UNIVERSITETET I BERGEN
Det matematisk-naturvitenskapelige fakultet

Modellering av overdispersjon i populasjonsdata

Master i statistikk
Finansteori og forsikringsmatematikk

Brage Grindheim
November 2023

Takk

Først og fremst vil jeg takke veilederen min, Hans Julius Skaug. Tusen takk for at du har vært tålmodig. Jeg setter stor pris på kommentarene du har gitt meg underveis, og jeg håper du har følt deg hørt.

Jeg vil også benytte anledningen til å takke Yushu Li og Hans Karlsen, for deres bidrag til min faglige utvikling.

Sist, men ikke minst, vil jeg takke Kristine Lysnes for gode råd.

Brage Grindheim,
18. november, 2023

Abstrakt

I denne studien anvendte vi generaliserte lineære modeller (GLM) for å modellere populasjonsdata fra The Human Mortality Database (HMD) for Sverige. Dataene ble brukt til å predikere antall døde med alderstrinn og kalenderår som prediktorer. Ved å anta Poisson som responsfordeling for antall døde, viste modellen seg imidlertid å være overdispersert.

For å inkludere mer variasjon i modellen endret vi responsfordeling fra Poisson til negativ binomial. Denne endringen tillot oss å estimere både forventning og dispersjonsparameter, noe som forbedret modellens tilpasning. For å utbedre modellen vår ytterligere benyttet vi simultan modellering av dispersjonsparameter og forventning med Generalized Linear Models using Template Model Builder (glmmTMB). Det viste seg at modellering av dispersjonsparameteren som en log-lineær funksjon av predktorvariabler, hadde god effekt når vi sammenlignet AIC blant modeller med heterogen dispersjonsparameter, opp mot modeller med homogen dispersjonsparameter.

En simuleringsbasert metode fra Diagnostics for Hierarchical Regression Models (DHARMa) ble brukt for å avgjøre hvor godt de ulike modellene forklarte variabiliteten i data. Programvaren ga oss en ikke-parametrisk tilnærming som sammenlignet observerte og simulerte residualer. Testing av dispersjon viste at dispersjonsmodellering bidro til å inkludere overdispersjon i modellene, og at utfallet varierte avhengig av hvordan vi definerte prediktoren i dispersjonsmodellen.

Prognoser fra de ulike modellene antydet at modellering av dispersjon bidro til å økt styrke for predikert forventning, og resulterte i lavere dødssannsynligheter når vi knyttet estimert dødelighet opp mot overlevelsesanalyse.

Innhold

1 Introduksjon **1**
1.1 Populasjonsdata . . . 2
1.1.1 Variabler . . . 2
1.1.2 Datafeil og endringer . . . 4
1.2 Statistisk læring . . . 4

2 Responsfordelinger for dødelighet **5**
2.1 Poisson-fordelingen . . . 5
2.2 Negativ binomialfordeling . . . 6

3 Den eksponentielle familien og generaliserte lineære modeller **7**
3.1 Den eksponentielle familien . . . 7
3.1.1 Poisson-fordelingen . . . 7
3.1.2 Negativ binomialfordeling . . . 8
3.2 Momentfunksjon og kumulantfunksjon . . . 8
3.2.1 Poisson-fordelingen . . . 9
3.2.2 Negativ binomialfordeling . . . 9
3.3 Generaliserte lineære modeller . . . 9
3.3.1 Notasjon . . . 9
3.3.2 GLM . . . 9
3.3.3 Dispersjonsmodellering . . . 10
3.3.4 Polynomregresjon . . . 11
3.4 Overdispersjon . . . 11

4 Dispersjonsmodellering av svenske dødelighetsdata **13**
4.1 Modellseleksjon . . . 13
4.2 Kategoriske modeller . . . 14
4.2.1 Utvidede negativ binomiale modeller . . . 15
4.3 Polynomiske modeller . . . 17
4.4 Dispersjon . . . 19

5 Residualanalyse **20**
5.1 Pearson residualer . . . 20
5.2 Simulerte residualer . . . 24
5.2.1 Parametrisk bootstrap . . . 24
5.2.2 Kvantilresidualer . . . 25
5.3 QQ-plot . . . 25
5.4 Testing av dispersjon . . . 26
5.4.1 Pearson-χ^2 test . . . 26
5.4.2 DHARMa dispersjonstest . . . 27

6 Overlevelsesanalyse **32**
6.1 Overlevelsesanalyse . . . 32
6.2 Kaplan-Meier . . . 33
6.3 Livtabell . . . 33
6.3.1 Dødsintensitet, dødssannsynlighet og overlevelsessannsynlighet . . . 34

