol{b}_i|\boldsymbol{\vartheta}, \sigma^2) d\boldsymbol{b}_i,$$

der begge dei betinga fordelingane på høgre-sida er multivariat normale. I den betinga fordelinga til random effects-ledda benyttast så matrisa Δ til å uttrykke likelihooden $L(\boldsymbol{\beta}, \boldsymbol{\vartheta}|\boldsymbol{y})$ som ein funksjon av $\boldsymbol{\beta}, \boldsymbol{\vartheta}$ og σ^2. Maksimering av denne likelihooden vil teoretisk kunne gje oss maksimum likelihood estimat av $\boldsymbol{\beta}, \boldsymbol{\vartheta}$ og σ^2.

Reknemessig, har Pinheiro og Bates funne ut at det er lettare å maksimere likelihooden i likning (3.8) dersom ein først uttrykker funksjonen med omsyn på kun $\boldsymbol{\vartheta}$. Denne metoden kallar dei å *profilere* eller konsentrere likelihooden (Pinheiro og Bates 2000, side 64). Deretter finn dei betinga estimat, $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})$ og $\hat{\sigma}^2(\boldsymbol{\vartheta})$, av høvevis dei faste parametrane og residualvariansen.

Ei anna utledning, også foreslått av Pinheiro og Bates, er noko kortare og enklare å følge. Denne snur om på definisjonen (2.1) i kapittel 2 på side 16 ved å skrive

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^*, \quad i = 1, \ldots, n$$

der $\boldsymbol{\epsilon}_i^* = \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$. Desse nye «støyledda» vil ha fordeling $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T)$ der $\boldsymbol{\Psi}$ er kovariansmatrisa til random effects ledda i modellen ein ser på. Dette vil vere det same som å nytte ein marginal synsvinkel på fordelinga til responsen $\boldsymbol{y}_i$, uttrykt som

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i), \quad \text{der} \quad \boldsymbol{V}_i = \sigma^2\boldsymbol{I} + \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T.$$

Pinheiro og Bates uttrykker denne kovariansmatrisa som $\boldsymbol{V}_i = \sigma^2\boldsymbol{\Sigma}_i$, der $\boldsymbol{\Sigma}_i = \boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^T$.

Dermed vil ein likelihood-funksjon for observasjonar på n uavhengige individ, der vi som for `Orthodont` antar at vi har like mange observasjonar p for kvart individ, kunne uttrykkast som

$$L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2|\boldsymbol{Y}) = \frac{1}{(2\pi\sigma^2)^{np/2} \cdot |\boldsymbol{\Sigma}_i|^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right)$$

der $\boldsymbol{Y}$ representerer ein vektor bestående av alle målingane på alle individa.

Tar vi log av denne likelihood-funksjonen, får vi uttrykket

$$l(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2|\boldsymbol{Y}) = -\frac{np}{2}\log 2\pi\sigma^2 - \frac{n}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}).$$

Får så å finne maksimum likelihood estimat må vi derivere med omsyn på parameteren vi vil finne estimat for. Om vi først maksimerer likelihood-funksjonen med omsyn på vektoren $\boldsymbol{\vartheta}$, får vi profilerte maksimum likelihood estimat av høvevis $\boldsymbol{\beta}$ og σ^2 ved å derivere den *profilerte* log-likelihooden. Derivasjon av desse betinga likelihoodane gir

$$\frac{\partial l(\boldsymbol{\vartheta})}{\partial\boldsymbol{\beta}} = -\frac{2}{2\sigma^2}\sum_{i=1}^{n}\left(-\boldsymbol{X}_i^T\right)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

og

$$\frac{\partial l(\boldsymbol{\vartheta})}{\partial\sigma^2} = -\frac{np}{2}\frac{1}{\sigma^2} + \frac{1}{4\sigma^4}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

Ved å sette desse uttrykka lik null, har Pinheiro og Bates funne uttrykka

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i, \tag{3.9}$$

og

$$\hat{\sigma}^2(\boldsymbol{\vartheta}) = \frac{1}{np} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) \right)^T \boldsymbol{\Sigma}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) \right) \tag{3.10}$$

som maksimum likelihood estimat av dei faste parametrane og residualvariansen.

3.2.4 Restricted maksimum likelihood

Ei utledning av parameteresimat med utgangspunkt i marginale fordelingar vil, som vi såg av synsvinkelen som vart nytta i likning (3.2.3), gje maksimum likelihood estimat av parametrane. Dersom ein derimot vel å nytte ei betinga fordeling for observasjonane, vil ein få restricted, på norsk avgrensa, maksimum likelihood estimat, forkorta REML estimat, av parametrane (Laird og Ware 1982).

Som vi såg i kapittel 2, då vi skulle tilpasse ein blanda modell i R, kunne både maksimum likelihood, forkorta ML, og restricted maksimum likelihood nyttast. I følge Pinheiro og Bates blir varianskomponentane i ein blanda modell ofte underestimert ved ML. Metoden ML, lar dei faste parametrane vere *nuisance* parametrar, det vil seie «forstyrrande parametrar», under estimeringa av varianselementa, og vil dermed ikkje gje forventningsrette estimat. Fitzmaurice *et al.* (2004, side 100) forklarar at forventninga av ML estimatet av σ^2, $\hat{\sigma^2}$, vil vere

$$E(\hat{\sigma^2}) = \left(\frac{m-p}{m} \right) \sigma^2 \tag{3.11}$$

der m er totalt talet på observasjonar i datasettet, og p talet på parametrar i modellen. For å få eit forventningsrett estimat av σ^2 ved denne metoden, må ein multiplisere med $m/(m-p)$. Små datasett vil vere større påverka av denne forventningsskjeivheita enn store datasett. Vi ser av likninga at når differansen $m-p$ er liten i høve til m, vil estimatet vere tilnærma forventningsrett.

Metoden REML gjer opp for skjeivheita i teljaren på høgresida i likning (3.11) ved å ta omsyn til at ein ikkje kjenner verdien av dei faste parametrane. I REML

estimering blir dei faste parametrane i vektoren $\boldsymbol{\beta}$ eliminert frå likelihooden i likning (3.8) på side 37, slik at den består kun av dei ukjende varianselementa.

Pinheiro og Bates nyttar Laird og Ware (1982) sin definisjon av ein restricted likelihood funksjon, som er på forma

$$L_R(\boldsymbol{\vartheta}, \sigma^2 \mid \boldsymbol{y}) = \int L(\boldsymbol{\beta}, \boldsymbol{\vartheta}, \sigma^2 \mid \boldsymbol{y}) d\boldsymbol{\beta}. \tag{3.12}$$

Her blir dei faste parametrane integrert ut av likelihooden. Navnet *restricted* kjem altså av at ein maksimerer likelihooden over ei *avgrensa* parametermengde (Corbeil og Searle 1976).

Problemet med å nytte restricted maksimum likelihood er at den er sårbar for endringar i matrisa $\boldsymbol{X}_i$ i tilfeller der denne består av kontrastar som representerer faktorar (Pinheiro og Bates 2000, side 76). Ei følge av dette er at modellar med ulik struktur i dei faste komponentane ikkje kan samaliknast på grunnlag av sine REML estimat. Ein kan dermed ikkje utføre ein likelihood ratio-test, som den beskrive i følgande delkapittel, på to modellar berekna ved REML dersom fixed effect-delen til modellane ikkje er lik.

3.3 Likelihood ratio observator for blanda modellar

Å benytte log-likelihood ratio som observator i modellsamanlikning er vanleg i tilfeller der den eine modellen representerer eit spesialtilfelle av ein annan der ein eller fleire av parametrane er sett lik null. Pinheiro og Bates benyttar uttrykket *nøsta*, på engelsk «nested», om to slike modellar. Ein seier då at den spesifikke modellen, modellen med færrast parametrar, er nøsta i ein meir generelle med fleire parametrar. Dette er tilfellet for modellane med høvevis éin og to variable parametrar.

3.3.1 Utleiing av likelihood ratio

La $L(\boldsymbol{\theta}_0|\boldsymbol{y})$ vere likelihood til ein modell definert under H_0 for eit bestemt datasett $\boldsymbol{y}$, og $L(\boldsymbol{\theta}_1|\boldsymbol{y})$ likelihood til ein meir generell modell definert under H_1. Dei

respektive likelihoodane er då definert som (Casella og Berger 2002, side 488)

$$\begin{aligned} L(\boldsymbol{\theta}_0|\boldsymbol{y}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\Theta}_0) \\ L(\boldsymbol{\theta}_1|\boldsymbol{y}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\Theta}_1). \end{aligned} \tag{3.13}$$

der L står for likelihoodfunksjonen definert i likning (3.4) på side 34, vektorane $\boldsymbol{\theta}_0$ og $\boldsymbol{\theta}_1$ står for parameterrommet til modellen under H_0 og H_1 respektivt, og $\boldsymbol{Y}$ er datasettet med alle observasjonane som modellane skal tilpassast. Sidan den meir generelle modellen har fleire parametrar enn den spesifikke vil alltid likelihooden til modellen under H_1 vere større eller lik likelihooden til modellen under H_0.

Definisjon 3.3.1: Likelihood ratio

Likelihood ratioen er definert som

$$\boldsymbol{\lambda} = \frac{L(\boldsymbol{\theta}_0|\boldsymbol{Y})}{L(\boldsymbol{\theta}_1|\boldsymbol{Y})} \geq 1 \tag{3.14}$$

Lar vi l_0 vere log-likelihood til modellen under H_0, og l_1 vere log-likelihood til modellen under H_1 kan ein definere observatoren

$$\boldsymbol{\Lambda} = -2\log\boldsymbol{\lambda} = -2(l_0 - l_1), \tag{3.15}$$

som kallast log-likelihood ratio, eller berre *likelihood ratio*.

Likelihood ratioen, gjerne forkorta til LR, har ei tilnærma kji-kvadratfordeling på forma

$$\boldsymbol{\Lambda} \sim \chi^2(p - q), \tag{3.16}$$

der p er talet på parametrar i modellen under H_1 og q talet på parametrar til modell under H_1 slik at $p > q$, (Casella og Berger 2002).

Definisjonsområdet til likelihood ratioen, $\boldsymbol{\Lambda}$, ligg mellom null og uendeleg. Dersom ein observerar ein verdi $\hat{\boldsymbol{\Lambda}} > c$, der $c \in [0, \infty)$ er kalla kritisk verdi, vil vi forkaste H_0. Dette kan forklarast ved at store verdiar av $\boldsymbol{\Lambda}$ betyr små verdiar av $\boldsymbol{\lambda}$ som igjen betyr at likelihooden til den generelle modellen ikkje er mykje større enn likelihooden til den spesifikke modellen. Dermed gir det ikkje så stor forbetring å gå frå modellen under H_0 til modellen under H_1 og vi vil dermed behalde den spesifikke modellen, altså den enklaste, når vi observerar små verdiar av $\hat{\boldsymbol{\Lambda}}$, altså

$\hat{\mathbf{\Lambda}} \leq c$. Det er vanleg å nytte notasjonen aksteptområde om verdiar i mengda $[0, c]$, og forkastningsområde, eller kritisk område, om verdiar i mengda $\langle c, \infty)$.

Den kritiske verdien c bestemmes ut i frå fordelinga til likelihood ratioen. Ved først å velge eit signifikansnivå α, altså kor mange observasjonar i det kritiske område ein vil tåle utan å forkaste H_0, finn vi den kritiske verdien $c = \chi^2_p(\alpha)$, der p er talet på frihetsgrader og α er signifikansnivået. Det er vanleg å benytte $\alpha = 0.05$, og det vil eg gjere her.

Ein kan også nytte metoden REML i estimeringa av likelihood ratioen så lenge modellane ein testar har same fixed effects-del. I så fall vil ein kalle testobservatoren for ein *restricted* likelihood ratio, eller berre RLR.

Før eg går vidare til å teste dei to blanda modellane i likning (2.4) og (2.6) på side 22, vil eg utføre ein likelihood ratio-test mellom modellen utan random effects ledd i likning (1.6) på side 10 og modellen med eit random effects-ledd, i likning (2.4) på side 20. Av tabell 2.1 på side 24 ser vi at den ordinære modellen har fem parametrar, medan random intercept-modellen har seks. Vi forventar då ved likning (3.16) på førre side at LR-observatoren har ei χ^2_1-fordeling. Eg vil benytte notasjonen $\hat{\mathbf{\Lambda}}$ om observerte verdiar av likelihood ratioen $\mathbf{\Lambda}$.

Eksempel 3.3.2: Hypotesetest for random intercept-modell

Eg ynskjer i dette døme å nytte likelihood ratio-observatoren til avgjere om variansen σ^2_b til det stokastiske konstantleddet i random intercept-modellen er statistisk signinfikant. Hypotesa vi testar er

$$H_0 : \sigma^2_b = 0 \quad \text{mot} \quad H_1 : \sigma^2_b > 0.$$

I følge likning (3.16) på førre side vil likelihood ratioen til desse modellane ha ei χ^2_1 fordeling. Vi reknar ut verdien av likelihood ratioen ved formelen (3.15) i R manuelt ved

```
> l0 <- logLik(fit0)
> l1 <- logLik(fit1)
> -2*(l0 - l1)


49.6027
```

Vi får verdien $\hat{\mathbf{\Lambda}} = 49.6$ av testobservatoren, og denne verdien er statistisk signifikant i ei χ^2_1-fordeling. Vi vil forkaste hypotesa $H_0 : \sigma^2_b = 0$, og behalde den alternative hypotesa som seier at den variable parameteren b_i er ulik null.

Eg vil no nytte teorien om likelihood ratio som modellobservator for å samanlikne dei to blanda modellane eg har definert i kapittel 2 for `Orthodont`-dataa.

3.3.2 Likelihood ratio test for Orthodont dataa

For individa i `Orthodont` kan vi formulere ein test mellom random intercept-modellen i likning (2.4), og modellen med både stokastisk konstantledd og stigningstal i likning (2.6), der random intercept-modellen er nøsta i modellen med to variable parametrar, som

$$H_0 : \psi_2 = 0 \quad \text{mot} \quad H_1 : \psi_2 > 0. \tag{3.17}$$

Sidan $\psi_2 = 0$, som betyr at det variable stigningstalet er lik null, medfører at $\psi_{12} = \psi_{21} = 0$, skriv vi ikkje desse med i hypoteseformuleringa. Dette er ei ei-sidig hypotese.

Som nemnt i kapittel 2, vil ein marginal modell tillate negativ korrelasjon, medan ein betinga modell vil anta at matrisa $\mathbf{\Psi}$ er positivt definitt, noko som medfører at ein antar at alle diagonalelementa er større enn null. Å teste ei hypotese på form (3.17) vil medføre at ein under H_0 definerar varianselementet ψ_2 på grensa av definisjonsområdet sitt. Testobservatoren vi nyttar til å inferere om ψ_2 i hypotesa vil bli påverka av dette. Fordelinga til observatoren vil ikkje lenger følge den generelle teorien der ein subtraherar talet på parametrar i den generelle og den spesifikke modellen.

Eg vil først sjå kva konklusjonen blir når vi testar hypotesa i likning (3.17) for dei to blanda modellane eg har definert for `Orthodont`.

Eksempel 3.3.3

Vektoren $\boldsymbol{\theta}_0$, med dei ukjente parametrane til random intercept-modellen i likning (2.4), inneheld seks element, $\boldsymbol{\theta}_0 = (\beta_0, \ldots, \beta_3, \sigma_b, \sigma)$. Vektoren $\boldsymbol{\theta}_1$, med dei ukjente parametrane til to variable parametrar-modellen i likning (2.6), inneheld åtte element, $\boldsymbol{\theta}_1 = (\beta_0, \ldots, \beta_3, \psi_1, \psi_{12}, \psi_2, \sigma)$. Vi forventar dermed ved generell teori at likelihood ratio observatoren til datasettet `Orthodont` er fordelt som

$$\mathbf{\Lambda} \sim \chi^2_{8-6} = \chi^2_2. \tag{3.18}$$

Med funksjonen `anova` i R kan vi utføre ein ordinær likelihood ratio test.

```
> anova(fit1,fit2)
```

```
     Model  df    AIC      BIC   logLik   Test L.Ratio p-value
fit1     1  6 440.639 456.732 -214.319
fit2     2  8 443.806 465.263 -213.903 1 vs 2 0.833107  0.659
```

Vi ser at `anova` har gitt oss ein observert likelihood ratio på $\hat{\mathbf{\Lambda}} = 0.833$ som ikkje er signifikant i χ^2_2-tabellen på nivå 0,05. Vi kan ikkje forkaste H_0, og vi har fått ein p-verdi på 0,659.

Sidan og ein testobservator som føl fordelinga i likning (3.16) på side 41 vil ha eit større definisjonsområde enn den vi ynskjer å nytte til å teste hypotesa $H_0 : \psi_2 = 0$, vil ein p-verdi utrekna som i eksempel 3.3.3 vere *konservativ*. Med konservativ meiner eg at den mest truleg er større enn ein p-verdi rekna ut ved den sanne fordelinga til testobservatoren.

Problemet med å bestemme fordelinga til likelihood ratioen-observatoren og underestimering av parametrar ved maksimum likelihood metoden, gjer at modellsamanlikning for blanda modellar er meir omstendeleg enn for ordinære regresjonsmodellar. Eg vil prøve å simulere fordelinga til likelihood ratio-observatoren for dei to blanda modellane for `Orthodont`, for å undersøkje korleis ein testobservator blir påverka av å ha element på randa av definisjonsområdet sitt. Men først vil eg sjå litt på korleis observatoren AIC er definert.

3.4 Modellsamanlikning med AIC-verdi

Observatoren AIC, forkorting for «Akaike's information criterion», vart introdusert av Akaike i 1971. Noko forkorta er kriteriet eit forsøk på å måle godheita av den *predikative fordelinga* til ein modell som er tilpassa eit gitt datasett ved forventa log-likelihood. Den predikative fordelinga h til ein modell er fordelinga til ein framtidig observasjon gitt noværande observasjonar (Ishiguro *et al.* 1997).

Akaike sitt kriterium blei raskt populært som eit informasjonskriterium ved val av modell, grunna at det hadde eit generelt bruksområde og ei enkel formulering. Det er blitt nytta mellom anna innan geofysikk, hydrologi, psykometri og medisin. Men nettopp på grunn av sitt ukonvensjonelle bruksområde, har for AIC observatoren ikkje blitt fullt godkjent av røynde statistikarar.

Ved samanlikning av to modellar tilpassa same datasett, vil modellen med minst AIC verdi vere den beste i følge Akaike sin teori. I kapittel 2 såg vi i tabell 2.3 på side 26 at random intercept-modellen for `Orthodont`-dataa fekk minst AIC verdi.

I følge teorien til Akaike høver denne då best til å modellere den ortopediske avstanden for individa. Eg vil avrunde kapittelet om modellsamanlikning for blanda modellar med å sjå litt nærare på definisjonen av Akaike sitt kriterium.

La $\boldsymbol{y}$ vere eit datasett vi ynskjer å tilpasse ein modell til. Modellen har parametervektor $\boldsymbol{\theta}$, og vi definerar fordelinga til modellen som $f(\boldsymbol{y} \mid \boldsymbol{\theta})$. Estimatet av denne parametervektoren, basert på datasettet, er $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$. La så $\boldsymbol{y}_{n+1}$ stå for éin eller fleire framtidige observasjonar. Den predikative fordelinga til observasjonane $\boldsymbol{y}_{n+1}$ gitt dei noværande observasjonane $\boldsymbol{y}$ kan med dette definerast som

$$h(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}) = g(\boldsymbol{y}_{n+1} \mid \hat{\boldsymbol{\theta}}(\boldsymbol{y})) \tag{3.19}$$

der, dersom $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ er ML estimatet av parametervektoren $\boldsymbol{\theta}$, vil

$$h(\boldsymbol{y} \mid \boldsymbol{y}) = f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}(\boldsymbol{y})) = L(\boldsymbol{\theta}, \boldsymbol{y}) \tag{3.20}$$

vere maksimum likelihood estimatet til modellen (Ishiguro *et al.* 1997).