6.3.2 Fremskrivelser 34

7 Diskusjon **37**
7.1 Videre arbeid 38

A Figurer **41**
A.1 Pearson residualer 42
A.2 Livtabell 45

B Kode **46**
B.1 Modeller 46
B.2 Livtabell 49
B.3 Overlevelseskurver 52

Tabeller

4.1 Parameter estimater for dispersjon fra de kategoriske modellene. 19

5.1 AIC for modellene som ble tilpasset i kapittel 4. 20
5.2 Pearson-χ^2 test 26
5.3 DHARMa dispersjonstest 30

Figurer

1.1 Rapporterte antall dødstilfeller blant svenske menn som funksjon av alderstrinn i perioden 1951 til 2005. . . . 2
1.2 Svensk dødelighet fra 1. januar 1951 til 31. desember 2022. . . . 3

3.1 Logaritmen av antall døde som funksjon av alderstrinn og kalenderår. . . . 10

4.1 AIC fra modeller som ble tilpasse under seleksjonsprosessen, som funksjon av polynomgrad a for alderstrinn og b for kalenderår. Til venstre ser vi AIC fra modeller med $b = 4$, og til høyre for modeller med $a = 25$. . . . 14
4.2 Predikert dødsrate per person i middelbefolkningen for 80-åringer fra de polynomiske modellene, inkludert konfidensintervaller. De sorte punktene er observasjoner fra den samme perioden. . 15
4.3 AIC fra seleksjonsprosess som ble gjennomført for å forsøke å finne den optimale polynomiske kurven. Til venstre ser vi AIC fra modeller med dispersjonsmodeller som tilpasset dataene til kurver av grad $b = 8$ til venstre, og til $a = 10$ til høyre. . . . 16
4.4 Predikert dødsrate for 0-åringer fra polynomiske modeller, med konfidensintervaller. De sorte punktene er observert dødsrate. . . . 17
4.5 Sammenligning av Poisson-overdispersjon modellert av M_2, M_3 og M_4 sammenlignet med M_1. 18
4.6 Sammenligning av dispersjon mellom M_5 og M_6 (204 punkter utelatt). . . . 19

5.1 Pearson residualer som funksjon av kalenderår for M_0 og M_1 . . . 21
5.2 Pearson residualene for alderstrinn 0 for M_0 og M_1 preges av *sporing*. . . . 22
5.3 Sammenligning av standardiserte residualer mellom hver av modellene. . . . 22
5.4 Pearson residualer for M_4, M_5 og M_6. . . . 23
5.5 Linjeplot med residualer for alderstrinn 0 fra modellene som funksjon av kalenderår. . . . 23
5.6 QQ-plot for M_0 og M_1 . . . 26
5.7 QQ-plot for kategoriske modeller. . . . 27
5.8 QQ-plot for polynomiske modeller. . . . 28
5.9 DHARMa dispersjonstest for kategoriske modeller. . . . 29
5.10 DHARMa dispersjonstest for polynomiske modeller. . . . 30

6.1 Kumulative overlevelsessannsynligheter fra livtabeller beregnet ved å bruke prediksjoner fra M_5 og M_6 sammenlignes med livtabeller basert på observerte verdier. . . . 35
6.2 Forventet levealder. . . . 36

A.1 Pearson residualer som funksjon av kalenderår. . . . 42
A.2 Pearson residualer som funksjon av alderstrinn. . . . 43
A.3 Variabler fra livtabell. Livtabellen ble konstruert fra prediksjoner for 2005 fra M_6. . . . 45

Kapittel 1

Introduksjon

Generaliserte lineære modeller (GLM) bruker parametriske antagelser for til å forklare variasjon i en avhengig variabel ved hjelp av en eller flere uavhengige variabler. Vår studie tar sikte på å undersøke årsaker og konsekvenser av overdispersjon. Overdispersjon oppstår når variasjonen i data er større enn det som samsvarer med antagelsene GLM gjør om fordelingen til den avhengige variabelen. Det fins flere mulige kilder til overdispersjon, men allikevel kan identifisering av overdispersjon være en utfordring da det ikke alltid er åpenbart hva den skyldes. Derfor er det nødvendig å utforske alternative metoder for å evaluere behovet for tilpasninger i modellen for å håndtere overdispersjon.

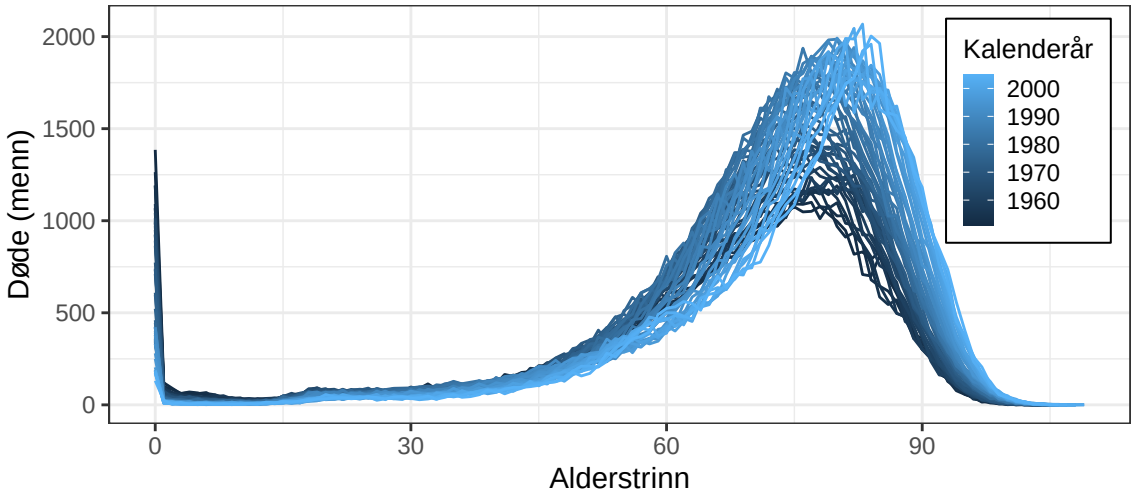
Generalized Linear Models using Template Model Builder (glmmTMB) er en R-pakke som gir muligheten til å tilpasse et bredt spekter av Generalized Linear Models (GLM) og generaliserte lineære blandede modeller (Brooks et al., 2017). Et nøkkelelement i glmmTMB er dispersjonsmodellen, som tillater modellering av dispersjonsparametrene ved å anta et lineært forhold mellom en logaritmisk transformasjon av dispersjonsparameteren og en eller flere prediktorvariabler. Dispersjonsmodellen muliggjør simultan modellering av både forventning og dispersjon, og var spesielt nyttig når vi analyserte populasjonsdata fra The Human Mortality Database (HMD, 2022).

Tidligere har Kolve (2022) undersøkt dispersjonsmodellering av seldata ved hjelp av Conway-Maxwell-Poisson fordelingen, mens Evensen (2018) har utforsket ulike metoder for dispersjonsmodellering, som inkluderte bruk av glmmTMB. Denne studien omhandler hvordan vi kan løse problemstillingen tilknyttet identifisering av overdispersjon ved bruk av tester, og visualisering. Det innebærer at fokuset vil primært være rettet mot reduksjon av overdispersjon gjennom modellering av dispersjonsparameteren i negativ binomiale modeller. Vi skal også benytte ulike transformasjoner av prediktorvariabler. Metodene vil bidra til økt forståelse av komplekse variasjonsmønstre i populasjonsdata.

Selv om det er vanlig å bruke Pearson- og responsresidualplott for å undersøke feilspesifikasjoner i Generalized Linear Models (GLM), er residualplot ikke alltid en pålitelig metode for å evaluere modellspefikasjoner, spesielt når man må velge blant nestede modeller. Residual Diagnostics for Hierarchical (multi-level/mixed) Regression Models (DHARMa) er en R-pakke som tilbyr muligheten til å teste for overdispersjon ved hjelp av en ikke-parametrisk tilnærming, som sammenligner variasjonen i observert og simulert utvalg(Hartig, 2022a). I forbindelse med bruken av denne R-pakken, vil vi utforske teoretiske prinsipper og formuleringer for å bedre forstå metodene som er beskrevet i pakkens dokumentasjon.

I dette kapittelet skal vi først presentere variablene i datasettet vårt, etterfulgt av en bred formulering av statistisk læring. I kapittel 2 vil vi greie ut om Poisson- og negativ binomialfordeling, som legger grunnlaget for parametriske antakelser som er nødvendige når vi skal modellere diskrete telledata. I kapittel 3 formuleres den nødvendige teorien fra de eksponentielle familiene, som vi deretter setter i sammenheng med GLM og overdispersjon.

I kapittel 4 tilpasser vi populasjonsdata for å modellere årlige dødsfall blant svenske menn ved ulike aldertrinn. I kapittel 5 gjennomfører vi en standard residualanalyse av Pearson residualer, før vi utleder funnene som er gjort i forbindelse med simulering og ikke-parametrisk dispersjonstest i DHARMa. Prediksjoner blir benyttet til overlevelsesanalyse i kapittel 6, hvor vi påviser forskjeller som oppstår avhengig av hvordan dispersjon er estimert av modellene.



Figur 1.1: Rapporterte antall dødstilfeller blant svenske menn som funksjon av alderstrinn i perioden 1951 til 2005.

1.1 Populasjonsdata

For å undersøke dispersjonsmodellen skal vi benytte populasjonsdata som er hentet fra The Human Mortality Database (HMD, 2022). HMD er en omfattende samling av data som inneholder informasjon om dødelighet og demografi fra ulike land rundt om i verden. I oppgaven blir det brukt svenske populasjonsdata. Datasettene kan brukes til en rekke forskningsformål, inkludert studier om folkehelse, demografi og epidemiologi.

I vår studie brukte vi data for perioden 1951 til 2022. Vi trente modellene våre på data fra 1951 til 2005, og brukte observasjoner fra 2006 til 2022 til å undersøke treffsikkerhet blant modellenes prediksjoner. Totalt sett omfatter det fullstendige datasettet 7701 observasjoner, hvorav 5864 ble brukt til å trene modellene, mens de resterende 1837 observasjonene ble brukt til validering.