Godheita til den predikative fordelinga kan estimerast ved

$$E_{\boldsymbol{y}_{n+1}} \log h(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}), \tag{3.21}$$

men det krev at vi har kjennskap til den sanne fordelinga dei framtidige observasjonane $\boldsymbol{y}_{n+1}$ er henta frå, noko vi ikkje har.

Eit estimat av forventninga i (3.21) er den forventa log-likelihooden i likning (3.20). Skjeivheita til denne likelihooden, som oppstår fordi vi nyttar dataa $\boldsymbol{y}$ både til å estimere likelihooden og til å estimere parametrane, er definert ved

$$C = E_{\boldsymbol{y}} \left[\log h(\boldsymbol{y} \mid \boldsymbol{y}) - E_{\boldsymbol{y}}(\boldsymbol{y}_{n+1} \mid \boldsymbol{y}) \right]. \tag{3.22}$$

Det forventningsrette estimatet av forventninga i likning (3.21), vil dermed vere

$$\log f(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}(\boldsymbol{y})) - C = l(\boldsymbol{y} \mid \hat{\boldsymbol{\theta}}(\boldsymbol{y})) - C \tag{3.23}$$

Akaike viste at eit godt estimat av skjeivheita C kunne gjerast ved dimensjonen av vektoren $\boldsymbol{\theta}$, det vil seie $C = |\boldsymbol{\theta}|$. Slik kom han fram til informasjonskriteriet sitt som i symbol uttrykkast som

$$\text{AIC} = -2 \cdot l(\boldsymbol{\theta}, \boldsymbol{y}) + 2 \cdot |\boldsymbol{\theta}|. \tag{3.24}$$

I ord kan Akaike sitt informasjonskriterium formulerast som:

Definisjon 3.4.1: AIC-verdi

$$\begin{aligned}\text{AIC} = &-2 \times (\text{maksimum log likelihood til modell}) \\ &+2 \times (\text{talet på parametrar i modellen}).\end{aligned}$$

Kriteriet BIC, også kalla SBC for «Schwarz's Bayesian Criterion», er definert på same måte med unntak av det siste leddet der $2 \times (\text{talet på parametrar})$ er erstatta med log(talet på observasjonar) multiplisert med antall parametrar, det vil seie $\log \dim(\boldsymbol{y}) \times$ (talet på parametrar i modellen). For både AIC og BIC gjeld jo mindre verdi jo betre modell.

Det er nokre ting som trekk ned for AIC-verdi som observator i modellsamanlikning. Den første er at estimering skjer ved maksimum likelihood. Denne antakinga gjer at AIC berekna ved ein avgrensa likelihood, REML, kun er samanliknbar for to modellar med like faste parametrar. Ein annan ting som trekk ned for AIC i modellsamanlikning er at den er best i tilfeller der talet på observasjonar er stort. Ein AIC-verdi er dermed ikkje ein så god observator i samanlikninga av to modellar, dersom ein har ei avgrensa mengde data ein skal dra konklusjonar frå.

The generation of random numbers is too important to be left to chance.

Robert R. Coveyou

4

Simulering av likelihood ratio for Orthodont dataa

Eg ynskjer i dette kapittelet å undersøke om p-verdien som vart utrekna i eksempel 3.3.3 på side 44 er konservativ.

Det er i hovudsak to situasjonar som er ugunstige for ein likelihood ratio som testobservator for to blanda modellar. Den eine er at fordelinga likelihood ratioen blir samanlikna med, som i følge generell teori er fordelinga i likning (3.18) på side 43, ikkje baserar seg på ein observator som har verdien null på randa av definisjonsområdet sitt. Den andre er at metoden ML som nyttast til å rekne ut observatoren, underestimerar varianselementa i modellane, og dermed også verdien av testobservatoren. Begge desse situasjonane er medverkande til at ein p-verdi som utreknast på generell måte, slik eg gjorde i eksempel 3.3.3, er konservativ. Eg ynskjer å angripe desse problema, og sjå om eg kan finne ein ikkje-konservativ p-verdi for hypotesa i likning (3.17) på side 43.

Pinheiro og Bates (2000) har utført fleire simuleringar av likelihood ratio. Dei har nytta ein R-funksjon kalla `simulate.lme` som dei har utvikla for denne typen problemstillingar. Denne funksjonen fantes ikkje i den versjonen av R som eg nyttar. Det er likevel av interesse for meg å utvikle eit eige program for simulering

av likelihood ratioen, for å ha kontroll over kva for antakingar som leggast til grunn.

4.1 Bakgrunn

Observatoren likelihood ratio er ein *suffisient* observator for parameteren ψ_2 i hypotesa i likning (3.17). Suffisiensprinsippet for ein observator seier at dersom $\mathbf{\Lambda}$ er ein suffisient observator for ψ_2, så vil all inferens omkring ψ_2 avhenge av datasettet vårt $\boldsymbol{Y}$ kun gjennom observatoren $\mathbf{\Lambda}$ (Casella og Berger 2002, side 272). Når ein testar to hypoteser for å avgjere til dømes kva for ein av to modellar som er passar best til observasjonar i eit datasett, vil p-verdien reknast ut etter formelen $\text{p} = P(\mathbf{\Lambda}_i \geq \hat{\mathbf{\Lambda}} \mid H_0 \text{ er sann})$, der $\hat{\mathbf{\Lambda}}$ er verdien av testobservatoren vi har berekna på grunnlag av dataa våre, og $\mathbf{\Lambda}_i$ er generelle observasjonar frå ei χ^2_2-fordeling.

Som nemnt er problemet for testobservatoren $\hat{\mathbf{\Lambda}}$ i hypotesetesten for elementet ψ_2 at den mest truleg ikkje kjem frå ei χ^2_{p-q}-fordeling der $p - q$ er differansen i varianselement for modellane. Dermed vil ein p-verdi utrekna med ei slik fordeling til grunn ikkje gje oss den reelle p-verdien.

Problemet med fordelinga ein samanliknar testobservatoren i ei hypotese for to blanda modellar med, kan angripast ved å simulere likelihood ratioen utifrå data estimert ved modellen under H_0. Problemet knytta til underestimerte varianselement, vil innlemmast i denne metoden. Dersom varianselementa i dei to blanda modellane for `Orthodont` er blitt underestimerte, vil dette påverke utgangspunktet for simuleringa mi. Sidan vi, for dei to modellane, har lik fast effekt-del, vil ein ifølge Pinheiro og Bates kunne samalikne REML estimat for modellane, og også estimere ein restricted likelihood ratio. Eg vil derfor prøve å simulere likelihood ratioen ved å nytte både ML og REML i estimeringa av modellparametrane. Det vil vere interessant å sjå om ein p-verdi berekna ved REML er mindre konservativ enn ein p-verdi estimert ved ML. Tabell 4.1 på neste side viser at parameter-estimata for dei to blanda modellane, er større når ein nyttar metoden REML i estimeringa.

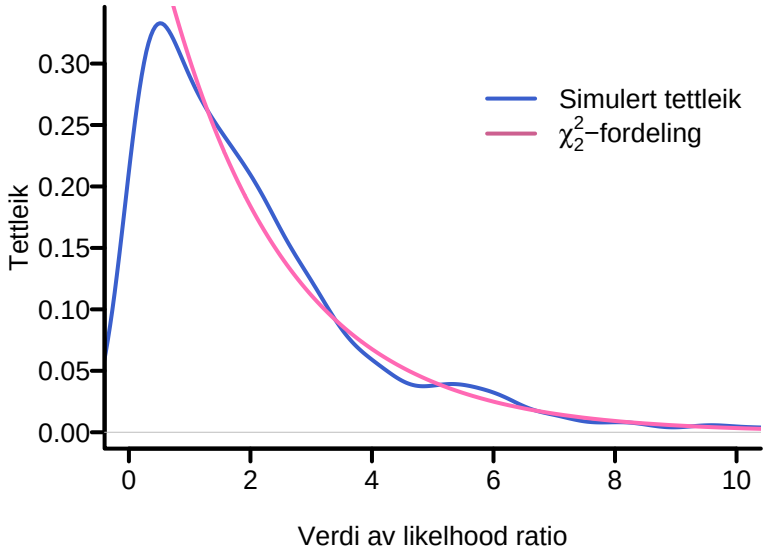
Tabellen viser at varianselementa er estimert til å ha høgare verdiar ved metoden REML enn ved ML.

For å gje eit døme på at den generelle teorien held for dei to ordinære regresjonsmodellane for `Orthodont` i likning (1.3) og (1.6), vil eg simulere ein likelihood ratio som testar hypotesa

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{mot} \quad H_1 : \quad \text{minst éin ulik null.}$$

	Modell (2.4) Estimat	Modell (2.6) Estimat
β_0	16,341	16,341
β_1	0,784	0,784
β_2	1,032	1,032
β_3	−0,305	−0,305
σ	1,386	1,310
σ_b	1,816	2,406
ψ_2	0	0,180
logLik	−216,879	−216,291
AIC	445,76	448,58

Tabell 4.1: Parameterestimat, verdi av log-likelihood funksjon, og AIC-verdi for random intercept modell og to random effect-modell tilpassa `Orthodont` ved metoden REML.



Figur 4.1: Simulert likelihood ratio for dei to ordinære modellane for `Orthodont` i kapittel 1. Figuren syner at den simulerte tettleiken og tettleiken til χ^2_2-fordelinga, er nokså like.

Ein likelihood ratio for desse modellane vil i følge generell teori ha ei χ^2_2-fordeling. Frå kapittel 1 hugsar eg at modellen under H_0 vart forkasta, og at p-verdien blei omlag lik null. Eg forventar ved ei simulering av likelihood ratioen med antaking om at modell (1.3) på side 8 er sann, at svært få av dei simulerte verdiane er større enn den observerte likelihood ratioen, som er $\hat{\mathbf{\Lambda}} = 27{,}1$.

Eg simulerer 1000 ratioar utrekna ved formel (3.15) på side 41. Koden eg nytta står i tillegg B. Resultatet av den simulerte fordelinga er vist i figur 4.1 på førre side, og som vi ser høver denne godt til tetthetskurva frå χ^2_2-fordelinga. Faktisk er ingen av dei simulerte verdiane større enn den observerte, og dette høver godt til ein p-verdien som ved generell teori var tilnærma lik null.

Eg vil no anta at random intercept-modellen er den beste for `Orthodont`-dataa, og vil simulere nye datasett ved dei predikerte modellverdiane til denne som står i tabell 2.1. Deretter vil eg simulere likelihood ratioen for hypotesa i likning (3.17), og samanlikne dei simulerte verdiane med verdiar frå ei χ^2_2-fordeling. Dersom den generelle teorien held mål, vil ein forvente at ein p-verdi på form

$$p = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Lambda}_i \geq \hat{\mathbf{\Lambda}}. \tag{4.1}$$

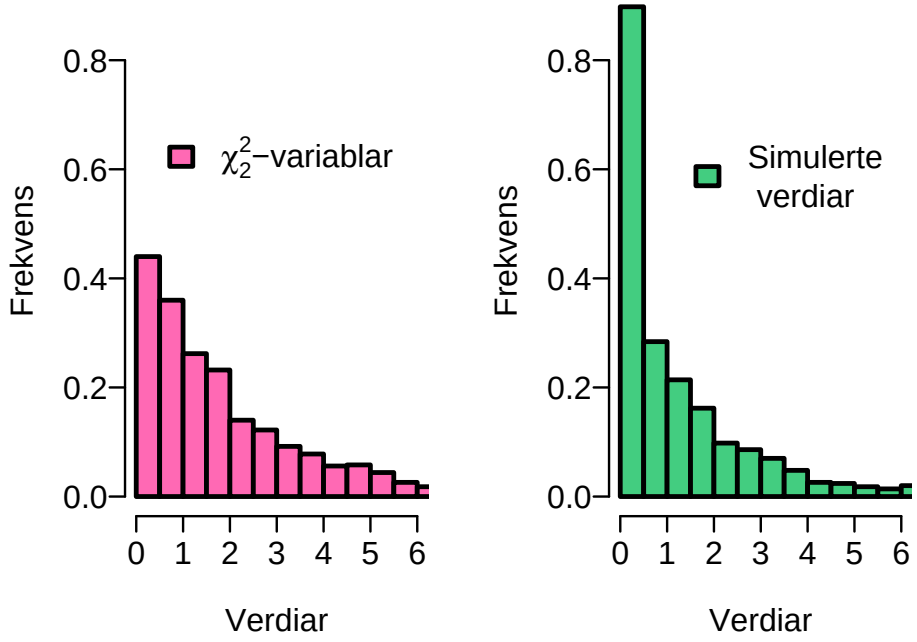
der n er antall simulerte verdiar, vil vere tilnærma 0,65. Programmet mitt med tilhøyrande kommentarar finnast i tillegg B.

4.2 Analyse av simulerte data

Eg ynskjer no å undersøke dei simulerte dataa. Eg forventar at dersom den generelle teorien held, så vil omlag 659 av 1000 ratioar vere større eller lik den observerte ratioen, $\hat{\mathbf{\Lambda}}$, som i eksempel 3.3.3 vart berekna til å vere $\hat{\mathbf{\Lambda}} = 0{,}833$.

Figur 4.2 på neste side viser frekvensar av verdiar av 1000 variablar frå ei χ^2_2-fordeling til venstre, og frekvensar av dei 1000 simulerte verdiane av likelihood ratioen eg fekk ved programmet mitt i tillegg B. I histogrammet har eg kutta den vertikale aksen for verdiar større enn seks. Sidan kun 29 av dei 1000 simulerte verdiane var større enn seks, mistar vi ikkje viktig informasjon i figuren ved denne justeringa.

Eg fekk mange simulerte verdiar som var tilnærma lik null. Jamfører vi dei to histogramma i figur 4.2 på neste side, ser vi at min andel av verdiar i intervallet

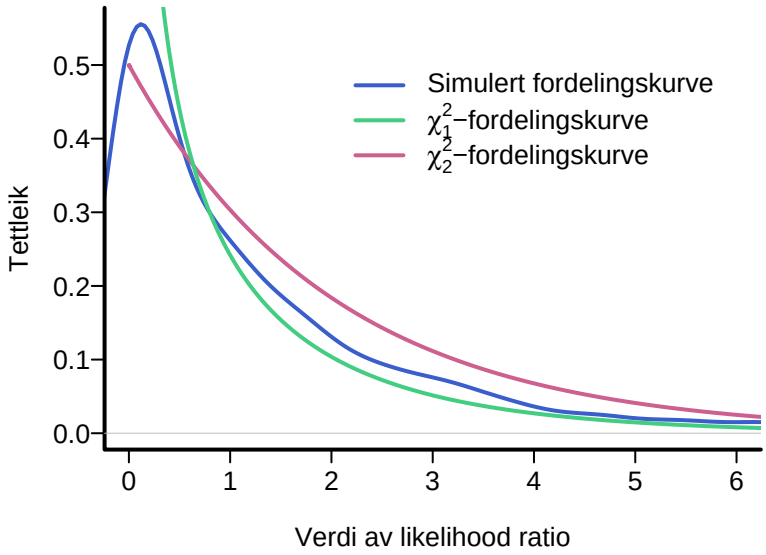


Figur 4.2: Histogrammet til venstre viser frekvens av verdiar i ei χ^2_2-fordeling, mens histogrammet til høgre viser frekvens av verdiar til den simulerte likelihood ratioen for `Orthodont` modellane.

$(0, 0.5]$ er mykje større enn i χ^2_2-fordelinga. Samstundes ser det ut til at dei simulerte verdiane har eit lettare haleparti enn det som er konsistent med χ^2_2-fordelinga.

Eg bereknar p-verdien ved likning (4.1) på førre side, og får med $n = 1000$ at p $= \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mathbf{\Lambda}}_i \geq 0.833\} = 0.464$. Dette er ein betydeleg mindre p-verdi enn den vi fekk i eksempel 3.3.3. Resultatet mitt tydar på at det at varianselementet ψ_2, som vi ynskjer å inferere om, ligg på randa av definisjonsområdet sitt under H_0, påverkar fordelinga til testobservatoren, og gjev ein konservativ p-verdi.

I følge Stram og Lee, vil å teste ei hypotese på form (3.17) på side 43, gje ein likelihood ratio-observator som har ei 50:50 blanding av ei χ^2_1- og ei χ^2_2-fordeling. I figur 4.3 på neste side har eg plotta tettleikskurva av observasjonane mine saman med tettleikskurva til høvevis χ^2_1- og χ^2_2-fordelinga. Den simulerte kurva fell mellom χ^2_1- og χ^2_2-fordeingane, i tråd med ei 50:50 blanda fordeling, men med muligens litt meir vekt på χ^2_1-fordelinga. Figuren viser at inferens for varainskomponenten ψ_2 med χ^2_2-fordelinga som fordeling for testobservatoren vil gje ein høgare p-verdi enn det som er reelt.



Figur 4.3: Vi ser her at den simulerte kurva fell mellom χ^2-tettleikane med 1 og 2 friheitsgradar. Kurvene til fordelingane er indikerte ved fargane i ramma.

4.3 Fordelinga til likelihood ratio for blanda modellar

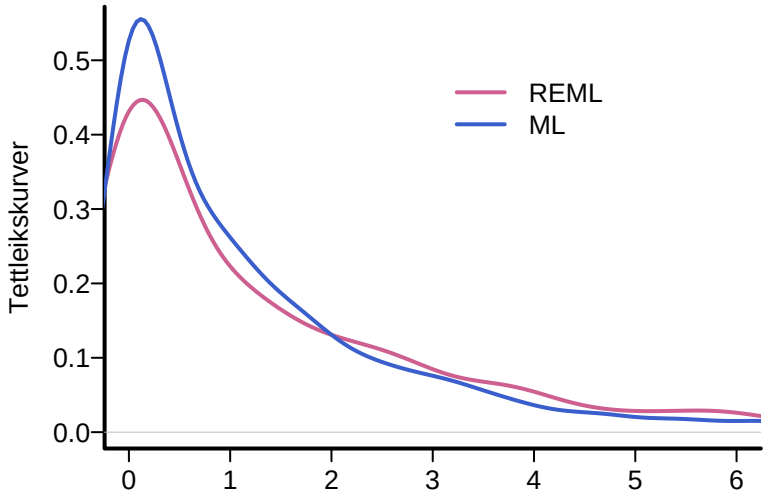
Stram og Lee (1994) generaliserte sine resultat til å kunne gjelde for to vilkårlege nøsta modellar. Dei meiner i sin artikkel at likelihood ratioen til to nøsta modellar med høvevis k og $k+1$ random effects vil ha ei fordeling på form

$$\mathbf{\Lambda}_B \sim 0.5\chi^2_k + 0.5\chi^2_{k+1}. \tag{4.2}$$

Eg nyttar notasjonen $\mathbf{\Lambda}_B$ for denne likelihood ratioen for å understreke at det gjeld likelihood ratioen til modellar der ein av dei har varianselement som under H_0 er sett lik null. For modellar der ein kun har gjort endringar i dei faste parametrane, er den generelle fordelinga i (3.16) på side 41 ikkje konservativ.

Figur 4.3 viser ikkje at mi simulerte kurve følger ei 50:50 fordeling eksakt. Dette vil føre til at dersom vi samanliknar testobservatoren med ei 50:50 blanding, vil vi også få ein noko konservativ p-verdi. Dei blanda modellane vi har tilpassa `Orthodont` har begge hatt samme faste parametrar, det vil seie at designmatrisa $\boldsymbol{X}_i$ er den samme i dei to modellane. I følge Pinheiro og Bates kan ein då samalikne REML-estimat for modellane, og dermed også berekne ein restricted likelihood ratio. For å undersøke om ein restricted likelihood ratio vil gje ei betre tilnærming til fordelinga til Stram og Lee, vil eg simulere ein restricted likelihood ratio for modellen med éin variabel parameter og modellen med to variable parametrar.

Eg nyttar då parameterestimata i tabell 4.1 på side 49 til å simulere nye datasett. Koden er gitt i tillegg B.



Figur 4.4: Tettleiksurvar til verdiar simulert ved høvevis REML og ML.