Datasettet gir informasjon om dødsfallene i befolkningen, og er strukturert på en 1x1 basis. Dette datasettet inkluderer variabler som alder, kjønn, middelbefolkning og antall dødsfall for hvert alderstrinn, og presenterer et detaljert bilde av dødelighetsmønstre og -trender over tid. Variablene er viktige når vi skal utforske demografiske trender, aldersspesifikke dødsrater og analyser av dødelighet i en befolkning. Slik data tilbyr verdifull innsikt i hvordan dødelighet og middelbefolkning varierer mellom ulike aldersgrupper og kjønn, og hvordan disse mønstrene endrer seg over tid.

1.1 viser antall døde som funksjon av alderstrinn. Vi ser at det forekommer betydelig variasjon mellom ulike kalenderår, men også mellom alderstrinn. 1.2 viser at antall døde følger en tilfeldig gange over tid. Selv om tidsseriedata ofte modelleres med stokastiske differensialligninger ønsker vi å undersøke hvordan vi kan predikere dødelighet med GLM, samt predikere variasjon som oppstår over tid og alderstrinn ved hjelp av dispersjonsmodellen i glmmTMB.

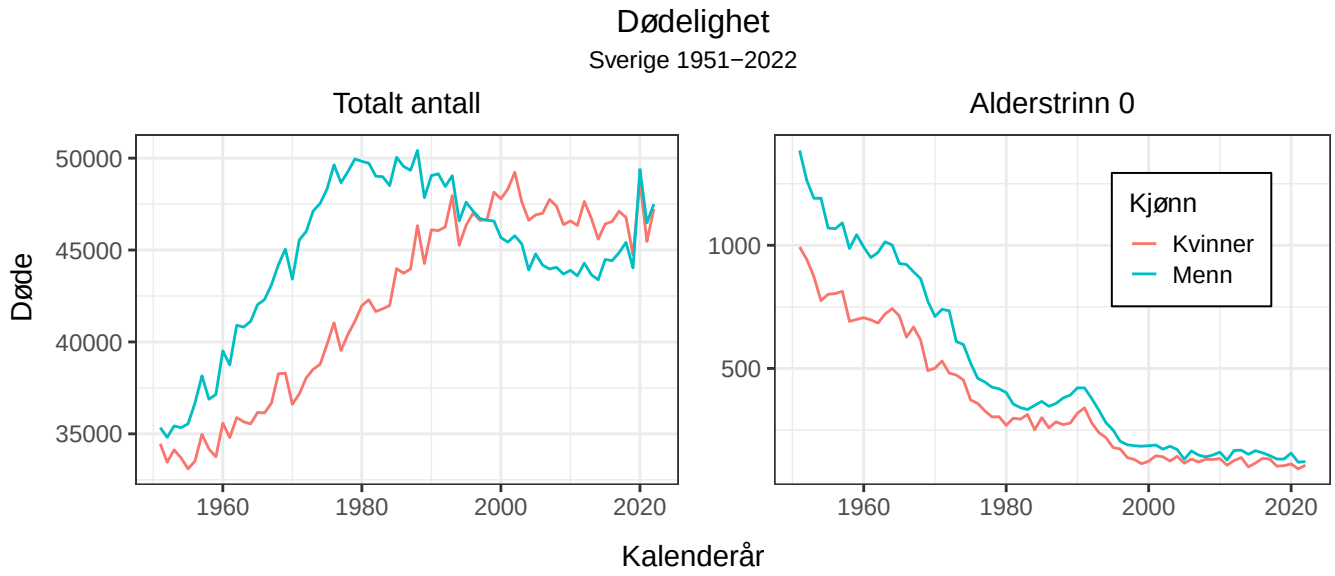
1.1.1 Variabler

Døde

En variabel som oppgir antall døde er en essensiell variabel i et demografiske datasett, da den registrerer antall dødsfall i befolkningen for en bestemt tidsperiode. Denne variabelen er avgjørende for å få innsikt i dødelighet og er grunnleggende for å forstå befolkningsdynamikk, helsestatus og demografiske trender. Når man arbeider med dødstall variabelen, er det viktig å være oppmerksom på mulige utfordringer, som manglende registrering av dødsfall, feilrapportering av alder eller underrepresentasjon av spesielle grupper i datasettet. Dette kan påvirke nøyaktigheten av dataene og analysens validitet.

Middelbefolkning

Middelbefolkning er en viktig variabel i demografiske datasett, da den gir informasjon om størrelsen på befolkningen ved midten av året. Denne variabelen er avgjørende for å beregne dødsrater, analysere demografiske



Figur 1.2: Svensk dødelighet fra 1. januar 1951 til 31. desember 2022.

trender og for å forstå befolkningsstørrelsen som er utsatt for risiko. Ved å arbeide med eksponeringsdata kan helseplanleggere bedre forstå befolkningens sammensetning og helsebehov.

Alderstrinn

Alder er en fundamental faktor som påvirker dødsrater betydelig. Ved å dele befolkningen inn i ulike alderstrinn, for eksempel barndom, ungdom, voksenliv og eldre, kan man analysere hvordan dødsrisiko varierer i forskjellige livsfaser. Dette gir verdifull informasjon om aldersspesifikke dødelighetsmønstre og hjelper til med å identifisere helseutfordringer som er unike for ulike aldersgrupper. For eksempel kan barn ha høyere risiko for infeksjonssykdommer, mens eldre voksne kan være mer utsatt for kroniske sykdommer som hjerte- og karsykdommer og kreft. Ved å inkludere alderstrinn i en statistisk modell kan man også beregne aldersjusterte dødsintensiteter, som gir en mer rettferdig sammenligning av dødelighet mellom ulike befolkningsgrupper med forskjellig aldersstruktur.

Kalenderår

Kalenderår er en kritisk variabel i statistiske modeller for dødsfall, da det gjør det mulig å identifisere trender og endringer i dødsrater over tid. Ved å analysere dødelighet over flere år kan man avdekke langvarige utfordringer, identifisere spesielle hendelser som pandemier eller epidemier, og vurdere effekten av folkehelseintervensjoner over tid. For eksempel har kalenderår spilt en avgjørende rolle i å overvåke og forstå effekten av HIV/AIDS-pandemien, samt evaluere vaksineprogrammer mot infeksjonssykdommer som meslinger og polio. Ved å inkludere kalenderår som variabel kan man også studere sesongvariasjoner i dødelighet og forberede seg på mulige helseutfordringer knyttet til sesongvise klimaendringer.

Kjønn

Kjønn er en annen viktig variabel i statistiske modeller for dødsfall, da det ofte er betydelige forskjeller i dødsrater mellom menn og kvinner. Ved å inkludere kjønn som en variabel, kan man undersøke kjønnsforskjeller i helse og dødelighet, samt identifisere spesifikke helseutfordringer knyttet til hvert kjønn. For eksempel har menn vanligvis høyere risiko for dødsfall knyttet til ulykker og vold, mens kvinner kan ha høyere risiko for dødsfall relatert til svangerskap og fødsel. Kjønn som variabel gir også muligheten til å utforske hvordan sosiale og kulturelle faktorer påvirker dødelighetsmønstre mellom kjønn.

1.1.2 Datafeil og endringer

HMD melder at måle- og rapporteringsfeil er uunngåelig. Mulige feil og dataendringer er rapportert på deres hjemmesider. Hver gang en ny metode blir oppdaget kan det også skje endringer som påvirker tidligere befolkningshistorikk. Rader hvor middelbefolkningen er 0 har blitt fjernet fra datasettet for å unngå problemer knyttet til relative frekvenser. Data som er brukt i denne studien ble lastet ned fra HMD 21.07.2023.

1.2 Statistisk læring

Statistisk læring referer til et sett med verktøy som vi kan bruke for å forstå data. Disse verktøyene kan klassifiseres som veiledet eller ikke-veiledet læring(James, 2013, s. 1). I bred forstand handler veiledet statistisk læring om å bygge statistiske modeller for å predikere, eller estimere, utdata basert på inndata. Ikke-veiledet statistisk læring omhandler metoder for å undersøke forhold og struktur i data. Vi forestiller oss at vi er i en situasjon hvor vi ønsker å vite mer om forholdet mellom en kvantitativ respons Y og p forskjellige prediktorer $Z_1, ..., Z_p$. Vi antar at det eksisterer et forhold mellom y og $\boldsymbol{z} = (Z_1, Z_2, ..., Z_p)$ som kan skrives på den generelle formen

$$Y = f(\boldsymbol{z}) + \epsilon. \tag{1.1}$$

Her er f en ukjent funksjon av $Z_1, ..., Z_p$, og ϵ er et tilfeldig feilledd (James, 2013, 16). Vi betegner ϵ som en uavhengig tilfeldig variabel med forventning 0 og standardavvik σ_ϵ, det vil si

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon^2). \tag{1.2}$$

I (1.1) representerer f den systematiske informasjonen som $\boldsymbol{z}$ oppgir om Y. Vi ønsker å estimere f for å kunne predikere og inferere om Y. En prediksjon for Y antar at forholdet til $\boldsymbol{z}$ kan beskrives av

$$\hat{Y} = \hat{f}(\boldsymbol{z}), \tag{1.3}$$

hvor $\hat{f}$ er et estimat av f, og $\hat{Y}$ representerer prediksjonen for Y gitt at vi kjenner $\boldsymbol{z}$ (James, 2013, s. 17).