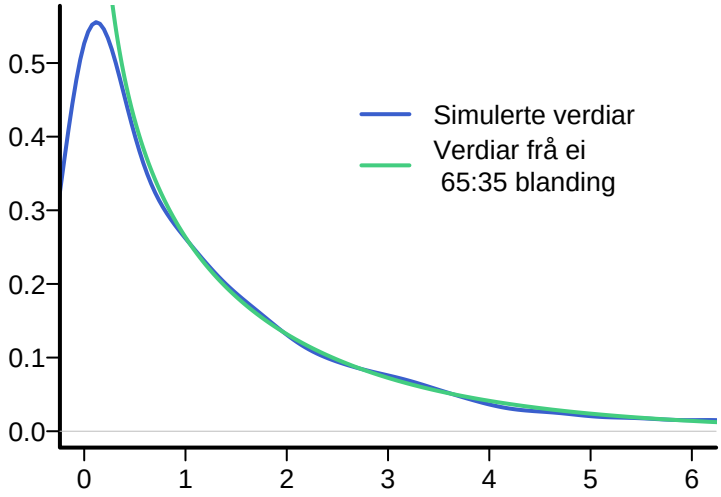
I figur 4.4 ser vi at tettleikskurva til verdiane simulert ved REML har ein mindre topp i null, og ein noko tyngre hale enn verdiane simulert ved ML. Den observerte restricted likelihood ratio-verdien for `Orthodont`-modellane er $\hat{\mathbf{\Lambda}}_R = 1{,}176$, og den tilhøyrande p-verdien ved samanlikning med ei χ^2_2-fordeling er 0,556. Denne p-verdien er lågare enn den eg fekk i eksempel 3.3.3 ved ML estimering, men jamført med dei simulerte ratioane, er p-verdien konservativ sidan 373 av 1000 simulerte restricted likelihood ratioar er større eller lik 1,176.

Pinheiro og Bates (2000) viser til at ei 50:50 fordeling på form (4.2) på førre side ikkje alltid er ei god tilnærming. For eit datasett dei har sett på, der dei har testa ein likelihood ratio av likelihooden til ein ordinær regresjonsmodell og ein random intercept-modell, viser simulering av likelihood ratioen at ei $0.65\chi^2_0 + 0.35\chi^2_1$-fordeling passar betre. Inspirert av dette prøver eg å plotte ei $0.65\chi^2_1 + 0.35\chi^2_2$-fordelingskurve saman med den simulerte kurva. Dette gav meg figur 4.5 på neste side.

Som vi ser i figuren ser denne kurva ut til å høve omlag eksakt. Det er høgre hale som er viktig for bestemminga av p-verdi.

Eg oppsumerar p-verdiane eg har berekna ved REML og ML, med dei ulike χ^2-fordelingane eg har nemnt hittil, i tabell 4.2 på neste side.

Vi ser at p-verdiane estimert med REML er jamnt lågare enn dei med ML. For begge metodane treff ei $0.65\chi^2_1 + 0.35\chi^2_2$-fordeling best den simulerte p-verdien.



Figur 4.5: Den simulerte kurva til verdiar estimert ved ML plotta saman med fordelingskurve til ei $0.65\chi^2_1 + 0.35\chi^2_2$-fordeling.

Metode	Simulerte data	χ^2_2	Blanda 50:50	Blanda 65:35
ML	0,464	0,659	0,510	0,466
REML	0,373	0,556	0,417	0,375

Tabell 4.2: Ei samanlikning av estimert p-verdi under ML og REML.

Med desse resultata ser det ut til at ei $0.65\chi^2_1 + 0.35\chi^2_2$-fordeling høver best for likelihood ratioen til modellen med éin variabel parameter i likning (2.4) og modellen med to variable parametrar i likning (2.6) for `Orthodont`. Ei jamføring av den observerte ratioen på 0,833 med denne fordelinga, gjev ikkje ein konservativ p-verdi i høve til p-verdien vi fann ved å jamføre den observerte ratioen med ei simulert fordeling. Ei analyse av den avgrensa likelihood ratioen RLR, gjev eit tilsvarande resultat.

Med p-verdiar på høvevis 0,464 og 0,373 for ein simulert LRT og ein simulert RLRT, kan vi ikkje forkaste modellen under H_0 med éin variabel parameter. Dette betyr at observasjonane av den ortopediske avstanden i `Orthodont` ikkje gjev grunnlag for å anta at denne avstanden aukar med eit stigningstal som varierer mellom individ. Basert på desse dataa aukar avstanden i følge ein lineær blanda regresjonsmodell med éin variabel parameter b_i, som modellerar endringar i startverdi mellom individ som stokastisk.

Det kan vere at dersom ein hadde hatt fleire målingar per individ og målingar på fleire born, at dataa hadde vist ei betre tilpassing til den 50:50 vekta fordelinga til

Stram og Lee (1994) i likning (4.2). Ein ville då mogeligeins fått meir nøyaktige estimat av varianselementa til modellane, og dermed fått eit betre utgangspunkt for simuleringa.

Konklusjonen av simuleringa mi, er at generell teori for hypotesetesting med ein LR (og ein RLR) der modellen under H_0 har varianselement som er sett lik null, vil gje konservative p-verdiar dersom ein samanliknar testobservatoren med ei χ^2_{p-q}-fordeling, der $p - q$ er talet på varianselement som er sett lik null i modellen under H_0.

All generalizations are dangerous, even this one.

Alexandre Dumas (1802 - 1870)

5

Generaliserte lineære blanda modellar

I dette kapittelet skal vi utvide den lineære blanda modellen til å kunne modellere responsvariablar med andre fordelingar enn normalfordelinga. Eg ynskjer å analysere longitudinelle observasjonar av antall krav innan yrkesskadeforsikring. Ein normalantaking er ikkje rimeleg for desse observasjonane. Når, i tillegg, observasjonar er grupperte, til dømes om ein har observasjonar frå ulike selskap, vil ein ordinær regresjonsmodell ikkje fange opp variasjon mellom grupper. Ein blanda modell vil vere berettiga for slike observasjonar, men responsvariabelen vil ikkje lenger følge ei normalfordeling slik den har gjort i dei førre kapittela.

Datasettet eg vil studere i dette kapittelet består av longitudinelle observasjonar av talet på krav frå uføreforsikra arbeidarar i ulike risikogrupper henta frå the National Council on Compensation Insurance (USA). Variabelen av interesse er antall krav, og ei Poisson fordeling vil vere rimeleg å anta for denne. I ein artikkel av Antonio og Beirlant (2006) er ein *generalisert lineær blanda modell* med Poisson fordeling for responsen definert og analysert, og denne artikkelen er mitt utgangspunkt for å analysere datasettet, kalla «Worker's compensation insurance», i lys av teori om generaliserte lineære blanda modellar.

Før eg kjem så langt vil eg definere ein ordinær generalisert lineær modell.

5.1 Generaliserte lineære modellar

Ein *generalisert lineær regresjonsmodell*, forkorta GLM, utvidar den normale lineære regresjonsmodellen i likning (1.1) på side 4 til å kunne modellere responsvariablar med fordelingar frå heile eksponensialfamilien.

Ei fordeling som er medlem av *eksponensialfamilien* er ei fordeling som kan skrivast på forma

$$f(y) = \exp\left\{\frac{y\theta + b(\theta)}{\phi} + c(y, \phi)\right\}, \tag{5.1}$$

der $b(.)$ og $c(.)$ er kjente funksjonar, og θ og ϕ parametrar (Antonio og Beirlant 2006). Parameteren θ kallast for den *naturlige* eller *kanoniske* parameteren, og parameteren ϕ for dispersjonsparameter eller *skaleringsparameter*, på engelsk *scale* parameter.

Variablar med fordelingar i eksponensialfamilien kan uttrykke forventning og varians ved parametrane ϕ, θ, og funksjonen $b(\theta)$. Forventning til ein variabel y med fordeling i eksponensialfamilien, kan uttrykkast ved

$$\mu = E(y) = b'(\theta),$$

og variansen til y som

$$\text{Var}(y) = \phi b''(\theta) = \phi V(\mu). \tag{5.2}$$

I uttrykk (5.2) er derivasjon med omsyn på den naturlige parameteren θ, parameteren ϕ er skaleringsparameteren, og funksjonen $V(.)$ er variansfunksjonen. Sidan variansen kan uttrykkast som ein funksjon av μ, det er mogeleg å undersøkje eit høve mellom forventning og varians for variablar med fordeling i eksponensialfamilien. I følgande eksempel vil eg vise at normalfordelinga er eit medlem av eksponensialfamilien, og at ein kan finne forventning og varians i fordelinga ved formlane over.

Eksempel 5.1.1: Normalfordelinga

Fordelingsfunksjonen til ein nomalfordelt variabel $y \sim \mathcal{N}(\mu, \sigma^2)$, er definert som

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Ved å ta log av uttrykket, opphøge i eksponenten e og multiplisere ut utrykket i parantesen, ser ein at

$$\begin{aligned} f(y) &= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \end{aligned}$$

der då μ er den naturlege parameteren, og σ^2 skaleringsparameteren.
Ved å nytte funksjonane for forventning og varians, finn ein at

$$E(y) = b'(\mu) = \frac{d}{d\mu}\left(\frac{1}{2}\mu^2\right) = \mu,$$

og for variansen at

$$\text{Var}(y) = \phi b''(\mu) = \sigma^2 \cdot 1 = \sigma^2,$$

som er konsistent med ein normalfordelt variablel $y \sim \text{N}(\mu, \sigma^2)$.

Ein generalisert lineær modell, forkorta GLM, er ein regresjonsmodell der responsvariabelen har ei fordeling som er medlem av eksponensialfamilien. I tillegg til å kunne modellere ikkje-normale fordelingar kan ein generalisert lineær modell binde ein funksjon g av forventninga $E(y) = \mu$, til forklaringsvariablane. Denne funksjonen kallast *link*-funksjon, og er ofte ikkje-lineær. Eit krav for denne funksjonen er at den er deriverbar og har ein invers g^{-1}. Oppsummert består ein GLM av tre element: ein respons med fordeling frå eksponensialfamilien, forklaringsvariablar og ein link-funksjon g. Definisjon 5.1.2 oppsummerar krava som må oppfyllast for at ein modell skal vere ein generalinsert lineær modell.

Definisjon 5.1.2: Ein generalisert lineær modell

Ein modell er ein generalisert lineær modell dersom den kan uttrykkast på forma

$$g(\mu_i) = g(E(y_i)) = \boldsymbol{X}_i\boldsymbol{\beta} \quad i = 1, \ldots, n$$

der

i) responsvariablane $y_1, \ldots, y_n$ er på kanonisk form og kjem frå samme fordeling $f(y_i, \theta)$ i eksponensialfamilien,

ii) høgre sida i modellen består av ein $p \times 1$ vektor $\boldsymbol{\beta}$ av parametrar, og ei $p \times n$ matrise $\boldsymbol{X}_i$ av forklaringsvariablar,

iii) og desse er forbunde til forventninga av responsen på venstre sida ved ein monoton link-funksjon g, der $g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ og $\mu_i = E(y_i)$.

Det er vanleg å uttrykke regresjonskoeffisientane ved ein *lineær prediktor* η_i, på forma

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} = \eta_i. \tag{5.3}$$

Ein seier gjerne at link-funksjonen g er bindeleddet mellom den lineære prediktoren η_i og forventninga μ_i.

Ein GLM har eit større definisjonsområde enn ein ordinær lineær modell, sidan den ikkje berre kan anvendast i tilfeller der normalantakinga må vike, men sidan den også kan modellere ein ikkje-lineær transformasjon av forventninga μ_i. Tabell 5.1 på neste side oppsummerar nokre fordelingar i eksponensialfamilien med døme på tilhøyrande link-funksjon, invers link-funksjon $\mu_i = g^{-1}(\eta_i)$, variansfunksjon og skaleringsparameter ϕ.

Fordeling	μ_i	Link	Invers link	$V(\mu_i)$	ϕ
Normal(μ, σ)	μ	Identitet	η	1	σ^2
Poisson(μ)	μ	Log	e^{η}	μ_i	1
Gamma(α, β)	$\alpha\beta$	Invers	$1/\eta$	μ_i^2	α^{-1}
Neg bin(r, p)	$r(1-p)/p$	Log	e^{η}	$\mu_i + \mu_i^2/r$	$1/p$?
Bin$(n, p)/n$	p	Logit	$e^{\eta_i}/(1+e^{\eta_i})$	$\mu_i(1-\mu_i)$	$1/n$
		Probit	$\Phi(\eta_i)$		

Tabell 5.1: Nokre vanlege fordelingar frå eksponensialfamilien med tilhøyrande link-funksjonar, uttrykk for variansfunksjonen og skaleringsparameter.

Den ordinære lineære regresjonsmodellen i likning (1.6) på side 10 for `Orthodont`-dataa, er av den mest vanlige typen generaliserte lineære modellar. Ein ordinær lineær regresjonsmodell er ein GLM med identitetsfunksjonen som link-funksjon og normalfordelte responsvariablar. Den kan skrivast i GLM notasjon som

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{X}_i\boldsymbol{\beta} \quad \boldsymbol{\mu}_i = \boldsymbol{\eta}_i \\ \boldsymbol{y}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma) \quad i = 1, \ldots, n, \end{aligned}$$

der link-funksjonen g er lik identitetsfunksjonen, $\mu = \eta$.

Generaliserte lineære modellar har eit stort bruksområde innan både livs- og skadeforsikring. Som aktuar er ein interessert i å estimere risiko knytta til ulike forsikringsobjekt. Ein ynksjer å ha kjennskap til *tapsfordelinga* til objektet, for å vite noko om risikoen det er å ta på seg eit forsikringsansvar for objektet. Det vil ofte ikkje vere rimeleg å anta at observasjonar av tap er normalfordelte.

Ei forutsetnad for å tilpasse ein GLM til eit datasett er at datasettet består av enkle, uavhengige observasjonar. I tilfeller der ein ynskjer å modellere ein resonsvariabel på tvers av grupper eller objekt vil ikkje uavhengigheit vere rimeleg å anta. Til dømes når ein har longitudinelle observasjonar forventar ein større variasjon for observasjonar i ulike grupper eller objekt, enn for observasjonar i samme gruppe. I kapittel 2 utvida vi ein LM til ein LMM for å modellere longitudinelle observasjonar med antatt normalfordeling. Vi skal i følgande delkapittel sjå korleis vi kan utvide ein GLM til å kunne modellere longitudinelle observasjonar med fordelingar frå heile eksponensialfamilien.

5.2 Generaliserte lineære blanda modellar

Ein *generalisert lineær blanda modell*, forkorta GLMM frå den engelske terminologien «generalized linear mixed model», fåast dersom ein, i den lineære prediktoren i likning (5.3) på side 59, lar nokre av parametrane i vektoren $\boldsymbol{\beta}$ variere mellom eksperimentelle einingar. Som i overgangen frå ein LM til ein LMM, legg ein til variable parametrar til den delmengda av regresjonskoeffisientane ein antar at varierer mellom einingar. Desse vil vi også i ein GLMM uttrykke ved ein vektor $\boldsymbol{b}_i$ og ei matrise $\boldsymbol{Z}_i$ som er ei delmatrise av designmatrisa $\boldsymbol{X}_i$.

Inferens for ein GLMM ofte er svært komplisert. Eg vil snart vise at den generelle lineære blanda modellen i likning (2.1) på side 17 er eit spesialtilfelle av ein GLMM der dei særskilde antakingane om link-funksjon og fordeling gjer inferens for ein LMM mindre komplisert enn i andre tilfeller.

Først vil eg uttrykke elementa i ein GLMM på grunnlag av Fitzmaurice *et al.* (2004) sine definisjonar.

Definisjon 5.2.1: Ein generalisert lineær blanda modell

Ein modell er ein generalisert lineær modell dersom og berre dersom den kan uttrykkast på forma

$$g(E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) = \boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \quad i = 1, \ldots, n, \tag{5.4}$$

der

i) den *betinga* fordelinga til responsen gitt vektoren med variable parametrar, $\boldsymbol{y}_i \mid \boldsymbol{b}_i$, er ei fordeling som er medlem av eksponensialfamilien uttrykt i (5.1) på side 57 slik at forventninga kan uttrykkast ved $E(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = b'(\theta)$ og variansen ved $\text{Var}(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \phi b''(\theta) = \phi V(\boldsymbol{\mu}_i)$, og der forklaringsvariablane $\boldsymbol{X}_i$ er uavhengig av dei variable parametrane $\boldsymbol{b}_i$,

ii) komponentane på høgre sida i modellen er både faste og variable, representert ved ei designmatrise $\boldsymbol{X}_i$ $(p \times n)$ med tilhøyrande faste parametrar $\boldsymbol{\beta}$, og ei matrise $\boldsymbol{Z}_i$ $(q \times n)$, der $q \leq p$, med tilhøyrande variable parametrar $\boldsymbol{b}_i$, og desse kan uttrykkast ved den betinga forventninga $E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)$ via ein monoton link-funksjon g som i likning (5.4),

iii) og der dei variable parametrane i vektoren $\boldsymbol{b}_i$ antas å ha ei multivariat sannsynsfordeling,

$$\boldsymbol{b}_i \sim F(\boldsymbol{0}, \boldsymbol{\Psi}) \quad i = 1, \ldots, n. \tag{5.5}$$

I prinsippet kan dei variable parametrane ha ei vilkårleg multivariat fordeling, men for reknemessige grunnar antas ofte denne å vere ei multivariat normalfordeling, $\boldsymbol{b}_i \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Psi})$. I dei programfunksjonane eg nyttar er denne antakinga lagt til grunn.

Til forskjell frå ein ordinær lineær blanda modell, LMM, blir ein GLMM definert ved den betinga fordelinga til responsen. I den lineære blanda modellen såg vi at ei marginal fordeling av responsen, $E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, gav ei populasjonsretta tolkning, medan den betinga fordelinga gav ei individ-spesifikk tolkning. Den marginale forventninga av responsen kunne finnast direkte frå den betinga ved

$$\begin{aligned} E(\boldsymbol{y}_i \mid \boldsymbol{b}_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ E(\boldsymbol{y}_i) &= E\left(E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)\right) = E\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i\right) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i E(\boldsymbol{b}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta}. \end{aligned}$$

Altså er den betinga og den marginale fordelinga i ein LMM den samme.

For den generaliserte lineære blanda modellen vil vi ikkje ei slik utledning vere mulig så lenge link-funksjonen ikkje er lik identitetsfunksjonen, sidan

$$\begin{aligned} g(E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ g(E(\boldsymbol{y}_i)) &= g\left(EE(\boldsymbol{y}_i \mid \boldsymbol{b}_i)\right) = g\left(E\left(g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)\right)\right) \\ &\neq \boldsymbol{X}_i\boldsymbol{\beta} \end{aligned}$$

Den marginale forventninga har altså ikkje samme tolkning i ein GLMM som i ein LMM. Ein GLMM på forma (5.4) på førre side vil vere best dersom ein er interessert i å undersøkje individa sin påverknad på ein populasjon istadenfor populasjonssnittet (Antonio og Beirlant 2006).

Ein LMM er eit spesialtilfelle av ein GLMM der den betinga fordelinga til responsen, og fordelinga til dei variable parametrane, begge er multivariat normal, og der den lineære prediktoren er forbunde til forventningsvektoren ved identitietsfunksjonen. Vi kan sjå at den lineære blanda modellen i kapittel to kan skrives på GLMM form som

$$\begin{aligned} E(\boldsymbol{y}_i \mid \boldsymbol{b}_i) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ \boldsymbol{y}_i \mid \boldsymbol{b}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad \boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_i), \end{aligned}$$

der $\boldsymbol{\mu}_i = \boldsymbol{\eta}_i$. Dette er modellen som eg kalla ein betinga modell i delkapittel 2.2.

Både den ordinære generaliserte lineære modellen, og den generaliserte lineære blanda modellen har eit stort bruksområde for forsikringsdata. Haberman og Renshaw (1996) nevner emner som levetidsmodellering, tilstandsmodellar for helseforsikring, risikoklassifisering og modellering av antall krav i skadeforsikring, som tilfeller der ein GLM gir god tilpassing. Eg vil snart ta for meg eit datasett med observasjonar av antall krav innan yrkesskadeforsikring.