Nøyaktigheten av $\hat{Y}$ som en prediksjon for Y avhenger av to mengder som vi kaller reduserbar og ureduserbar feil. Generelt sett vil ikke $\hat{f}$ være et perfekt estimat for f. Vi kaller differansen mellom $\hat{f}$ og f for en reduserbar feil. Det vil si at vi forbinder forskjellen mellom funksjonene som mulig å redusere ved å velge metoden fra statistisk læring som egner seg best for å estimere f.

Allikevel, når $\hat{f}$ i (1.3) er et perfekt estimat for f slik at prediksjonen vår er

$$\hat{Y} = f(\boldsymbol{z}), \tag{1.4}$$

vil den fortsatt inneholde feil fordi Y også er en funksjon av ϵ. Som nevnt tidligere er ϵ en tilfeldig variabel, og uavhengig av $\boldsymbol{z}$. Derfor kan vi ikke predikere ϵ som funksjon av $\boldsymbol{z}$ (James, 2013, s. 18).

Variasjonen i ϵ er det vi kaller for ureduserbar feil. Ureduserbar feil forekommer at ϵ inneholder informasjon som er nyttig for å predikere Y, men som vi ikke har målt. ϵ kan også innholde umålbar variasjon.

Fra differansen mellom (1.1) og (1.3) finner vi at

$$\begin{aligned} \text{E}(Y - \hat{Y})^2 &= \text{E}\left[f(\boldsymbol{z}) + \epsilon - \hat{f}(\boldsymbol{z})\right]^2 \\ &= \left[f(\boldsymbol{z}) - \hat{f}(\boldsymbol{z})\right]^2 + \text{V}\left[\epsilon\right], \end{aligned} \tag{1.5}$$

hvor $\text{E}(Y - \hat{Y})^2$ er gjennomsnittlig, eller forventet, kvadrat av differansen mellom predikert og faktisk verdi for Y. $\text{V}[\epsilon]$ er variansen som assosieres med feilleddet(James, 2013, s. 19). Her er $\text{V}[\epsilon]$ den ureduserbare feilen og $\left[f(\boldsymbol{z}) - \hat{f}(\boldsymbol{z})\right]^2$ den reduserbare feilen. Når vi har valgt metoden som fjerner den reduserbare feilen følger det at

$$\text{E}(Y - \hat{Y})^2 = \text{V}\left[\epsilon\right] = \sigma_\epsilon^2. \tag{1.6}$$

Et vanlig analyseverktøy for GLM er residualer. Residualer beskriver avstanden $Y - \hat{Y}$, og en residualanalyse kan utføres for å avgjøre om den reduserbare feilen er redusert tilstrekkelig. Det gjøres ved å undersøke om det kan være rimelig å anta at (1.6) . Residualplot brukes ofte for å gjennomføre residualanalyse, og selv etter standardisering kan det være vanskelig å skille mellom nestede modeller for å avgjøre hvilke av dem som har estimert f best.

I kapittel 5 skal vi vise funnene vi gjorde når vi benyttet simulering til å å avgjøre om $\hat{Y}$ er et godt estimat av Y ved å teste residualene. Testen vurderer om utfallet for residualvariansen er sannsynlig gitt antagelsene om f.

Kapittel 2

Responsfordelinger for dødelighet

I dette kapittelet skal vi greie ut om Poisson, og negativ binomialfordeling som fordelinger for å beskrive utfallene for en stokastisk variabel Y. Fordelingene brukes ofte som responsfordelinger for å modellere telledata, det vil si data som oppgir frekvenser, og er ofte brukt som responsfordelinger når vi bruke GLM. Vi går nærmere inn på teorien som ligger til grunn for generaliserte lineære modeller (GLM) i kapittel 3.

Vi ønsker å modellere dødelighet, og lar Y betegne antall døde. Med data fra HMD skal vi predikere antall døde som funksjon av alderstrinn x og kalenderår t. Vi antar at antallet som dør ved de ulike aldrene skyldes interaksjon mellom middelbefolkningen n og intensiteten μ.

Dersom vi antar at individene i middelbefolkningen dør uavhengig av hverandre, har vi at Y er summen av n Bernoulli forsøk, $I_j \sim \text{Bernoulli}(\mu)$ $(0 \leq j \leq n)$, det vil si

$$Y = I_1 + ... + I_n = \sum_{j=1}^{n} I_j. \tag{2.1}$$

Det fins flere fordelinger for å modellere Y, og i dette kapittelet skal vi greie ut om hvilke som er relevant i vår sammenheng. Binomisk fordeling er ofte brukt til å modellere utfallet av en sum av Bernoulli forsøk, men på grunn av en kollektiv datastruktur må vi bruke responsfordelinger som egner seg for å beskrive kollektive frekvenser.

I kapittel 6 forklarer vi fremgangsmåten vi brukes for å estimere intensiteten μ fra Y. For å estimere dødssannsynligheten q fra μ bruker vi ulikheten

$$\mu = -\log(1-q) > q, \tag{2.2}$$

for å være på den sikre siden (Kaas, 2008, s. 59).

2.1 Poisson-fordelingen

La f være en vilkårlig tetthetsfunksjon. Sannsynligheten for å observere utfallet y fra Y er

$$f_Y(y) = P(Y = y). \tag{2.3}$$

Om $Y \sim \text{Poisson}(\lambda)$ er tetthetsfunksjonen (Dobson, 2018, s. 197)

$$f_Y(y;\lambda) = \frac{(\lambda)^y}{y!}e^{-\lambda}, \quad \lambda \geq 0, y \geq 0. \tag{2.4}$$

Poisson-fordelingen er en fordeling som passer godt til å modellere antall dødsfall, da den tar hensyn til at antall dødsfall er en positiv, diskret variabel. Alder og kalenderår er to viktige prediktorer som kan påvirke dødeligheten betydelig. Alderstrinn er en fundamentalt viktig variabel som kan avdekke aldersspesifikke dødelighetsmønstre, mens kalenderår gir muligheten til å identifisere demografiske trender og endringer over tid. Ved å inkludere alder og kalenderår som prediktorvariabler i en Poisson-modell, kan vi undersøke hvordan dødeligheten varierer i ulike aldersgrupper og hvordan den endrer seg over tid.

I relasjon til estimering av overlevelse fra en Poisson modell skal vi bruke Poisson approksimasjonen av binomisk fordeling. Poisson approksimasjonen sier oss at dersom

$$W \sim \text{Binomial}(n, q), \text{ og } Y \sim \text{Poisson}(\lambda), \tag{2.5}$$

med $\lambda = nq$, så har vi at

$$f_Y(y) \approx f_W(w), \tag{2.6}$$

for stor n og liten q (Casella, 2002, s. 66-67) .

De teoretiske egenskapene for Poisson-fordelingen sier oss at forventning og varians for en Poisson(λ) variabel skal være like (Kaas, 2008, s. 45). I tilfeller hvor

$$\frac{\mathrm{V}[Y]}{\mathrm{E}[Y]} > 1, \tag{2.7}$$

sier vi at Poisson-fordelingen er overdispersert, og i et slikt tilfelle kan en alternativt bruke negativ binomialfordeling til å modellere variasjonen som medfører ulikheten.

2.2 Negativ binomialfordeling

Om $Y \sim$ Negativ binomial(λ, τ) er tetthetsfunksjonen,

$$f_Y(y) = \binom{y+\tau-1}{\tau-1} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(1-\frac{\tau}{\tau+\lambda}\right)^{y}, \tag{2.8}$$

hvor $\lambda \geq 0$ er forventning, og $\tau \geq 0$ er en dispersjonsparameter (Hardin, 2007, s. 202).

Negativ binomialfordeling egner seg godt til å modellere overdispersjon når en Poisson-modell viser seg å være overdispersert. Dette er vanlig i telledata, som dødstall, da variasjonen i antall dødsfall kan være høyere enn det som kan forklares av en Poisson-fordeling. Den negativ binomiale fordelingen gir oss muligheten til å justere for denne ekstra variasjonen, og dermed kan vi få mer realistiske og nøyaktige modeller for dødelighet.

Å modellere telledata som antall dødsfall ved hjelp av negativ binomialfordeling gir en kraftig metode som tar hensyn til overdispersjon og gir en mer fleksibel tilnærming enn den vanlige Poisson-fordelingen. Denne fordelingen gir oss mulighet til å forstå variasjonen i dødstallene bedre og gir mer realistiske prognoser. Når vi står overfor utfordringer med hensyn til variasjon og kompleksitet i telledata, er den negativ binomiale fordelingen et verdifullt verktøy for å oppnå mer nøyaktige og informative modeller.

Kapittel 3

Den eksponentielle familien og generaliserte lineære modeller

Før vi går videre til dispersjonsmodellering må vi formulerer generaliserte lineære modeller (GLM). GLM bygger videre på teorien om lineære modeller (LM) hvor forskjellen mellom rammeverkene hovedsakelig omfatter at GLM tillater transformasjoner av responser for at antagelser om et lineært forhold mellom respons og forklaringsvariabler skal være tilfredsstilt. GLM gjør det dermed mulig å modellere ikke-lineære sammenhenger, men er i prinsippet fortsatt en lineær modell ettersom den er lineær i parameterne (Dunn, 2018, s. 13).

Allikevel må vi ta til etterretning at responsfordelingene som antas for variabler i GLM ofte ikke er normale slik at vi må være varsomme med hvilke teknikker vi bruker for å avgjøre hvor godt modellen modellerer dataene våre.