5.3 Estimeringsmetodar og inferens i ein GLMM

Som nemnt, er inferens og estimering av parametrar i ein GLMM komplisert. Unntak er modellar der vi antar at både den betinga fordelinga til responsen og vektoren med variable parametrar har ei multivariat normalfordeling. Dette kallast for normal-normal tilfellet. Eit anna tilfelle som gjer inferens enklare er Poisson-gamma tilfellet (Antonio og Beirlant 2006), men ei multinormalfordeling for dei variable parametrane blir likevel ofte foretrukke sidan det gjev ein meir allsidig modell.

Vi tek utgangspunkt i ein generell GLMM for n eksperimentelle einingar på forma

$$\begin{aligned} g(E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)) &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i \\ \boldsymbol{y}_i \mid \boldsymbol{b}_i \sim f_p(\boldsymbol{\mu}_i, \phi V(\boldsymbol{\mu}_i)) \quad \boldsymbol{b}_i &\sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Psi}), \quad i = 1, \ldots, n \end{aligned} \tag{5.6}$$

der f er ei p-variat fordeling i eksponensialfamilien på forma i (5.1) på side 57, vektoren $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i \mid \boldsymbol{b}_i)$ er forventningsvektoren som er forbunde til forklaringsvariablane ved ein link-funksjon g slik at $\boldsymbol{\mu}_i = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i)$, den betinga variansen kan uttrykkast ved $\text{Var}(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \phi V(\boldsymbol{\mu}_i)$, og kovariansmatrisa $\boldsymbol{\Psi}$ til vektoren med variable parametrar $\boldsymbol{b}_i$ har elementa

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_{12} & \cdots & \psi_{1q} \\ \psi_{21} & \psi_2 & \cdots & \psi_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{q1} & \psi_{q2} & \cdots & \psi_q \end{bmatrix} .$$

For elementa i kovariansmatrisa $\boldsymbol{\Psi}$ skal vi, som for den generelle LMM, anta at $\psi_{jk} = \psi_{jk}$ slik at matrisa er symmetrisk, og at dersom eit diagonalelement $\psi_j = 0$ vil elementa i den tilhøyrande rada ψ_{jk}, $k = 1, \ldots, q$, og kolonna ψ_{kj}, $k = 1, \ldots, q$, vere lik null. Fordelingane i tabell 5.1 på side 60 med tilhøyrande link-funksjon er også ofte nytta for longitudinelle data.

Dersom kan gjere ei antaking om fordelinga f til responsvariabelen med ein tilhøyrande link-funksjon g som bind forventningsvektoren $\boldsymbol{\mu}_i$ til den lineære prediktoren $\boldsymbol{\eta}_i$, er dei ukjente parametrane i modellen vektoren $\boldsymbol{\beta}$ med dei faste parametrane, elementa i kovariansmatrisa $\boldsymbol{\Psi}$ til dei variable parametrane som ofte definerast i ein vektor $\boldsymbol{\psi}$, og skaleringsparameteren ϕ. Likelihood-funksjonen for den samla fordelinga til responsen vil dermed i det balanserte tilfellet kunne uttrykkast ved

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}, \phi) = \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{b}_i). \tag{5.7}$$

Sidan vi i ein GLMM kjenner både den betinga fordelinga $f(\boldsymbol{y}_i \mid \boldsymbol{b}_i)$ til responsen, og fordelinga til vektoren med variable parametrar $\boldsymbol{b}_i$ som er normalfordelinga, kan vi uttrykke likelihood-funksjonen uavhengig av vektoren med variable parametrar ved å integrere over $\boldsymbol{b}_i$. Dermed kan vi skrive

$$L(\boldsymbol{\beta}, \boldsymbol{\psi}, \phi) = \prod_{i=1}^{n} \int f(\boldsymbol{y}_i \mid \boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i. \tag{5.8}$$

Denne likelihood-funksjonen (Fitzmaurice *et al.* 2004, side 338) kan estimerast ved kvadratur metodar, som gaussisk kvadratur og Laplace approksimajonar, eller ved pseudo-likelihood metodar som penalized quasi-likelihood estimering, PQL (Antonio og Beirlant 2006). Eg vil ikkje gå i detaljar om teorien bak desse metodane, med i analysar i R vil eg sjå på kva som skil metodane.

Som nemnt i innleiinga er ei utfordring ved estimering i GLMM at ein i diskrete fordelingar vil oppleve at responsvariabelen er null. For nokre av link-funksjonane vil dette medføre at ein i tilpassinga av modellen deler på null og får estimat som går mot uendeleg. Den enklaste måten å løyse dette problemet på er å fjerne observasjonar som gir oss vanskar med å tilpasse modellen. Eit anna problem for GLMM-modellar kan vere å gjere dei rette antakingane om fordelinga til responsvariabelen i eit datasett. Det same gjeld for antaking av link-funksjon. Eg vil no analysere eit datasett med ikkje-normale observasjonar, der eg vil sjå på om antakingane mine om fordelinga til responsen er rimeleg.

5.4 Ein Poisson GLMM for skadeforsikringsdata

Feltet skadeforsikring har mange undergrupper, som bilforsikring, eigendomforsikring og innbuforsikring for å nemne nokre. Felles for desse undergruppene er

at ein ynskjer å estimere *risikoen* for eit selskap som sel ei bestemt kontrakt, kalla ei *polise*, der dei tek på seg eit bestemt forsikringansvar til ein eigendel for kunden mot å få eit innskudd, kalla *premie*.

For å berekne ein rimeleg premie for ei polise, ynskjer ein å vite mest mulig om risikoen til polisa. Kredibilitetsteori bygger på antakinga om at alle poliser har ein underliggande *risikoparameter* som avgjer kor stor risiko det er for eit selskap å selge ei bestemt polise. Det er hovudsakeleg to prosessar ein ynskjer kjennskap til. Den eine er hyppigheita av krav i poliser, og den andre er storleiken av krav i poliser. Desse prosessane er gjerne avhengige. Tildømes vil ein bilførar som køyrer risikofylt ikkje berre ha stor sjanse for å krasje bilen sin, men òg få store skadar på bilen når uhellet først er ute. Kjennskapen om desse prosessane er innehaldt i risikoparameteren Θ, og i dens tilhøyrande *struktur*-fordeling, U (Sundt 1999, side 30).

Ein klassifiserer ofte polisene i undergrupper som følge mellom anna av *kva* dei skal forsikre: person, eigendom, gjeld etc.; *årsak* til forsikringskrav: brann, innbrudd etc.; og *betingelsar* for at kravet vil godtas: under reise, på jobb, etc., for å nevne nokre (Sundt 1999, side 7). Desse punkta vil gjerne bli delt opp i mindre klassar. Ein antar så at kvar gruppe har sine individuelle risikokarakteristikkar, men også karakterestikkar som er felles.

Dersom ein, for ei polise i, kjente fordelinga U_i til den tilhøyrande risikoparameteren Θ_i, ville det vore lett å berekne risikoen til polisa. Men sidan Θ_i er uobserverbar vil ein aldri med sikkerheit kjenne fordelinga U_i. Ein er derfor interessert i å estimere Θ_i for ei polise basert på observasjonar av risikoen i polisa, og observasjonar av risikoen i liknande poliser. Ein blanda modell er dermed aktuell sidan den kan formulere ein populasjonsretta modell som er felles for gruppene, samt gje parametrar som fortel kor mykje kvar undergruppe skil seg frå populasjonen.

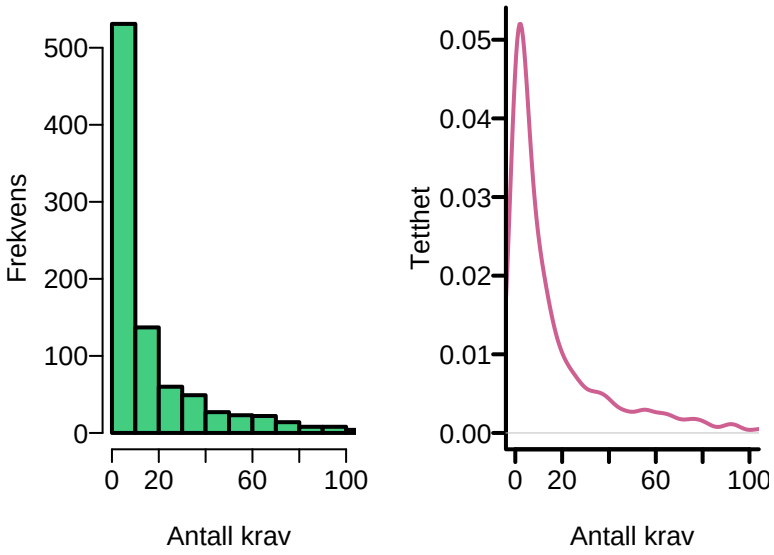
Eg vil studere eit datasett der det er talet på krav ein er interessert i. Poisson fordelinga er som regel antatt for telle-observasjonar. Ei Poisson fordeling med parameter μ er definert som

$$P(Y_i = y \mid \mu) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, \ldots. \tag{5.9}$$

I Poisson fordelinga er variansen lik forventninga, derfor antar ein som regel, for Poisson-fordelte variable, at skaleringsparameteren er lik éin, altså $\phi = 1$. Nokre reknealgoritmar gjev eit estimat av ϕ etter at modellen er tilpassa. Som regel vert log-funksjonen i tabell 5.1 på side 60 nytta som link-funksjon i ein Poisson GLMM.

5.4.1 Worker's compensation insurance

I undergruppa person, er forsikring for yrkesuførheit eit viktig område. Ulike yrkesgrupper vil ha ulik risiko, men samstundes vil ein forvente somme likheitar. Når ein skal berekne risiko til ei yrkesgruppe, vil det derfor vere informasjon å hente i observasjonar frå liknande yrkesgrupper. Eg vil no studere eit datasett som har registrert antall krav i ulike yrkesgrupper, kalla «Worker's compensation insurance».



Figur 5.1: Histogram og tettleikskurve for responsvariabelen Count_i i datasettet `Klugman`.

Datasettet «Worker's compensation insurance» består av talet på krav frå forsikringspoliser i 133 ulike yrkesgrupper over den samme sjuårsperioden. Datasettet blei, såvidt eg veit, først trekt fram av Klugman (1992). Eg vil derfor eg referere til datasettet som `Klugman`-dataa heretter.

Talet på krav i dei ulike yrkesgruppene er registrert ved variabelen Count_i, der i er indeks for kva for gruppe vi ser på. Det rimeleg å anta ei Poisson fordeling for variabelen Count_i. Mulige forklariangsvariablar for Count_i er året krava dukka opp, registrert ved variabelen Year_i, og observasjonar av lønningslister, registrert under Payroll_i, som skal gje eit mål på talet av arbeidarar i yrkesgruppa.

Av histogramma i figur 5.1 ser vi at dei låge verdiane er sterkast representert. I histogramma er observasjonar der Payroll_i er lik null fjerna, slik at vi ser på 895 frekvensar av variabelen Count_i. Histogramma viser ikkje frekvensen til antall krav større enn hundre. Den høgaste observasjonen av Count_i er 228. Kun 16

observasjonar av Count_i ligg i intervallet $[100, 228]$, derfor har eg kutta aksen i histogramma ved 100. Vidare er omlag ein femtedel av verdiane til Count_i lik null , og omlag halvparten av dei registrerte krava i løpet av eit år i ei yrkesgruppe i mindre eller lik fem (jamfør tillegg C).

Tabell 5.2 viser snittet og variansen av antall krav per år samt høve av dei. Som vi

År	1	2	3	4	5	6	7
Snitt	14,90	16,24	16,63	17,54	22,80	17,30	16,71
Varians	690,49	624,60	632,08	790,22	1181,82	720,31	649,07
Høve	46,33	38,45	38,02	45,06	51,84	41,63	38,84

Tabell 5.2: Snitt, varians og ratio av snitt over varians for antall krav per år i settet Klugman.

ser er det teoretiske høve éin, mellom snitt og varians i Poisson fordelinga, tydeleg overskride.

Desto fleire arbeidarar det er i ei gruppe, jo fleire krav vil ein forvente. I følge Haberman og Renshaw (1996) er det rimeleg å sjå på storleiken av lønningslistene, som dei har kalla *exposures*, og antallet krav som para observasjonar. Dermed ser vi på observasjonar av $(\text{Count}_i, \text{Payroll}_i)$ for dei ulike yrkesgruppene, $i = 1, \ldots, 133$. Eg nyttar notasjon $\boldsymbol{y}_i$ om ein vektor med observasjonar av Count_i for yrkesgruppe i i den aktuelle sjuårs-perioden, og Payroll_i for den tilsvarande lønningslista for gruppe i i samme perioden.

Dei to modellane Antonio og Beirlant (2006) har definert i sin artikkel, som modellerar antall krav $\boldsymbol{Y}_i$, er ein Poisson GLMM med variabelt konstantledd definert som

$$\begin{aligned} \log(\boldsymbol{\mu}_i) &= \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_i \\ \boldsymbol{y}_i \mid b_i &\sim \text{Poisson}(\boldsymbol{\mu}_i) \quad b_i \sim \mathcal{N}(0, \psi_0), \\ i &= 1, \ldots, 133, \end{aligned} \tag{5.10}$$

og ein Poisson GLMM med ein variabel parameter lagt til både konstantledd og stigningstal definert ved

$$\begin{aligned} \log(\boldsymbol{\mu}_i) &= \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_{i0} + \text{Year}_i \cdot b_{i1} \\ \boldsymbol{y}_i \mid \boldsymbol{b}_i &\sim \text{Poisson}(\boldsymbol{\mu}_i) \quad \boldsymbol{b}_i \sim \mathcal{N}_4(\boldsymbol{0}, \boldsymbol{\Psi}), \\ i &= 1, \ldots, 133 \end{aligned} \tag{5.11}$$

der vektoren med dei variable parametrane og tilhøyrande kovariansmatrisa er definerte som

$$\boldsymbol{b}_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \quad \text{og} \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_{12} \\ \psi_{21} & \psi_2 \end{bmatrix}.$$

Link-funksjonen i modellane er log-funksjonen. Variabelen $\log(\text{Payroll})_i$ har ingen parameter foran seg. Dette er fordi ein i modellane antar at høve mellom $\log(\text{Count})_i$ og $\log(\text{Payroll})_i$ er lik éin, og er ikkje interessert i estimere ein eventuell parameter $\beta \neq 1$ for $\log(\text{Payroll})_i$ i modellen. Ein slik parameter kallast ein *offset*-parameter (McCulloch 2003), og i denne samanhengen betyr inkluderinga av denne at ein ynskjer å sjå på antall krav per arbeidar. Klugman (1992, side 116) seier at dei registrerte tala på krav i dei ulike klassane er registrerte i same periode, slik at lønningane har vore utsette for same endringar i dollarkurs og kan slik fungere som eit mål på storleiken av dei ulike yrkesgruppene. Vi ser altså på ein modell

$$g\left(\frac{\boldsymbol{\mu}_i}{\text{Payroll}_i}\right) = \log\left(\frac{\boldsymbol{\mu}_i}{\text{Payroll}_i}\right) = \boldsymbol{\eta}_i, \tag{5.12}$$

der log-funksjonen er link mellom den lineære prediktoren og brøken på venstresida. Løyser vi opp brøken får vi

$$\log(\boldsymbol{\mu}_i) - \log(\text{Payroll}_i) = \boldsymbol{\eta}_i,$$

som gir oss den generelle formuleringa av dei to modellane i likning (5.10) og (5.11) på førre side som er

$$\log(\boldsymbol{\mu}_i) = \log(\text{Payroll})_i + \boldsymbol{\eta}_i.$$

Dersom ein undersøkjer ein modell der $\log(\text{Payroll})_i$ har ein parameter foran seg, vil ein truleg få ein betre tilpassa modell, sidan fleire parametrar som regel gir betre tilpassing. Men dette ville ikkje ha modellert det vi er interessert i som er antall krav per arbeidstakar. Ein tabell over varians over snitt til $\text{Count}_i/\text{Payroll}_i$ vises i tabell 5.3 på neste side.

Dei ukjente parametrane i modell (5.10) på førre side er parametrane β_0 og β_1, variansen til den variable parameteren b_i, ψ_0, og skaleringsparameteren ϕ. Skaleringsparameteren vil gjerne bli sett lik éin i reknealgoritmar, derfor tek eg ikkje med denne blandt dei ukjende parametrane. Eg oppsummerar dei ukjende elementa i modell (5.10) i ein vektor, $\boldsymbol{\theta}$, med tre element, $\boldsymbol{\theta} = (\beta_0, \beta_1, \psi_0)$. Ei analyse av denne modellen viser at variansen ψ_0 er ulik null i modellen. Det er altså ein forskjell i konstantledd mellom grupper i modell (5.10).

År	Snitt	Varians	Varians/Snitt
1	0,0362	0,0017	0,0457
2	0,0438	0,0037	0,0835
3	0,0398	0,0019	0,0473
4	0,0593	0,0438	0,7383
5	0,0393	0,0028	0,0707
6	0,0962	0,3451	3,5852
7	0,0483	0,0057	0,1171

Tabell 5.3: Snitt, varians og høve mellom snitt og varians for antall krav per arbeidstakar i datasettet «Klugman».

I modell (5.11) på side 67 antar vi at yrkesgruppene, i tillegg til å ha individuelle startverdiar, også har individuell endring i antall krav fra år til år. Dei ukjente parametrane i denne modellen er $\beta_0, \beta_1, \psi_1, \psi_2$ og ψ_{12}. Dersom $\psi_2 = 0$ får vi modell (5.10). Vi er dermed interessert i undersøkje om $\psi_2 > 0$.

5.4.2 Inferens for ein Poisson GLMM

Berekningsmessig er ei gamma fordeling for dei variable parametrane gunstig å anta. Poisson fordelinga er relatert til gamma fordelinga, $G(X) = \Gamma(\alpha, \beta)$, ved

$$P(X \leq x) = P(Y \geq \alpha)$$

for ein vilkårlig gamma fordelt x, der $Y \sim \text{Poisson}(x/\beta)$ og α er eit heiltal. Men normalfordeling blir ofte antall for å få ein meir generell og fleiksibel modell (McCulloch 2003). I reknefunksjonane eg nyttar i R er det lagt til grunn ei normalfordeling for dei variable parameterane.

Som nemnt er eit problem ved å nytte Poisson fordeling for observasjonar av ein tellevariabel, at antakinga $E(\boldsymbol{y}_i \mid b_i) = \text{Var}(\boldsymbol{y}_i \mid b_i) = \boldsymbol{\mu}_i$ ikkje alltid er oppfylt. Dersom $E(\boldsymbol{y}_i \mid b_i) < \text{Var}(\boldsymbol{y}_i \mid b_i)$ seier vi at dataa er *overdispersert*. Modellar tilpassa overdisperserte data kan få forventningsskjeive estimat av modellparametrane.

Overdispersjon er eit vanleg problem i Poisson modellar, og det er fleire måtar å forbetre ein modell for overdisperserte data på. Ved å gå frå ein GLM til ein GLMM vil vi ha gjort rede for varians mellom individ og slik forbetra kovariansstrukturen dersom vi har longitudinelle data. Om modellen likevel ikkje fangar opp tilstrekkeleg av variasjonen i dataa, det vil seie at skaleringsparameteren $\phi > 1$, kan det vere at ei Poisson fordeling ikkje stemmer for dataa våre. Eg vil no sjekke

kor godt modellane (5.10) og (5.11) på side 67 passar `Klugman`-dataa, og jamføre mine resultat med Antonio og Beirlant sine.

5.4.3 Analysar av modellar for Klugman data

Når eg skal analysere dei to modellane i R, må eg gjere nokre justeringar av datasettet. For det første må eg fjerne observasjonar av Payroll_i som er lik null, sidan vi ellers vil få null under brøkstreken på venstresida i likning (5.12). Vi taper ingen viktig informasjon ved dette sidan Count_i også er null der Payroll_i er null. Dette er fornuftig sidan ingen arbeidarar vil medføre ingen krav. Det står ingen plass om desse dataa er «missing data» eller om det er reelle observasjonar. Ved å fjerne desse dataa, reduserast antall yrkesgrupper frå 133 til 130, og datasettet er ikkje lenger balansert.

Eg har nytta pakken `lme4` og funksjonen `lmer` for å tilpasse desse Poisson modellane. Eg har nytta to ulike metodar i `lmer`, maksimum likelihood og Laplace approksimasjon, i estimeringa av modellparametrane, og eg ynskjer å samanlikne verdiar frå desse metodane med kvarandre, og med estimata til Antonio og Beirlant. Programkoden eg har nytta vises i tillegg C.2.

Tabell 5.4 på neste side oppsummerar resultata av analysene mine samt resultat av Antonio og Beirlant sine analysar frå bruk av SAS og WINBUGS. Rada i tabellen merka $|\boldsymbol{b}_i|$ viser antall variable parametrar, og rada merka $|\boldsymbol{\theta}|$ viser antall ukjende parametrar i modellane.

Med metoden ML, fekk eg meir negative likelihood-verdiar og dermed høgare AIC-veridar enn eg fekk i dei samme modellane analysert ved Laplace-metoden. Estimatet av variansen til random effects-parametrane, ψ_0 i random intercept modell, og ψ_1 og ψ_2 i to random effects-modellen, blei mindre ved ML-metoden enn ved Laplace-metoden. Sistnemnte metode gav eit estimat av ψ_0 som er liknande det Antonio og Beirlant har fått i sine analysar. Det kan dermed sjå ut til at ML-metoden gjev noko konservative parameterestimat.

Eg fekk eit svært estimat av ψ_2, variansen til det stokastiske stigningstalet i denne modellen. Men som vi ser av estimatet til parameteren β_1 er denne sjølv estimert til å vere svært liten.

Estimata av variansen til dei variable parametrane i modellen med random stigningstal er noko større enn det Antonio og Beirlant har fått i sine analysar. Det er rimeleg at det er ulikheiter i estimat sidan Antonio og Beirlant har nytta ei gamma fordeling for dei variable parametrane i Bayes-estimata, medan det i `lmer`

	ML	Laplace	Antonio, PQL	Antonio, Bayes
Modell (5.10)				
$\lvert\boldsymbol{b}_i\rvert$	1	1	1	1
$\lvert\boldsymbol{\theta}\rvert$	3	3	3	3
β_0	−3,540	−3,594	−3,5407	−3,602
β_1	0,0126	0,013	0,01263	0,014
ψ_0	0,304	0,797	0,7097	0,7984
ϕ	1,430	1,430		
logLik	−1041,1	−1010,2		
AIC	2088,3	2026,5		
Modell (5.11)				
$\lvert\boldsymbol{b}_i\rvert$	2	2	2	2
$\lvert\boldsymbol{\theta}\rvert$	5	5	5	5
β_0	−3,545	−3,599	−3,542	−3,583
β_1	0,013	0,012	0,012	0,0079
ψ_1	0,336	0,875	0,702	0,786
ψ_2	0,00066	0,0053	0,0013	0,0013
ψ_{12}	−0,169	−0,292		
ϕ	1,308	1,308		
logLik	−1014,3	−967,8		
AIC	2038,7	1945,5		

Tabell 5.4: Avrunda parameterestimat, estimert ved `lmer`, i dei to Poisson modellane for Klugman-dataa. Kolonna merka ML viser estimat gjort ved maksimum likelihood metode i `lmer`, medan kolonna merka Laplace viser estimat gjort ved Laplace approksimajonar i `lmer`.

vert nytta normalfordeling for desse. Eg fekk eit svært estimat av ψ_2, variansen til det stokastiske stigningstalet i denne modellen. Men som vi ser av estimatet til parameteren β_1 er denne sjølv estimert til å vere svært liten.

Antonio og Beirlant har ikkje vist sine verdiar av log-likelihood og AIC, men konkluderar med at variansen til det stokastiske stikningstalet i modell (5.11) ikkje er signifikant. I mine analysar kjem modellen med variable parametrar både for konstantledd og stigningstal best ut med ein AIC-verdi på 1945,5 estimert under Laplace-metoden, og ein differanse i log-likelihood på 42,4 til modellen med variabel parameter kun for konstantleddet.

Vi ser av tabell 5.4 at både modellen med stokastisk konstantledd og modellen med stokastisk stigningstal for `Klugman` hadde verdiar av skaleringsparameteren ϕ som var større enn 1. Dette betyr at responsvariabelen Count_i ikkje er heilt

konsistent med Poisson-fordelinga. Ein kan i slike tilfeller undersøkje tilpassinga av modellar frå to alternativ. Det første baserar seg på anta ei anna fordeling for responsvariabelen. Det andre alternativet går ut på å anta ei binomisk fordeling for verdiar som er null og ikkje null dersom eit datasett består av fleire nullverdiar enn det som er konsistent med fordelinga ein har antatt. Eg vil først sjå nærmare på alternativet der ein undersøkjer tilpassinga ei anna fordeling har til dataa.

5.5 Negativ binomisk fordeling for telledata

Når ein ynskjer å undersøkje korleis ein tellevariabel varierer som ein funksjon av forklaringsvariablar, kan det vere at det finnes fleire uobserverbare faktorar som påverkar responsen, og som medfører at tellevariabelen er utsett for større variasjon enn det ein ordinær Poisson prosess antydar. Ved å utelate desse faktorane frå modellen, vil ein observere at variansen til tellevariabelen er overdispersert (Booth *et al.* 2003), slik vi har gjort for modellane (5.10) og (5.11) på side 67. Ein måte å korrigere for denne overdispersjonen, er å anta ei anna fordeling for tellevariabelen.

Ei fordeling som ofte vert nytta for overdisperserte telledata, er den negative binomiske fordelinga. Denne passar også godt for telledata, og unngår problemet med overdispersjon ved å legge til ein ekstra parameter i variansen. Ei negativ binomisk fordeling ser på antall fiaskoar, Y, før r-te suksess, og er definert som

$$P(Y = y) = \binom{r+y-1}{y} p^r (1-p)^{1-r} \quad y \geq 0. \tag{5.13}$$

Variabel i den negative binomiske fordelinga har forventning $E(Y) = r\frac{1-p}{p^2}$ og varians $\text{Var}(Y) = \frac{r(1-p)}{p^2} = E(Y) + \frac{1}{r}(E(Y))^2$ (Casella og Berger 2002, side 96).

Poisson fordelinga er relatert til den negative binomiske fordelinga som eit grensetilfelle. Når $r \to \infty$ og $p \to 1$ vil $r(1-p) \to \mu$, der $0 < \mu < \infty$, slik at

$$EY = \frac{r(1-p)}{p} \to \mu$$
$$\text{Var}\, Y = \frac{r(1-p)}{p^2} \to \mu$$

der μ er parameter i Poisson fordelinga.

Ein effekt av overdispersjon for ein responsvariabel i ein Poisson GLMM, er at parameterestimata i modellen ikkje vil vere forventningsrette. Ein GLMM basert på ei betinga binomisk fordeling for responsvariabelen derimot, vil i tillegg til å ha mykje av tolkninga til ein Poisson GLMM, ha ein meir fleksibel kovariansstruktur, og slik redusere ein eventuell forventningsskjeivhet i parameterestimata (Booth *et al.* 2003).

I den negativt binomiske fordelinga som eg vil nytte, er det ikkje eit krav at α er eit heiltal. Den følgande definisjonen er ein definisjon i ein artikkel av Booth *et al.*.

Definisjon 5.5.1: Negativ binomisk fordeling for ein tellevariabel

La $V_1, V_2, \ldots, V_n$ vere uavhengig og identisk gamma fordelte variablar med parameter α og fordelingsfunksjon $f(v_i, \alpha)$, slik at

$$f(v_i, \alpha) \quad \propto \quad v_i^{\alpha-1} e^{-\alpha v_i} \cdot I(v_i > 0), \quad i = 1, \ldots, n$$

Anta så at betinga på v_i har tellevariabelen Y_i ei Poisson fordeling med parameter $v_i\mu_i$, på form

$$Y_i \mid v_i \sim \text{Poisson}(v_i\mu_i), \quad i = 1, \ldots, n.$$

Då har tellevariabelen Y_i marginalt ei uavhengig negativ binomisk fordeling med tetthetsfunksjon

$$P(Y_i = y_i, \alpha, \mu_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!} \left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_i} \quad y_i = 0,1,2,\ldots \tag{5.14}$$

Vi skriver då at $Y_i \sim \text{nbin}(\alpha, \mu_i)$.

Forventninga til tellevariabelen er $EY_i = \mu_i$, og varians $\text{Var}\, Y_i = \mu_i + \frac{1}{\alpha}\mu_i^2$. Parameteren α er dermed eit mål på overdispersjon i henhald til Poisson fordelinga, der $\alpha = \infty$ gir $\text{Var}\, Y_i = \mu_i$ som betyr ingen overdispersjon.

For ein longitudinal tellevariabel $\boldsymbol{Y}_i$ av repeterte målingar av $i = 1, \ldots, n$ grupper kan vi formulere ein negativ binomisk GLMM med utgangspunkt den lineære prediktoren i likning (5.4) på side 61 som

$$\begin{aligned} \log(\boldsymbol{\mu}_i) &= \boldsymbol{\eta}_i \\ \boldsymbol{Y}_i \mid \boldsymbol{b}_i &\sim \text{nbin}(\alpha, \mu_i) \\ \boldsymbol{b}_i &\sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Psi}). \end{aligned} \tag{5.15}$$

Som for ein Poisson GLMM er log-funksjonen vanleg å nytte i ein negativt binomiske GLMM. I tillegg til dei faste parametrane i vektoren $\boldsymbol{\beta}$, og elementa i kovariansmatrisa $\boldsymbol{\Psi}$, er parameteren α ein ukjent parameter i ein generell negativt binomisk GLMM.

5.5.1 Negativ binomisk GLMM for Klugman-dataa

Dersom $\boldsymbol{y}_i$ er antall krav i gruppe i, kan vi anta at variabelen $\boldsymbol{y}_i$ er negativt binomisk fordelt betinga på vektoren med variable parametrar $\boldsymbol{b}_i$ ved $\boldsymbol{y}_i \mid \boldsymbol{b}_i \sim \text{nbin}(\alpha, \boldsymbol{\mu}_i)$. Vektoren $\boldsymbol{\mu}_i$ vil då vere forventa antall krav i gruppe i.

Ein negativ binomisk GLMM med random intercept for `Klugman`-dataa kan formulerast som

$$\begin{aligned} \log(\boldsymbol{\mu}_i) &= \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_i \\ \boldsymbol{y}_i \mid b_i &\sim \text{nbin}(\alpha, \boldsymbol{\mu}_i) \\ b_i &\sim \mathcal{N}(0, \psi_0) \\ i &= 1, \ldots, n \end{aligned} \tag{5.16}$$

I denne modellen er dei ukjente parametrane lik, oppsummert i ein vektor θ, $\theta = (\beta_0, \beta_1, \psi_0, \alpha)$, medan $\log(\text{Payroll})_i$ er ein offset-parameter som i Poisson modellen.

Legg vi til ein variabel parameter til stigningstalet, har vi ein negativ binomisk GLMM med to variable parametrar for `Klugman`:

$$\begin{aligned} \log(\boldsymbol{\mu}_i) &= \log(\text{Payroll})_i + \beta_0 + \text{Year}_i \cdot \beta_1 + b_{i0} + \text{Year}_i \cdot b_{i1} \\ \text{der} \quad \boldsymbol{b}_i &= (b_{i0}, b_{i1})^T \\ \boldsymbol{y}_i \mid \boldsymbol{b}_i &\sim \text{nbin}(\alpha, \boldsymbol{\mu}_i) \\ \boldsymbol{b}_i &\sim \mathcal{N}_4(\mathbf{0}, \boldsymbol{\Psi}) \\ i &= 1, \ldots, n. \end{aligned} \tag{5.17}$$

Kovariansmatrisa til vektoren med variable parametrar $\boldsymbol{b}_i$ har elementa

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \psi_{12} \\ \psi_{21} & \psi_2 \end{bmatrix}$$

der $\psi_{12} = \psi_{21}$ og $\psi_2 = 0$ gir modell (5.16). Også her er $\log(\text{Payroll})_i$ ein offset-parameter. Dei ukjente elementa i denne modellen er oppsummert i ein vektor θ lik, $\theta = (\beta_0, \beta_1, \psi_1, \psi_{12}, \psi_2, \alpha)$.

Eg vil no samanlikne estimat i dei to Poisson modellane i likning (5.10) og (5.11) på side 67 med estimat frå dei negative binomiske modellane i likning (5.16) og (5.17) på førre side for `Klugman`. For å tilpasse modellane benyttar eg pakken `glmmADMB`, sjå Skaug *et al.* (2006), og funksjonen `glmm.admb`. Framgangsmåten vises i tillegg C.3.

For funksjonen `glmm.admb`, er metoden `easyFlag` default. Denne er ein raskare, men mindre presis algoritme enn alternativet `easyFlag=F`. Eg vil oppsummere resultata frå begge metodane i tabell 5.5. I tabellen er AIC-verdien til modellane rekna ut ved formelen i definisjon 3.4.1 på side 46. Koden eg har nytta kan studerast i tillegg C.

	Poisson		Negativ binomisk	
	`easyFlag`	`easyFlag=F`	`easyFlag`	`easyFlag=F`
$\lvert\boldsymbol{b}_i\rvert$	1	1	1	1
$\lvert\boldsymbol{\theta}\rvert$	3	3	4	4
β_0	−3,607	−3,613	−3,594	−3,595
β_1	0,014	0,015	0,013	0,013
ψ_0	0,883	0,881	0,877	0,877
α	∞	∞	23,34	23,53
logLik	−2526,26	−2521,83	−2445,14	−2445,02
AIC	5058,5	5049,7	4898,3	4898,0
$\lvert\boldsymbol{b}_i\rvert$	2	2	2	2
$\lvert\boldsymbol{\theta}\rvert$	5	5	6	6
β_0	−3,556	−3,613	−3,585	⊛
β_1	0,006	0,015	0,009	⊛
ψ_1	0,845	0,622	0,872	⊛
ψ_2	0,067	0,000045	0,056	⊛
α	∞	∞	31,70	⊛
logLik	−2591,96	−2521,84	−2437,00	⊛
AIC	5193,9	5053,7	4886,0	⊛

Tabell 5.5: Avrunda parameterestimat til modell med to variable parametrar for `Klugman`-dataa med høvevis Poisson og negativ binomisk fordeling for responsen berekna med funksjonen `glmm.admb`. Kolonnene merka `easyFlag`, har nytta ein raskare, men mindre robust reknealgoritme.

Ved den nøysame metoden, `easyflag=F`, fekk eg ikkje funksjonen `glmm.admb` til å konvergere i estimeringa av parametrane til den negative binomiske modellen med to variable parametrar. Eg har merka plassane med ⊛ , der desse estimata skulle stått. Dermed har eg kun resultatet av estimering ved den mindre nøysame metoden `easyFlag` for denne modellen. Sistnemnte estimering, for modell med

to variable parametrar, gav meg den største verdien av log-likelihood for dei fire modellane, og den minste verdien av AIC.

Vi legg merke til at den nøysame metoden gjev større log-likelihood-verdiar for samtlege av modellane der metoden konvergerte. For Poisson modellen med to variable parametrar observarar vi at den hardføre metoden gjev eit noko større estimat av den faste parameteren til stigningstalet, men eit mykje mindre estimat av variansen til den variable parameteren til stigningstalet, enn det den mindre nøysame modellen gjev. For den mindre nøysame metoden fekk eg for Poisson modellane ei negativ endring i likelihood-verdi ved inklusjon av ein variabel parameter for stigningstalet. Ved den nøysame metoden, fekk eg derimot knapt nokon forskjell i likelihood-verdi for dei to Poisson modellane. Eg observerar også at for den nøysame metoden er estimatet av dei faste parametrane i Poisson modellane uendra ved inklusjon av ein variabel parameter for stigningstalet.

Forskjellen i log-likelihood-verdi er ikkje like stor for random intercept-modellen med negativ binomisk fordeling mellom den meir og den mindre hardføre metoden. Vi ser at den lille auka i log-likelihood for den nøysame metoden gjev utslag i ein noko mindre AIC-verdi.

Funksjonen `glmm.admb` oppgir ikkje noko estimat av ϕ etter tilpassing. Det kan vere at denne blir antatt lik ein for å vere i samsvar med Poisson fordelinga.

Ein hyppig årsak til forventningsskjeive estimat dukkar opp i tilfeller der vi har eit større antall verdiar av null enn det som er konsistent fordelinga vi har lagt til grunn. Longitudinale observasjonar av telledata, som vårt datasett `Klugman`, er ofte utsett for ein høg frekvens av nullverdiar. For slike datasett er det definert ein type modellar der ein antar at det er ein eigen underliggande prosess som avgjer om vi observerar nullverdiar eller verdiar frå fordelinga vi har lagt til grunn. Eg vil no undersøkje tilpassinga av ein slik modell til `Klugman`.

5.6 Nullforhøgde modellar

Modellar som har tilnavnet *zero-inflated* er modellar der ein antar at responsvariabelen har null- og ikkje-nullverdiar som føl ei eiga sannsynsfordeling med parameter ω. Når ein i tillegg har variable parametrar i den lineære prediktoren i modellen, har ein «zero-inflated» blanda modellar, på norsk oversett til nullforhøgde blanda modellar.

Ein nullforhøgd modell, er ein modell der ein antar ei nullforhøgd fordeling for responsvariabelen i modellen. I ei nullforhøgd fordeling, $f(y_i, \theta, \omega_i)$, er variabelen Y_i binomisk fordelt som

$$Y_i \sim \begin{cases} 0 \quad \text{med sannsyn} \quad \omega_i \\ f(y_i, \theta) \quad \text{med sannsyn} \quad 1 - \omega_i \end{cases} \tag{5.18}$$

der fordelinga $f(y_i, \theta)$ er ei sannsynsfordeling, og $0 < \omega_i < 1$ (Hall 2000). Ein seier då at variabelen Y_i har ei nullforhøgd $f(y_i, \theta, \omega_i)$ fordeling. Det er også mulig at $\omega_i < 0$. I såfall har vi ei *nullredusert* fordeling.

Observasjonar av ein nullforhøgd variabel som er lik null, $Y_i = 0$, vil ved likning (5.18) kunne oppstå på to måtar som følge av to ulike underliggande prosessar. Den første prosessen produserar nullar med sannsyn ω_i. Nullar frå denne prosessen kallas *strukturerte* nullar, på engelsk *structural zeros* (Jansakul og Hinde 2002). Den andre prosessen opptrer med sannsyn $1 - \omega_i$, og produserar nullar ifølge sannsynsfordelinga $f(y_i, \theta)$. Nullar produsert på denne måten kallast *observerte* nullar, på engelsk *sampling zeros*. Dermed er sannsynet for verdiar lik null, og sannsynlegheita for verdiar ulik null i ei null-forhøgd fordeling, lik høvevis

$$P(Y_i = 0) = \omega_i + (1 - \omega_i) \cdot f(0, \theta) \tag{5.19}$$

$$P(Y_i = y_i) = (1 - \omega_i) \cdot f(y_i, \theta), \quad y_i = 1, 2 \ldots \tag{5.20}$$

Ved å tilpasse ein nullforhøgd modell til skadeforsikringsdata undersøkjer ein ei hypotese om at ein har observasjonar av kundar med minimal risiko.

For skadeforsikringsdata vil ei nullforhøgd fordeling for responsvariabelen sei at ein antar at nokre av observasjonane stammar frå kundar med minimal risiko (Hall 2000). Ved å tilpasse ein nullforhøgd modell til skadeforsikringsdata kan ein undersøkje ei hypotese om at nokre av kundane i datasettet har minimal risiko, medan andre kundar har ein kravsfrekvens som følger ei sannsynsfordeling $f(y_i, \theta)$. Eg ynskjer å undersøkje ei slik hypotese for `Klugman`-dataa. Tabell 5.6 på neste side viser antall verdiar av variabelen Count_i i `Klugman` som er lik 0 og ikkje.

Det ser ut til at andelen krav lik null er rimeleg konstant gjennom dei sju åra observasjonane er gjort. Andelen observasjonar av null krav ligg jamnt på $1/5$ av totalen. Eg vil undersøkje om ein zero-inflated Poisson GLMM, forkorta ZIP, og ein zero-inflated negativ binomisk GLMM, forkorta ZINB kan passe `Klugman`-dataa. På

År	$\text{Count}_i = 0$	$\text{Count}_i > 0$
1	21	105
2	20	107
3	23	105
4	25	105
5	21	107
6	21	107
7	21	107

Tabell 5.6: Registrerte verdiar av responsvariabelen Count som er lik null og større enn null.

norsk kan desse oversetjast til ein nullforhøgd Poisson modell, og ein nullforhøgd negativ binomisk modell. Eg nyttar likevel dei engelske forkortingane.

5.6.1 Ein ZIP-modell

I tilpassinga av ein negativ binomisk GLMM var eg interessert i å undersøkje om variabelen Count_i kunne forklarast betre ved å anta denne fordelinga, sidan den negative binomiske fordelinga tillet ein større varians enn Poisson fordelinga. Ved å tilpasse ein nullforhøgd Poisson GLMM derimot, ynskjer eg å undersøkje om avviket frå Poisson fordelinga skuldast ein for høg andel av antall krav lik null.

I ein ZIP-modell antar vi at responsvariabelen har ei nullforhøgd Poisson-fordeling slik at vi observerer høvevis $Y_i = 0$ og $Y_i = y_i > 0$ med sannsyna

$$P(Y_i = 0) = \omega_i + (1 - \omega)e^{-\mu_i}$$
$$P(Y_i = y_i) = (1 - \omega_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 1, 2, \ldots,$$

der $0 < \omega_i < 1$ er ukjend, og μ_i er parameteren i Poisson fordelinga (Jansakul og Hinde 2002).

Ein blanda ZIP-modell vil dermed, i tillegg til dei faste parametrane og variansen til dei variable parametrane, ha ω_i som ukjent parameter. Forventning til ein ZIP fordelt variabel, $y_i \sim \text{Po}(\mu_i, \omega)$, vil i følge Jansakul og Hinde vere definert som

$$E(y_i) = (1 - \omega_i)\mu_i = \tilde{\mu}_i,$$

og varians ved

$$\begin{aligned}\text{Var}(y_i) &= E(y_i) + E(y_i)\left[\mu_i - E(y_i)\right] \\ &= \tilde{\mu}_i + \frac{\omega_i}{1-\omega_i}\tilde{\mu}_i^2.\end{aligned}$$

Dersom $\omega_i = 0$ så har vi ei ordinær Poisson fordeling. Ved tilpassing av ein ZIP-modell ynskjer ein derfor å undersøkje om $\omega_i > 0$, i tillegg til å inferere om parametrane i den lineære prediktoren.

5.6.2 Ein ZINB-modell

I ein ZINB-modell definert for ein tellevariabel, antar vi at responsvariabelen kjem frå ei nullforhøgd negativ binomisk fordeling, og inntek verdiane $y_i = 0$ og $y_i > 0$ ved sannsyna

$$\begin{aligned}P(Y_i = 0) &= \omega_i + (1-\omega_i)\cdot\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha} \\ P(Y_i = y_i) &= (1-\omega_i)\cdot P(Y_i = y_i, \alpha, \mu_i) \quad y_i = 1, 2, \ldots,\end{aligned} \tag{5.21}$$

der sannsynet $P(Y_i = y_i, \alpha, \mu_i)$ er uttrykt i likning (5.14) på side 73, og uttrykket $\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}$ fåast ved å sette $y_i = 0$ i denne likninga. Som i ZIP fordelinga er forventninga til ein nullforhøgd negativ binomisk variabel lik $E(y_i) = (1-\omega_i)\mu_i$. Variansen derimot har eit lengre uttrykk som eg ikkje tek med her.

5.6.3 Inferens for nullforhøgde modellar

Eg ynskjer å undersøkje om ei ZIP eller ei ZINB fordeling passar for responsvariabelen Count_i i `Klugman`.

Likelihood-funksjonane for nullforhøgde fordelingar vil inkludere sannsynsparameteren ω_i. Dersom vi kun er interesserte i å teste om parameteren ω_i er større enn null, det vil seie at vi utelukkar muligheita for ein nullredusert modell, vil vi teste ei hypotese på form

$$H_0 : \omega_i = 0 \quad \text{mot} \quad H_1 : \omega_i > 0. \tag{5.22}$$

Her vil, som i hypotesa i likning (3.17) på side 43, parameteren ω_i ligge på randa av definisjonsområdet under H_0. Dermed vil ein likelihood-ratio basert på log-likelihooden til den ordinære og den nullførhøgde modellen ha ei fordeling som avviker frå den generelle fordelinga for ein slik testobservator som i følge likning (3.16) på side 41 ville vore χ^2_1. Jansakul og Hinde hevdar at testobservatoren for hypotesa (5.22) på førre side har ei $0.5\chi^2_0 + 0.5\chi^2_1$-fordeling.

5.6.4 Analyse av nullforhøgde modellar

I R kan nullforhøgde modellar analyserast av funksjonen `glmm.admb` i pakken `glmmADMB`, som eg nytta til å anaysere ein negativ binomisk GLMM med. Eg viser til tillegg C.3 for kode. Tabell 5.7 oppsummerar parameterverdiar til ZIP og ZINB-modellar med høvevis éin og to variable parametrar.

	ZIP		ZINB	
	`easyFlag`	`easyFlag=F`	`easyFlag`	`easyFlag=F`
$\lvert\boldsymbol{b}_i\rvert$	1	1	1	1
$\lvert\boldsymbol{\theta}\rvert$	4	4	5	6
ω_i	6,54$*$	5,86$*$	4,05$*$	4,05$*$
β_0	$-3{,}607$	$-3{,}613$	$-3{,}594$	$-3{,}595$
β_1	0,014	0,015	0,013	0,013
ψ_0	0,883	0,881	0,877	0,877
α	∞	∞	23,334	23,526
logLik	$-2526{,}25$	$-2521{,}82$	$-2445{,}12$	-2445
AIC	5060,5	5051,6	4900,24	4900
$\lvert\boldsymbol{b}_i\rvert$	2	2	2	2
$\lvert\boldsymbol{\theta}\rvert$	6	6	7	7
ω_i	$\circledast$	5,86$*$	3,57$*$	4,04$*$
β_0	$\circledast$	$-3{,}613$	$-3{,}585$	$-3{,}595$
β_1	$\circledast$	0,015	0,009	0,013
ψ_1	$\circledast$	0,622	0,872	0,620
ψ_2	$\circledast$	0,00004	0,056	0,00004
α	∞	∞	31,697	23,526
logLik	$\circledast$	$-2521{,}83$	$-2436{,}97$	-2445
AIC	$\circledast$	5055,7	4887,9	4904

Tabell 5.7: Parameterverdiar til nullforhøgde modellar for Klugman data. Verdiane merka * er i skalaen 10^{-6}.

For ZIP-modellen med random intercept ser eg at estimatet av ω_i er svært lite, og at endringa i likelihood-verdi er, i høve til Poisson modellen som ikkje er nullforhøgd,

omlag lik null. For ZINB-modellen med random intercept er estimatet av ω_i også svært lite, og endringa i likelihood-verdi i høve til modellen utan nullforhøgd fordeling for responsen er svært liten.

Ein hypotesetest i R på parameteren ω_i, gjev resultatet

```
> anova(modNB.yr,nb2.zip)

Analysis of Variance Table

Model 1: Count ~ Year
Model 2: Count ~ Year
     NoPar    LogLik Df   -2logQ  P.value
1     5.00 -2437.00
2     6.00 -2436.97  1     0.06     0.81
```

Ein verdi på 0,06 av likelihood ratioen vil heller ikkje vere signifikant i ei $0.5\chi^2_0 + 0.5\chi^2_1$-fordeling. Dermed kan vi med rimeleg sikkerheit slå fast at variabelen Count ikkje har ei nullforhøgd fordeling.

For dei ZIP og ZINB-modellen med to variable parametrar er det kun ZINB-modellen eg får `glmm.admb`-funksjonen til å konvergere for. Denne har ein noko betre likelihood enn ZINB-modellen med éin variabel parameter.

Likevel er det klart frå desse analysene at ein nullforhøgd GLMM ikkje passar betre enn ein ordinær GLMM for variabelen Count i `Klugman`-data.

5.7 Konklusjon Klugman

Eg oppsummerar log-likelihood- og AIC-verdiar for modellane for `Klugman` eg har analysert i dette kapittelet i tabell 5.8 på neste side. Eg oppgjev estimata eg fekk ved den hardføre metoden for alle modellane med unntak av modellen med to variable parametrar og negativ binomisk fordeling for responsen i likning (5.17). For denne har eg kun estimata eg fekk ved den mindre hardføre metoden, og desse har eg ført inn i tabellen og ført opp ein `T` i kolonna `easyFlag`.

Av tabellen ser vi at modellane med Poisson fordeling for responsvariabelen Count_i jamnt over har større AIC-verdiar enn modellane med negativ binomisk fordeling.

Eg ynskjer ikkje å samanlikne verdiar funne i programmet `lme4` med verdiar funne i `glmmADMB`, sidan eg ikkje er sikker på om like antakingar er lagt til grunn i dei to programma. Sjølv om Antonio og Beirlant (2006) fekk....

	$\lvert b_i \rvert$	$\lvert \theta \rvert$	logLik	AIC	easyFlag
Poisson GLMM	1	3	−2521,83	5049,7	F
Neg. bin. GLMM	1	4	−2245,02	4898,0	F
Poisson GLMM	1	5	−2521,84	5053,7	F
Neg. bin. GLMM	2	6	−2437,0	4886,0	T
ZIP	1	4	−2521,82	5051,6	F
ZIP	2	6	−2521,83	5055,7	F
ZINB	1	5	−2445	4900	F
ZINB	2	7	−2445	4904	F

Tabell 5.8: Verdiar av observatorar for modellane eg har tilpassa `Klugman`-dataa ved funksjonen `glmm.admb`.

Vi ser av denne tabellen at modellen med negativ binomisk fordeling og to variable parametrar har fått den største likelihooden og best AIC-verdi.

Den beste verdien ved den hardføre metoden i `glmm.admb`, fekk eg for modellen med ein variabel parameter og ei negativ binomisk fordeling. Eg legg merke til at den hardføre metoden gjev liten auke i likelihood-verdi ved inklusjon av ein variabel parameter for stigningstalet for samtlege av modellane der metoden konvergerte for begge modellane. For Poisson modellen med to variable parametrar ser eg at likelihood blir meir negativ etter inklusjonen. Dette er i strid med tidlegare antakingar som eg har lagt til grunn. For to modellar med høvevis p og q parametrar der $p > q$ vil likelihooden til modellen med p parametrar L_2 vere større eller lik likelihooden til modellen der nokre av desse parametrane er sett lik null, L_1. Altså $L_2 \geq L_1$ når $|\theta_2| \geq |\theta_1|$. Dette tyder på at den mindre hardføre metoden moglegeins ikkje alltid er til å stole på.

Konklusjonen min er med desse resultata at ein GLMM med ei negativ binomisk fordeling for responsen og éin variabel parameter lagt til konstantleddet i den lineære prediktoren, er den beste modellen for variabelen Count_i i datasettet `Klugman`.

I det neste kapittelet vil eg oppsummere resultata eg har kome fram til for lineære blanda og generaliserte lineære blanda modellar.

The world is round and the place which may seem like the end may also be only the beginning.

Ivy Baker Priest, in Parade, 1958

6 Konklusjonar

I dei føregåande kapittela sett på korleis ein kan tilpasse modellar til longitudinelle observasjonar som ein antar er normalfordelte eller har ei anna fordeling i eksponensialfamilien. Desse modellane let ein ta høgd for at observasjonar på same eksperimentelle eining er korrelerte, men samstundes at dei har ein felles påverknad på responsen.

I ein lineær blanda modell kan ein studere både eit populasjonssnitt og individuelle modellar. I kapittel 2 viste eg korleis kovariansstrukturen til ein modell som inneheld variable parametrar kan ta høgd for individuell korrelasjon. I definisjonen av desse modellane kjem det fram at det er variansen til dei variable parametrane vi ynskjer å estimere, sidan denne vil fortelje oss om det er variasjonar i storleiken til parametrar mellom einingar. Når ein skal undersøkje om variansen til ein variabel parameter er ulik null, vil ein måtte velje om ein godtek at korrelasjonen mellom målingar på same individ kan vere negativ, eller om ein antar at denne er strengt positiv. Dersom vi antar at korrelasjonen er strengt positiv, vil ein test for å undersøkje om variansen er lik null, setje verdien av varianselementet vi undersøkjer på randa av definisjonsområdet sitt. I ein slik test såg eg i kapittel 4 at testobservatoren ikkje lenger har den generelle fordelinga til ein likelihood ratio-observator som eg definerte i likning (3.16) på side 41.

I definisjonen av dei blanda modellane i kapittel 5, der ein ikkje lenger avgrensar seg til å sjå på normalfordelte responsvariablar, støtte eg på endringar i formuleringa i høve til den lineære blanda modellen i kapittel 2. For det første kan ein GLMM modellere responsvariablar der ei normalfordeling ikkje lenger høver. For det andre kan ein i ein GLMM modellere den lineære prediktoren ved ein transformasjon av responsvektoren. Sist, men ikkje minst, skil ein GLMM seg frå ein ordinær GLM og ein LMM, ved å vere definert med si betinga fordeling, noko som vi såg i kapittel 2 at tilsvara ei individsretta tolkning. Vi såg også at den lineære blanda modellen er eit spesialtilfelle av ein GLMM, der den betinga fordelinga til responsen er normalfordelinga, og link-funksjonen er identitetsfunksjonen. Dette svarar til modellen som Molenberghs og Verbeke (2004) kalla ein betinga modell. I ein GLMM har ein dermed ikkje lenger ei populasjonsretta tolkning slik vi hadde i den lineære blanda modellen. Eg har, ved sidan av GLMM-ar med høvevis Poisson og negativ binomisk fordeling for responsvariabelen, sett på modellar der ein undersøkjer om nullverdiane i datasettet, i tillegg til å dukke naturleg opp som følge den antekne fordelinga, blir påverka av ein eigen prosess. Slike modellar vil ha endå ein parameter inkludert, nemleg sannsynet ω for at denne prosessen skal generere ein nullverdi.

I undersøkjinga av den beste modellen for observasjonane i `Orthodont`, var det konklusjon omkring fordelinga til testobservatoren til modellane som gav hovudbry. I prosessen med å finne modellen som høvde best for observasjonane av talet på krav i `Klugman`, var å avgjere fordelinga til responsvariabelen viktig å avgjere. I dei kommande avsnitta vil eg oppsummere mine konklusjonar for analysene i dei to datasetta, samt kva for problem eg støytt på, både teoretiske og praktiske, i analysene av blanda modellar.

6.1 Analyse av ein ortopedisk avstand

I analysen av dei to lineære blanda modellane i likning (2.4) og (2.6) på side 22, støtte eg på problemet med å avgjere fordelinga til ein testobservator som har parametrar på randa av definisjonsområdet sitt. Denne problemstillinga var starten på temaet som skulle bli masteroppgåva mi, nemleg analysar av blanda modellar i R. Ved starten av mine analyser var eg ukjend med artikkelen til Stram og Lee (1994), der den same simuleringa er blitt gjort. Etter å ha blitt gjort merksam på artikkelen, har eg ved jamføring sett at mine resultat er nokså nærme resultata til Stram og Lee. Stram og Lee fekk ved si simulering at 485 av 1000 simulerte ratioar var større enn den observerte verdien $\hat{\mathbf{\Lambda}} = 0{,}833$. Til samanlikning gjev

mitt resultat på 464 av 1000 ratioar større eller lik 0,833, ein noko mindre p-verdi enn det Stram og Lee får. Stram og Lee (1994) omtalte i samband med sine simuleringar at den 50:50 vekta fordelinga i likning (4.2) på side 52 høver best for ein likelihood ratio for to blanda modellar. Eg såg i plottet i figur 4.5 på side 54 at ei $0{,}65\chi^2_1 + 0{,}35\chi^2_2$-fordeling høvde best for mi simulerte tettleik. Men sidan det er få observasjonar eg har estimert modellverdiane mine for, vil grunnlaget mitt for å simulere fordelinga ikkje vere så solid og eg kan ikkje forvente at resultatet stemmer fullstendig med teorien til Stram og Lee, som er ein asymptotisk teori.

Problem undervegs har gjort at min simuleringsprosess sjølv har blitt påverka av tilfeldigheiter. Mellom anna, som eg nemner i tillegg B, fekk eg problem med å køyre simuleringsfunksjonen min i nyare versjonar av R. Under oppdateringa av maskinene, som dessverre skjedde midt i det siste semesteret for oppgåva mi, forsvann den gamle versjonen av R som gjorde at eg ikkje trong tilleggsfunksjonen `try`. Ved simulering i versjon `2.4.0` får eg svært mange `NA`-verdiar med funksjonen min, og resultatet mitt blir ikkje i samsvar med det Stram og Lee (1994) får.

6.1.1 Resultata mine

Eg har valgt å la simuleringane mine i den gamle versjonen av R stå som resultata mine. Desse er også mest like Stram og Lee (1994) sine. Under simuleringa av RLR hadde eg ikkje lenger tilgang på den gamle versjonen av R. Desse er derfor simulerte med eit program som gav mange `NA`-verdiar.

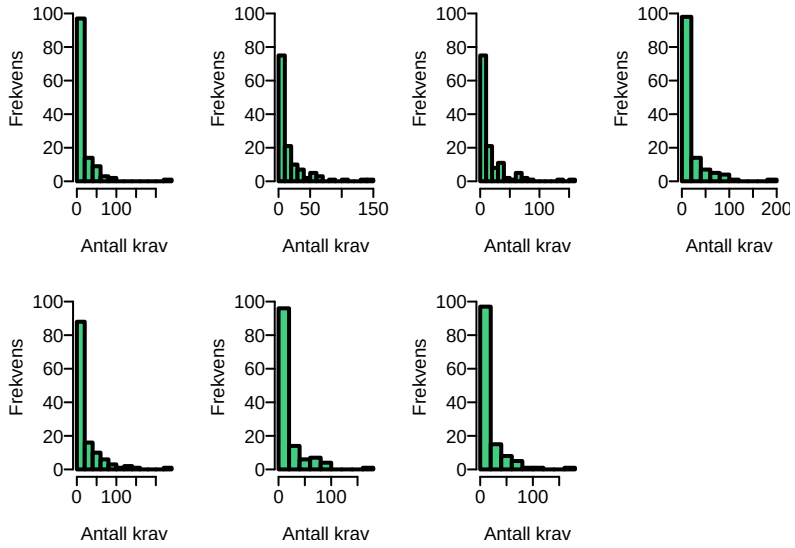
Simuleringa av LR til testobservatoren til hypotesa i likning (3.17) på side 43 gav meg innsikt i kva det medfører at ein testobservator inneheld ein verdi av ein parameter som er på grensa av definisjonsområdet sitt. For observasjonane i `Orthodont`, fann eg som nemnt at ei $0{,}65\chi^2_1 + 0{,}35\chi^2_2$-fordeling for testobservatoren til hypotesa $H_1 : \psi_2 > 0$. Sjølv om ikkje verdien av testobservatoren min blei signifikant i denne fordelinga, så blei p-verdien i forsøket mitt synleg mindre enn ved samanlikning med den konservative fordelinga χ^2_2. Eg kan dermed ikkje forkaste modellen med éin variabel parameter for `Orthodont`.

6.2 Analyse av talet på krav

Når fordelinga til ein respons ikkje lenger avgrensar seg til vere antatt normal, vil eit nytt landskap av mogelegheitar åpne seg. I definisjonen av ein GLMM, vil dei

mange mogelegheitene for fordeling og link-funksjon til responsen føre til at å finne den beste modellen for eit datasett er krevjande.

Eg har sett på datasettet `Klugman`, som med observasjonar av talet på krav, fell inn i ei stor gruppe tilfeller der ein ikkje kan nytte normalantaking for responsen. Det er kjent frå elementære statistikkurs at telle-variablar antas å ha ei Poisson fordeling. Men i tilfeller der ein har overdispersjon vil ikkje den generelle Poisson fordelinga der forventning er lik varians høve lenger. Eg såg for `Klugman` at nettopp dette var tilfellet. I litteraturen som omtalar blanda modellar for telledata, er det foreslått minst to mogelegheitar for å oppnå betre tilpassa modellar for slike data. Ei av dei er å nytte ei anna fordeling for responsen, og ei anna er å undersøkje om det finnes ein eigen prosess som er med på å generere nullverdiar. Slike modellar er aktuelle for skadedata fordi det tilsvarar å undersøkje om ein har observasjonar av svært «gode» risikoar, samt observasjonar av risikoar frå ein generell populasjon. Eg valde å studere tilsaman åtte modellar for `Klugman`-dataa. Fire av desse antok ei nullforhøgd fordeling, medan dei fire første antok høvevis ei Poisson og ei negativ binomisk fordeling for responsvariabelen.



Figur 6.1: Frekvenshistogram av det totale talet på krav dei sju åra i settet `Klugman`.

6.2.1 Resultat

Eg kom ved å studere verdiane i tabell 5.8 på side 82 fram til ein GLMM med negativ binomisk fordeling for responsen og éin variabel parameter høvde best til å modellere talet på krav for ulike risikogrupper i `Klugman`.

Om eg jamfører resultata frå analysene av dei negative binomiske modellane utan nullforhøgd fordeling i tabell 5.5 på side 75 med Antonio og Beirlant (2006) sine resultat i tabell 5.4 på side 71, ser eg at estimata av dei faste parametrane er nokså like. Eg legg merke til at kolonna med Bayes-estimat hos Antonio og Beirlant er dei som liknar mest på estimata av dei faste parametrane til den negative binomiske modellen.

Antonio og Beirlant (2006) kom i sin artikkel fram til at Poisson modellen med éin variabel parameter passa best for `Klugman`-dataa. Analyser i `glmmADMB` viser at mellom dei to Poisson modellane er modellen med éin variabel parameter den beste. Til kontrast gjev analyser i `lme4` ein motsett konklusjon. Der får modellen med to variable parametrar ein klart større likelihood. Det hadde vore interessant å vite kva som gjer at eg får ulike konklusjonar ved bruk av funksjonar som prøver å rekne ut det samme.

Fordi dataa ikkje held antakinga om $\phi = 1$ i Poisson fordelinga, er det rettmessig å undersøkje om ei negativ binomisk fordeling gjev betre resultat. Når, ved jamføring, ei negativ binomisk fordeling for responsen høver betre til dataa enn ei Poisson fordeling analysert ved same program, vil eg konkludere med at førstnemnte fordeling høver dataa best. Men sidan den hardføre metoden ikkje konvergerte i analysa av den negative binomiske modellen med to variable parametrar, kan eg ikkje utelukke at denne høver betre til å modellere talet på krav. Å få funksjonen `glmm.admb` til å konvergere for denne vil vere eit emne for vidare arbeid.

6.3 Til slutt

Eg vil avslutte dette kapittelet med å kommentere praktiske problem som dukka opp undervegs i analysene mine. Eg nemner i tillegg B nokre problem som dukkar opp når eg skal nytte simuleringsfunksjonen min `sim2` i nyare versjonar.

Eit anna problem omlegginga av datasystemet medførte var at versjonen av R som då var tilgjengeleg hadde gjort oppdateringar av `lme4`-pakken og `lmer`-funksjonen som eg nytta til å finne estimat av Poisson-GLMM-ar i tabell 5.4 på side 71. I den nye versjonen er kun metoden Laplace tilgjengeleg. Denne er dog den mest hardføre. Men for meg betydde oppdateringa at eg ikkje har fått dobbelsjekka verdiane eg fekk ved å nytte `method=«ML»` i den gamle versjonen av `lmer`.

Inventions reached their limit long ago,
and I see no hope for further development.

Julius Frontinus (det første hundreåret e.Kr.)

7 Vidare arbeid

I dette kapittelet vil eg presentere ei analyse som vil vere interessant å studere vidare. I analysa settes det fokus på å estimere den ukjende fordelinga eit datasettet er henta frå, og slik kunne bygge ein observator som betre vil kunne avgjere kva for ein modell som høver best til eit datasett.

7.1 Bootstrapping

Det engelske uttrykket *bootstrapping* er av statistikarar blitt eit navn på ei prosedyre som, med utgangspunkt i eit bestemt datasett, utfører ei «ny» innsamling av data ved å trekke tilfeldige observasjonar frå datasettet med tilbakelegging. Ein ynskjer ved bootstrapping å finne ut meir om fordelinga datasettet vårt er henta frå, ved å studere korleis dataa våre kan variere som følge av eit tilfeldig utval. Ordet «bootstrap» kan oversetjast til støvleskaft, og uttrykket «bootstrapping» er henta frå historia om Baron von Münchhausen, sjå øvst på side 106.

Som eg nemnte i kapittel 3 er kriteriet AIC ikkje så presist i å vurdere tilpassinga for blanda modellar. Mange stastistikarar har prøvd å forbetre informasjonskriteriet ved å, istadenfor å nytte talet på parametrar som ei justering for forventnings-skjeivheit, nytte bootstrapping for å estimere skjeivheita. Bakdelen med desse

observatorane er at dei krev ei meir omstendig implementering. Men det vert hevda at for små datasett, vil ei evaluering av desse bootstrap-kriteria kunne vere essensielt for å komme fram til den sanne konklusjonen. I ein artikkel av Yafune *et al.* (2005), blir datasettet `Orthodont` trekt fram som eit døme på at eit bootstrap-kriterium gjev ein annan konklusjon enn Akaike sitt kriterium gjev.

Informasjonskriterier eg har støtt på som nyttar bootstrap i sine kalkuleringar er TIC og GIC som er kort for «Takeuchi's information criterion» og «generalized information criterion» (Konishi og Kitagawa 1996), og eit kriterium kalla EIC som er kort for «extended information criterion». Sistnemnte kriterium er blitt nytta av Yafune *et al.* (2005) for modellar definert for guttane i `Orthodont`, og dei kom fram til at ein modell med to variable parametrar, både for konstantledd og stigningstal, høvde best. Dette har inspirert meg sjå litt nærare på kriteriet EIC.

7.2 Extended information criterion

Observatoren EIC bygger på eit utvida AIC, derav navnet *extended information criterion*, der ein nyttar bootstrapmetodar for å estimere forventningsskjeivheita i AIC-definisjonen. Dette kriteriet er blitt studert av mellom anna Yafune *et al.*, som hevdar at EIC er ein meir nøyaktig observator også i tilfeller der ML er blitt nytta, særleg dersom ein har få observasjonar.

Yafune *et al.* (2005) ynskjer, istadenfor å nytte tilnærminga $C = |\boldsymbol{\theta}|$ som Akaike har gjort, å utføre ei bootstrap-prosedyre for å estimere skjeivheita C i likning (3.22) på side 45. Dei definerar med dette observatoren EIC som

$$\text{EIC} = -2l(\boldsymbol{\theta}, \boldsymbol{y}) + 2\hat{C}^* \tag{7.1}$$

der det første leddet er log-likelihood verdien til modellen basert på datasettet og $\hat{C}^*$ eit bootstrap-estimat av skjeivheita i likning (3.22).

7.2.1 Ei bootstrappingsprosedyre

Eg har prøvd å nytte Yafune *et al.* (2005) si skildring av metoden samt funksjonen `boot` i R til å utføre ei bootstrapping av skjeivheita i den forventa log-likelihooden i likning (3.21) på side 45, men eg fekk svært ustabile resultat. Eg har lagt ved kode for det eg gjort i tillegg D.

Yafune *et al.* (2005) forklarar si prosedyre som følgjer:

For eit datasettet $\boldsymbol{y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ trekk ein B uavhengige tilfeldige observasjonar slik at ein får B «nye» datasett $\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_B^*$.

Basert på desse «nye» datasetta, $\boldsymbol{y}_b^* = \boldsymbol{y}_{b1}, \ldots, \boldsymbol{y}_{bn}$, $b = 1, \ldots, B$, estimerar ein skjeivheita $\hat{C}^*$ ved

$$\hat{C}^* = \frac{1}{B} \sum_{b=1}^{B} \left[\sum_{i=1}^{n} \log h(\boldsymbol{y}_{bi}^* \mid \hat{\boldsymbol{\theta}}^*) - \sum_{i=1}^{n} \log h(\boldsymbol{y}_i \mid \hat{\boldsymbol{\theta}}^*) \right]$$

der vektoren $\hat{\boldsymbol{\theta}}^*$ er eit REML estimat av parametervektoren $\boldsymbol{\theta}$ basert på bootstapdataa $\boldsymbol{y}_b^*$. Som for AIC vil ein modell med liten EIC-verdi sjåast på som å passe betre til dataa enn ein modell med større EIC. Tabell 7.1 viser resultat eg fekk ved å nytte programmet i tillegg D.

	$b = 100$	$b = 500$	$b = 750$	$b = 1000$
Verdiar	29,079	−13,325	−35,004	−7,574
av C_1^*	16,324	−2,668	4,624	−10,108
	9,667	20,116	7,383	1,297
	0,372	25,204	−9,805	
	31,580	8,824	−4,978	
	−29,619	1,822	−10,788	
	−16,860	25,643	2,832	
	−1,306	21,708	−9,499	
	−15,080	5,239	4,010	
	30,080	13,791	−6,789	
Snitt	5,424	10,27118	−5,801	
EIC	444,604	454,299	422.1545	

Tabell 7.1: EIC-verdiar for modell fit1.

Problemet mitt med å utføre denne simuleringa var at eg ikkje heilt kom til bunns i kva funksjonen `boot` gjorde for meg. Det var dermed vanskeleg å tolke det som kom ut av funksjonen. Eg vil ikkje kommentere resultata mine så mykje, for eg tvilar på at det er meininga at estimatet av C skal variere så mykje som i tabell 7.1. Men det som er noko oppsiktsvekkande for meg er at desse verdiane av EIC ikkje er så langt unna AIC-verdien eg fekk i tabell 2.1 på side 24, som er AIC = 440,639.

Men å nytte bootstrapping for å bestemme den ukjende fordelinga eit datasett er henta frå, verkar for meg som ein nyttig metode for modellanalyse, særleg i tilfeller der ein har få observasjonar. Eg innser at mi analyse av blanda modellar er endå i byrjinga, og at somme av metodane som nyttast for å analysere blanda

modellar er over mitt kompetansenivå. Likevel er eg blitt svært interessert i å kunne forstå meg på nye metodar, som EIC, for å sjekke godheita til ein modell tilpassa eit datasett. Eg tvilar på at dette er mitt siste møte med blanda modellar, og deira krav til testobservatorar. Særleg håpar eg å kunne undersøkje nærare nytta av blanda modellar i skadeforsikring.

It would appear that we have reached the limits of what it is possible to achieve with computer technology, although one should be careful with such statements, as they tend to sound pretty silly in 5 years.

John Von Neumann (omlag 1949)

A

Analysar av modellar for Orthodont

A.1 Lineære modellar

Den ordinære lineære modellen for `Orthodont`, tilpassar eg i R ved funksjonen `lm`. Modellen med både alder og kjønn som forklaringsvariablar, i likning (1.6) på side 10, formulerar eg som i utskrifta under. Eg nyttar kommandoen `summary` for å få utskrift av estimerte modell parametre.

```
> lm2 <- lm(distance~age*Sex,data=Orthodont)
> summary(lm2)

Call:
lm(formula = distance ~ age * Sex, data = Orthodont)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6156 -1.3219 -0.1682  1.3299  5.2469

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.3406     1.4162  11.538  < 2e-16 ***
age            0.7844     0.1262   6.217 1.07e-08 ***
SexFemale      1.0321     2.2188   0.465    0.643
```

```
age:SexFemale  -0.3048     0.1977  -1.542    0.126
---
Signif. codes:  0 "***" 0.001

Residual standard error: 2.257 on 104 degrees of freedom
Multiple R-Squared: 0.4227,     Adjusted R-squared: 0.4061
F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

Vi ser av utskrifta at variabelen kjønn har stor p-verdi. Ved å fjerne denne får eg at interaksjonsleddet er signifikant.

```
> fit0s <- lm(distance~age*Sex - Sex,data=Orthodont)
> summary(fit0s)

Call:
lm(formula = distance ~ age * Sex - Sex, data = Orthodont)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7424 -1.2424 -0.1893  1.2681  5.2669

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.76111    1.08613  15.432  < 2e-16 ***
age            0.74767    0.09807   7.624 1.16e-11 ***
age:SexFemale -0.21473    0.03923  -5.474 3.02e-07 ***
---
Signif. codes:  0 "***" 0.001

Residual standard error: 2.249 on 105 degrees of freedom
Multiple R-Squared: 0.4215,     Adjusted R-squared: 0.4105
F-statistic: 38.26 on 2 and 105 DF,  p-value: 3.31e-13
```

Men ei analyse med kommandoen `anova`, gjev oss ikkje grunnlag til å forkaste modell (1.6).

```
> anova(fit0,fit0s)

Analysis of Variance Table

Model 1: distance ~ age * Sex
Model 2: distance ~ age * Sex - Sex
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1    104 529.76
2    105 530.86  -1     -1.10 0.2164 0.6428
```

Derfor konkluderar eg i kapittel 1 med at modellen med både alder, kjønn og interaksjonsleddet mellom alder og kjønn, passar best til `Orthodont` av dei ordinære regresjonsmodellane.

A.2 Blanda modellar

Får å kunne analysere blanda modellar i R har eg nytta pakken `nlme` oppretta av Pinheiro og Bates (2000). I denne pakken ligg funksjonen `lme` som kan ta inn regresjonsmodellar med random effects-ledd.

Eg spesifiserar metoden som eg vil nytte til å estimere parametrane i den blanda modellen ved argumentet `method=«.»`. Eg ynskjer først å nytte maksimum likelihood som, og skriv då `method=«ML»`. Alternativet er metoden restricted maksimum likelihood, REML, som er default for `lme`. Random intercept-modellen i likning (2.4) på side 20, tilpassar eg ved

```
> fit1 <-
lme(distance~age*Sex,data=Orthodont,random=~1|Subject,method="ML")
```

og modellen med to random effects-ledd i likning (2.6) på side 22 ved

```
> fit2 <- update(fit1, random=~age|Subject)
```

Utskrift av dei estimerte parameterverdiane får eg ved å nytte

```
> summary(fit2)

Linear mixed-effects model fit by maximum likelihood
 Data: Orthodont
       AIC     BIC    logLik
  443.8060 465.263 -213.9030

Random effects:
 Formula: ~age | Subject
 Structure:
General positive-definite,Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.1346882 (Intr)
age         0.1541390 -0.603
Residual    1.3100396

Fixed effects: distance ~ age * Sex
                 Value Std.Error DF   t-value p-value
(Intercept)   16.340625 0.9987521 79 16.361042  0.0000
age            0.784375 0.0843294 79  9.301321  0.0000
```

```
SexFemale      1.032102 1.5647438 25  0.659598  0.5155
age:SexFemale -0.304830 0.1321188 79 -2.307238  0.0237
 Correlation:
              (Intr) age    SexFml
age           -0.880
SexFemale     -0.638  0.562
age:SexFemale  0.562 -0.638 -0.880

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.3360336 -0.4153984  0.0103917  0.4916952  3.8581929

Number of Observations: 108
Number of Groups: 27
```

Dette er utskrifta for modellen med både stokastisk konstantledd og stigningstal. Ein kan for denne få utskift av matrisa $\mathbf{\Psi}$ ved å nytte kommandoen

```
> getVarCov(fit2)

Random effects variance covariance matrix
            (Intercept)       age
(Intercept)     4.55690 -0.198250
age            -0.19825  0.023759
  Standard Deviations: 2.1347 0.15414
```

som er ein del av `nlme`-pakken sine kommandoar. I `summary(fit2)`, vises standandavviket til varianselementa.

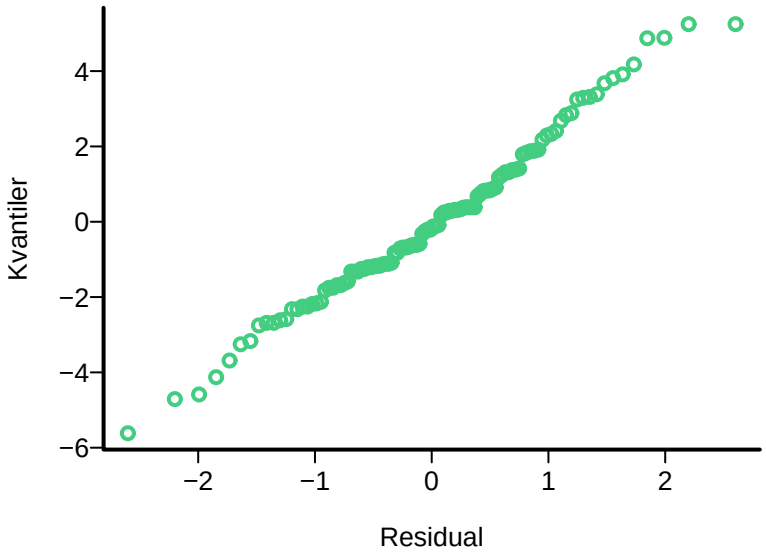
A.3 Plott av residual

For den ordinære regresjonsmodellen vil kommandoen

```
> plot(fit0)
```

gje oss fire ulike typar residualplott som vist i figur 3.1 på side 29. Som eg nemnte kjem ikkje grupperingane av residuala så godt fram i QQ-plottet i denne figuren. Ved å nytte funksjonen

```
> qqnorm(fit0$residuals)
```



Figur A.1: Eit plott av standardiserte residual for den ordinære regresjonsmodellen.

vil ein få eit plott kun av standariserte residual mot kvantiler i standard normalfordelinga, og dette plottet, i figur A.1, viser grupperingane betre.

For dei blanda modellane vil `plot`-kommandoen gje kun eit plott av residuala. I dette plottet er ikkje uteliggarar merka av. Får å identifisere uteliggarar må ein legge til argumentet `id`. Talverdien står for signifikansnivået (Pinheiro og Bates 2000, side 188).

```
> plot(fit1,form=resid(.,type="p")~fitted(.), id=0.05)
```

Denne kommandoen gav meg figur 3.2 på side 30. Eit QQ-plot for blanda modellar, der eventuelle uteliggarar er merka av, fåast ved kommando

```
> qqnorm(fit1, ~resid(.),id=0.05)
```

Dette gav meg figur 3.3 på side 32. Analogt ga koden

```
> plot(fit2,form=resid(.,type="p")~fitted(.), id=0.05)
```

meg figur 3.4 på side 33.

Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.

John von Neumann

B

Simulering av likelihood ratio

B.1 Simulering av ordinær likelihood ratio

Eg laga følgande funksjon for å simulere ein LR for dei ordinære lineære regresjonsmodellane i likning (1.3) på side 8 og (1.6) på side 10.

```
ordsim <- function(Di,omu,Osigmau)
 {
  Otmp <- rnorm(108,omu,Osigmau)

    Di$distance <- Otmp

    fit00ny <- lm(distance~age,data=Di)
    fit0ny <- lm(distance~age*Sex,data=Di)
    l0 <- logLik(fit00ny)[1]
    l1 <- logLik(fit0ny)[1]
    return( -2*(l0-l1) )

 }
```

Ved så å definere

```
> D <- Orthodont
> fit00 <- lm(distance~age,data=Orthodont)
> omu <- predict(fit00,D)
> Osigmau <-2.537
> Lsim <- replicate(1000,ordsim(D,omu,Osigmau))
```

kunne eg la funksjonen `replicate` utføre 1000 repitisjonar for `ordsim` med dei rette input-verdiane. Den opprinnelege LR til modellane fann eg at var

```
> fit0 <- lm(distance~age*Sex,data=Orthodont)
> Lordobs <- -2*(logLik(fit00)[1] - logLik(fit0)[1])
> Lordobs

[1] 27.33521
```

Eg fann ved å nytte funksjonen `mean` at andelen simulerte LR som var større eller like denne var

```
> mean(Lsim>=Lordobs)

[1] 0
```

Altså var ingen av dei simulerte LR større enn den observerte, noko som høver med den observerte p-verdien i kapittel 1 som er 1.9×10^{-6}.

B.2 Programkode for simulering av likelihood ratio

I simuleringa av testobservatoren til hypotesa i likning (3.17) på side 43, vil eg nytte funksjonen `anova` til å berekne likelihood ratioen til dei to blanda modellane for `Orthodont`. Eg kan hente kun den berekna likelihood ratioen for modellane ut av utsrifta ved å skrive

```
> anova(fit1,fit2)$L.Ratio[2]

[1] 0.8331072
```

Eg nyttar dei estimerte verdiane til random intercept-modellen, som synt i tabell 2.1 på side 24, til å simulere nye datasett. For kvart individ antar eg at observasjonane har same stokastisk konstantledd, b_i, men at støyledda varierar frå observasjon til observasjon. Den forventa avstanden til individa, let eg R estimere ved funksjonen `predict`. Eg kan slik simulere datasett ved å generere normalfordelte variablar der dei predikerte verdiane av avstanden er forventning, og variansen er den estimerte verdien av σ^2 mellom observasjonar, og den estimerte verdien av σ_b^2 mellom individ. Med dette simulerte datasettet som utgangspunkt, kan eg tilpasse

dei to blanda modellane på nytt, og nytte `anova` til å beregne likelihood ratioen for dei to modellane tilpassa dei simulerte dataa. Slik programmerte eg dette i R:

```
sim2 <- function(Di,mu,sigmab,sigmau) {
   tmp <- rnorm(108,mu,sigmab)
   ii = 1
   for (i in 1:27)
     {
       ui = rnorm(1,0,sigmau)
       for (j in 1:4)
         {
           tmp[ii] = tmp[ii] + ui
           ii = ii + 1
         }}
   Di$distance <- tmp

   fit1ny <- try(lme(distance~age*Sex,data=Di, random=~1|
       Subject,method="ML"),T)
   fit2ny <- try(update(fit1ny,data=Di, random=~age|Subject,
       method="ML"),T)

   if( data.class(fit1ny)=="try-error" | data.class(fit2ny)=="
       try-error" )
     return(NA)
   else return( anova(fit1ny,fit2ny)$L.Ratio[2] ) }
```

Eg nyttar `try` i estimeringa av dei blanda modellane, sidan utan denne vil, når eg skal utføre simuleringa repeterte gonger, programmet stogge kvar gong maskina ikkje får modellen `fit2ny` tilpassa ved `lme`. Dette var eit startproblem som viste seg å kun dukke opp i nyare versjonar av R. Men eg har lete `try` bli i programmet sidan den er nødvendig dersom ein ikkje har tilgang på ein bestemt versjon av R. Med ei løkke, `if`, vil eg istaden få NA-verdiar i tilfeller der `lme` får problem med å tilpasse modellen med to random effects til dei simulerte datasetta, og `try` gjev meldinga `data.class(.)==«try-error»`.

Eg utførte denne simuleringa 1000 gonger, ved å nytte

```
> L <- replicate(1000, sim2(D,mu,sigmab,sigmau))
```

der eg lot

```
> D <- Orthodont
> fit1 <- lme(distance~age*Sex,data=D,random=~1|Subject,method="ML")
> mu <- predict(fit1,D)
```

```
> sigmab <- sqrt(getVarCov(fit1)[1])
> sigmau <- summary(fit1)$sigma
```

Desse 1000 simulerte verdiane vises i det høgre histogrammet i figur 4.2 på side 51. Andelen av desse simulerte verdiane som var større eller lik den observerte ratioen frå det orginale datasettet berekna eg ved kommandoen `mean`.

```
> mean(L > anova(fit1,fit2)$L.Ratio[2])

[1] 0.464
```

For å jamføre med ei blanda 50:50 χ^2_1-, χ^2_2-fordeling, nytta eg R sin tabell av χ^2-fordelte verdiar. Den observerte likelihood ratioen frå det orginale datasettet har her fått navnet `Lobs`.

```
> 1 - ( 0.5*pchisq(Lobs,2) + 0.5*pchisq(Lobs,1) )

[1] 0.5103454
```

Likens nytta eg vekter med høve 65:35.

```
> 1 - ( 0.65*pchisq(Lobs,1) + 0.35*pchisq(Lobs,2) )

[1] 0.466
```

Desse p-verdiane er dei eg har oppsummert i den første rada i tabell 4.2 på side 54.

For p-verdiane eg berekna i rada merka REML, utførte eg ein tilsvarande simulering som den eg har synt for modellane tilpassa ved ML. Den einaste endringa i programmet er at eg fjernar `method="ML"` i koden, og at eg nyttar

```
> fitR1 <- lme(distance~age*Sex,data=D,random=~1|Subject)
> fitR2 <- update(fitR1,random=~age|Subject)
> muR <- predict(fitR1,D)
> sigmabR <- 1.816214
> sigmauR <- 1.386382
> RL <- replicate(1000, sim2(D,Rmu,sigmabR,sigmauR))
```

Dette gav med p-verdien

```
> mean(RL > anova(fitR1,fitR2)$L.Ratio[2])

[1] 0.373
```

Vidare jamførte eg den observerte restricted likelihood ratioen frå datasettet med ei 50:50 og ei 65:35 blanding av χ^2_1- og χ^2_2-fordelingane.

```
> RLobs <- anova(fitR1,fitR2)$L.Ratio[2]
> 1 - ( 0.5*pchisq(RLobs,1) + 0.5*pchisq(RLobs,2) )
> 1 - ( 0.65*pchisq(RLobs,1) + 0.35*pchisq(RLobs,2) )

[1] 0.4169038
[1] 0.3753095
```

Desse p-verdiane er oppsummert i rada merka REML i tabell 4.2 på side 54.

Eg vil til sist i dette tillegget gjere oppmerksam på at funksjonen `density` som eg har nytta til å plotte tettleikane gjev ein «knekk» i null for dei simulerte tettleikane. Denne knekken klarte eg ikkje å fjerne sjølv om sette `xlim=c(0,6)` i plotta. Dermed kan ein sjå vekk i frå kommentarar der eg har påstått at dei simulerte tettleikane ikkje høver godt i null.

The odds against there being a bomb on a plane are a million to one, and against two bombs a million times a million to one. Next time you fly, cut the odds and take a bomb.

Benny Hill

C

Analyse av generaliserte lineære blanda modellar

C.1 Innleiande analysar

Datasettet `Klugman` inneheld i Klugman (1992, side 185–196) 931 observasjonar av variablane Count_i, Year_i, Payroll_i og Class_i, som høvevis er talet på krav, året talet blei registrert, totale lønningslister og den tilhøyrande arbeidgruppa desse målingane blei gjort på. I si opprinnelege form finn eg at talet på høvevis krav og lønningslister som er lik null er

```
 > sum(krav.org$Count==0)

[1] 188

> sum(krav.org$Payroll==0)

[1] 36
```

Når eg fjernar dei 36 observasjonane der Payroll_i er lik null, er det att 152 observasjonar av null krav som tilsvarar eit høve på

```
> sum(krav$Count==0)/length(krav$Count)
```

```
[1] 0.1698324
```

Dermed er altså omlag 17 % av observasjonane av Count_i lik ingen registrerte krav. Verdien $\text{Count}_i = 0$ er den hyppigast observerte, etterfulgt av $\text{Count}_i = 1$.

C.2 Analysar ved lmer

For Poisson modellane i kapittel 5, nytta eg i første omgang pakken `lme4` og funksjonen `lmer` i estimaringa av modellparametrar. Denne let ein ha generaliserte modellar med variable parametrar.

```
> install.packages("lme4",dependencies=TRUE)
> library(lme4)
```

I funksjonen `lmer` kan ein ta inn offset-parametrar ved ganske enkelt å skrive `offset=«.»`, og navnet på offset-variabelen, i klammene til funksjonen. Funksjonen har tre metodar ein kan nytte i estimeringa av modellparametrar: PQL, Laplace og AGQ. Metoden PQL, kort for «penalized quasi-likelihood», og er default for `lmer`. Denne er den snøggaste av metodane, men kan gje unøyaktige estimat. Metoden AGQ, kort for «adaptive Gaussian quadrature», er den mest nøyaktige av dei tre. Derimot er den treg, og eg fekk ikkje denne til å virke på mine modellar. Metoden «Laplace» nyttar, som navnet indikerar, Laplace approksimasjonar i estimeringa. Denne er meir nøyaktig enn PQL-metoden, og raskare enn AGQ-metoden. For Poisson GLMM-ar med henhaldsvis éin og to random effects-ledd, og log som link-funksjon, tilpassa ved metoden PQL, blir koden slik:

```
> klug1 <-
lmer(Loss~Year+(1|Class),offset=log(Payroll),data=krav
+ family=poisson(link="log"),method="ML")
> klug2 <-
lmer(Loss~Year+(Year|Class),offset=log(Payroll),data=krav
+ family=poisson(link="log"),method="ML")
```

Verdiane som blir estimert ved denne metoden er oppsummert i tabell 5.4 på side 71. Eg nyttar så Laplace metoden i estimeringa.

```
> klug1.la <-
lmer(Count~Year+(1|Class),offset=log(Payroll),data=krav
+ family=poisson(link="log"),method="Laplace")
> klug2.la <-
lmer(Count~Year+(Year|Class),offset=log(Payroll),data=krav
+ family=poisson(link="log"),method="Laplace")
```

Differansen i estimerte verdiar frå desse to metodane kan sjåast i tabell 5.4 på side 71.

C.3 Analyse ved glmmADMB

I analyse av negative binomiske GLMM-ar, kan eg ikkje nytte `lmer`, sidan den negative binomiske fordelinga ikkje er definert i denne funksjonen. Eg vil istaden nytte pakken `glmmADMB` (Skaug *et al.* 2006), og den tilhøyrande funksjonen `glmm.admb`. Koden under viser korleis eg har tilpassa modellane.

```
> library(glmmADMB,lib.loc="/home/u5/skaug/lib/R")
> krav$logPayroll <- log(krav$Payroll)
> modNB <-
glmm.admb(Loss~Year+log(Payroll), random=~1,
+ group="Class",family="nbinom",data=krav,offset="logPayroll")
> modNB.yr <-
glmm.admb(Loss~Year+log(Payroll), random=~Year,
+ group="Class",family="nbinom",data=krav,offset="logPayroll")
```

I `glmm.admb` kan ein nytte ein meir robust metode ved å tilføye `easyFlag = F`. Eg har sett på estimat av modellar frå både den robuste og den mindre robuste metoden i tabell 5.5 på side 75. Men for modellen med negativ binomisk fordeling og to variable parametrar konvergerte ikkje funksjonen `glmm.admb` hos meg når eg nytta den robuste metoden, og meldingane

```
Error matrix not positive definite in choleski_decomp
The function maximizer failed
```

dukkar opp. Eg tilpassa likeins Poisson GLMM-ar i tabell 5.5 på side 75 ved

```
> modP <-
glmm.admb( Loss~Year+log(Payroll),random=~1, group="Class",
+ family="poisson",data=krav, offset="logPayroll")
> modP.yr <-
glmm.admb( Loss~Year+log(Payroll), random=~Year,group="Class",
+ family="poisson",data=krav,offset="logPayroll")
```

der eg også har nytta både default metoden, og `easyFlag = F`. Her konvergerte `glmm.admb` i alle fire tilfellene.

For dei nullforhøgde modellane i 5.6, tek eg med `zeroInflation=TRUE`. For Poisson-modellen med to variable parametrar konvergerte dessverre ikkje funksjonen `glmm.admb`, og eg fekk meldinga

```
> po2.zip <-
glmm.admb(Count~Year, zeroInflation=TRUE, random=~Year,
+ group="Class",family="poisson",data=krav,offset="logPayroll")

Hessian does not appear to be positive definite
The function maximizer failed
```

Derfor inneheld ikkje tabell 5.7 på side 80 estimat av denne modellen. Men då eg nytta den nøysame metoden fekk eg resultat, og likens for den nullforhøgde modellen med negativ binomisk fordeling for responsen og to variable parametrar. For denne konvergerte `glmm.admb` både utan og med `easyFlag = F`.

Når det gjeld verdien av AIC til modellane i tabell 5.5 på side 75, tabell 5.7 på side 80 og tabell 5.8 på side 82 har eg nytta funksjonen

```
aicverdi <- function(modellnavn,antparm)
  {
    l <- logLik(modellnavn)[1]
    -2*l+2*antparm
  }
```

som er basert på definisjon 3.4.1 på side 46. Her var det viktig for meg å vite talet på parametrar i dei respektive modellane. Eg har nytta verdien i rada $|\boldsymbol{\theta}|$ i dei respektive tabellane då eg nytta denne funksjonen.

Eg har nytta versjon `2.4.0` av R, ved desse berekningane.

I was still a couple of miles above the clouds when it broke, and with such violence I fell to the ground that I found myself stunned, and in a hole nine fathoms under the grass, when I recovered, hardly knowing how to get out again. Looking down, I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoist myself to the top and stepped out on terra firma without further ado.

Singular Travels, Campaigns and Adventures of Baron Munchausen (1786)

D

Bootstrap

Sitatet over er opprinnelsen til uttrykket bootstrap. Uttrykket «pulled himself up by his own bootstraps» har innan finans blitt nytta for selskap som har fått suksess på eiga hand, og innan datateknologi om å nytte enkle algoritmar til å starte større og meir kompliserte.

For modellen med éin variabel parameter laga eg følgane funksjon for likelihood.

```
Lh1 <- function(Di,a) {

 tmp <- rnorm(108,a[1],a[3])
   ii = 1
   for (i in 1:27)
     {
       ui = rnorm(1,0,a[2])
       for (j in 1:4)
         {
           tmp[ii] = tmp[ii] + ui
           ii = ii + 1
         }
     }
   Di$distance <- tmp
 fit1RL.sim <- lme(distance~age*Sex,data=Di,random=~1|Subject)
 return(logLik(fit1RL.sim)[1])
```

```
}
```

For å bootstrappe likelihooden definerte eg funksjonen

```
cstar <- function(b1)
{
t0 <- boot(D,Lh1,b1)$t0
t <- boot(D,Lh1,b1)$t

op <- cbind(t0, apply(t, 2, mean, na.rm = TRUE)-t0,
+ sqrt(apply(t, 2, function(t.st) var(t.st[!is.na(t.st)]))))
op[2]-op[1]
}
```

Så utførte eg simuleringar av denne ved

```
c.100 <- replicate(10,cstar(100))
```

og fann EIC-verdi ved

```
C1k <- mean(c.100)
eic1 <- function(C1k){ -2*logLik(fit1rl)[1]
+ 2*C1k }
```

Resultata av dette kan studerat i kapittel 7.

Litteratur

Antonio K. og Beirlant J. (2006). «Actuarial statistics with generalized linear mixed models». `http://www.econ.kuleuven.be/public/NDBAE81/GLMMRevisionIME.pdf`. Referert til på side 1, 56, 57, 62, 63, 64, 67, 70, 71, 81 og 87.

Booth J.G., Casella G., Friedl H. og Hobert J.P. (2003). «Negative binomial loglinear mixed models». *Stat. Model.*, **volum 3**, nummer 3, side 179–191. ISSN 1471-082X. Referert til på side 72 og 73.

Casella G. og Berger R.L. (2002). *Statistical Inference*. Duxbury. Referert til på side 19, 34, 35, 41, 48 og 72.

Corbeil R.R. og Searle S.R. (1976). «Restricted maximum likelihood (REML) estimation of variance components in the mixed model». *Technometrics*, **volum 18**, nummer 1, side 31–38. ISSN 0040-1706. Referert til på side 40.

Dobson A.J. (2002). *An Introduction To Generalized Linear Models, Second Edition*. Chapman & Hall/CRC. Referert til på side 10, 13, 16 og 29.

Fitzmaurice G.M., Laird N.M. og Ware J.H. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Inc. Referert til på side 2, 16, 18, 19, 21, 39, 61 og 64.

Haberman S. og Renshaw A.E. (1996). «Generalized linear models and actuarial science». *The Statistician*, **volum 45**, side 407–436. Referert til på side 63 og 67.

Hall D.B. (2000). «Zero-inflated Poisson and binomial regression with random effects: a case study». *Biometrics*, **volum 56**, nummer 4, side 1030–1039. ISSN 0006-341X. Referert til på side 77.

Ishiguro M., Sakamoto Y. og Kitagawa G. (1997). «Bootstrapping log likelihood and EIC, an extension of AIC». *Annals of the institute of Statistical Mathematics*, **volum 49**, side 411–434. Referert til på side 44 og 45.

Jansakul N. og Hinde J.P. (2002). «Score tests for zero-inflated Poisson models». *Comput. Statist. Data Anal.*, **volum 40**, nummer 1, side 75–96. ISSN 0167-9473. Referert til på side 77, 78 og 80.

Johnson R.A. og Wichern D.W. (2002). *Applied Multivariate Statistical Analysis, Fifth Edidtion*. Prentice Hall. Referert til på side 12.

Klugman S.A. (1992). *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*. Kluwer Academic Publishers. Referert til på side 1, 66, 68 og 102.

Konishi S. og Kitagawa G. (1996). «Generalised information criteria in model selection». *Biometrika*, **volum 83**, nummer 4, side 875–890. ISSN 0006-3444. Referert til på side 89.

Laird N.M. og Ware J.H. (1982). «Random-effects models for longitudinal data». *Biometrics*, **volum 38**, side 963–974. Referert til på side 33, 36, 39 og 40.

McCulloch C.E. (2003). «Generalized linear mixed models». side viii+84. Referert til på side 68 og 69.

Molenberghs G. og Verbeke G. (2004). «Meaningful statistical model formulations for repeated measures». *Statistica Sinica*, **volum 14**, side 989–1020. Referert til på side 16, 18, 19, 20, 22 og 84.

Pinheiro J.C. og Bates D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer. Referert til på side 1, 2, 6, 10, 11, 16, 18, 23, 31, 33, 36, 37, 38, 39, 40, 47, 48, 52, 53, 94 og 96.

Potthoff R.F. og Roy S.N. (1964). «A generalized multivariate analysis of variance model useful especially for growth curve problems». *Biometrika*, **volum 51**, side 313–326. Referert til på side 1 og 7.

R Development Core Team (2006). «R: A language and environment for statistical computing».
`http://www.R-project.org`. ISBN 3-900051-07-0. Referert til på side 1.

Skaug H., Fournier D. og Nielsen A. (2006). «glmmADMB: Generalized Linear Mixed Models using AD Model Builder».
`http://otter-rsch.com/admbre/examples/glmmadmb/glmmADMB.html`. R package version 0.3. Referert til på side 75 og 104.

Stram D.O. og Lee J.W. (1994). «Variance components testing in the longitudinal mixed effects model». *Biometrics*, **volum 50**, side 1171–1177. Referert til på side 51, 52, 55, 84 og 85.

Sundt B. (1999). *An Introduction to Non-Life Insurance Mathematics*. Karlsruhe. Referert til på side 65.

Walpole R.E., Myers R.H. og Myers S.L. (1998). *Probability and Statistics*. Prentice Hall International, Inc. Referert til på side 18, 28 og 36.

Yafune A., Funatogawa T. og Ishiguro M. (2005). «Extended information criterion (EIC) approach for linear mixed effects models under restricted maximum likelihood (REML) estimation». *Stat. Med.*, **volum 24**, nummer 22, side 3417–3429. ISSN 0277-6715. Referert til på side 89.