3.1 Den eksponentielle familien

For en tilfeldig variabel Y med utfallsrom S, kan den eksponentielle tettheten formuleres som (De Jong, 2008, s. 35)

$$f_Y(y;\theta,\phi) = c(y,\phi)e^{\frac{y\theta-\kappa(\theta)}{\phi}}, \tag{3.1}$$

hvor $\theta \in \Theta$ er den kanoniske parameteren og $\phi > 0$ er dispersjonsparameter. $\kappa(\theta)$ er en kjent funksjon og vi kaller den for kumulantfunksjonen. Domenet Θ omfatter alle verdier som oppfyller $\kappa(\theta) < \infty$. $c(y,\phi)$ er en normaliseringsfunksjon av y og ϕ, og sørger for at $\sum_y f_Y(y;\theta,\phi) = 1$ (Dunn, 2018, s. 212) slik at $f_Y(y;\theta,\phi)$ er en gyldig tetthet fra definisjonen av sannsynlighets tetthetsfunksjoner (Dobson, 2018, s. 53).

Alle sannsynlighetsfordelinger med tetthetsfunksjoner som kan skrives på formen(3.1), er medlemmer av den eksponentielle familien. Valget av $\kappa(\theta)$ og $c(y,\phi)$ avgjør nøyaktig hvilke sannsynlighetsfordeling Y følger (De Jong, 2008, s. 35).

3.1.1 Poisson-fordelingen

Poisson- og negativ binomialfordeling tilhører de eksponentielle fordelingene.

Om $Y \sim \text{Poisson}(\lambda)$ finner vi den eksponentielle formen,

$$f_Y(y;\lambda) = e^{y\log(\lambda)-\lambda-\log(y!)}, \tag{3.2}$$

ved å ta eksponenten av (2.4).

Fra (3.2) kan vi identifisere at den kanoniske linkfunksjonen og kumulantfunksjonen i Poisson-tilfellet henholdsvis er

$$\theta = \log(\lambda), \tag{3.3}$$

og

$$\kappa(\theta) = \lambda, \tag{3.4}$$

med $\phi = 1$.

3.1.2 Negativ binomialfordeling

Om $Y \sim$ Negativ binomial(λ, τ) finner vi den eksponentielle formen,

$$f_Y(y;\lambda,\tau) = e^{y\log\left(\frac{\lambda}{\tau+\lambda}\right)-\tau\log\left(1+\frac{\lambda}{\tau}\right)+\log\Gamma(y+\tau)-\log\Gamma(y+1)-\log\Gamma(\tau)}, \tag{3.5}$$

fra (2.8) (Hardin, 2007, s. 207). Her identifiserer vi at eksponentialfamiliens komponenter er

$$\theta = \log\left(\frac{\lambda}{\lambda+\tau}\right), \tag{3.6}$$

og

$$\kappa(\theta) = \tau\log\left(1+\frac{\lambda}{\tau}\right), \tag{3.7}$$

med $\phi = 1$.

3.2 Momentfunksjon og kumulantfunksjon

Eksponentielle familier har mange viktige og nyttige egenskaper. En av de nyttige egenskapene er at den momentgenererende funksjonen (MGF) alltid kan skrives på en enkel form, selv når tetthetsfunksjonen ikke kan skrives i lukket form (Dunn, 2018, s. 214). Forventningen og variansen kan identifiseres ved hjelp av den momentgenererende funksjonen $M_Y(s)$. Dersom Y er en stokastisk variabel med kumulativ fordelingsfunksjon $F_Y(y)$, så er

$$M_Y(s) = \int_{y\in S} f_Y(y)e^{sy}dy \tag{3.8}$$

MGF for Y ved alle verdier av s hvor forventningen eksisterer.

Den kumulantgenererende funksjonen (CGF) for Y er

$$K_Y(s) = \log M_Y(s) = \log \mathrm{E}\left[e^{sY}\right]. \tag{3.9}$$

En CGF kan brukes for a finne kumulanter av en fordeling. Den en i-te kumulanten for fordelingen til Y er

$$\kappa_i = \frac{d^i}{ds^i}K_Y(s)\bigg|_{s=0}, \tag{3.10}$$

hvor notasjonen indikerer at vi evaluerer den deriverte ved $s = 0$.

Gjennom den momentgenererende funksjonen finner vi at forventning og varians er

$$\mathrm{E}[Y] = \kappa_1 = \frac{d}{ds}K_Y(s)\bigg|_{s=0} \text{ og } \mathrm{V}[Y] = \kappa_2 = \frac{d^2}{ds^2}K_Y(s)\bigg|_{s=0}. \tag{3.11}$$

Når Y er medlem av de eksponentielle familiene med tetthet som i (3.1) finner vi at Y har MGF (Dunn, 2018, 215)

$$\begin{aligned} M_Y(s) &= \mathrm{E}[e^{sY}] \\ &= \int_S e^{sy}c(y,\phi)e^{\frac{y\theta-\kappa(\theta)}{\phi}}dy \\ &= \frac{\kappa(\theta')-\kappa(\theta)}{\phi}\int_S c(y,\phi)e^{\frac{y\theta'-\kappa(\theta')}{\phi}}dy \\ &= \frac{\kappa(\theta')-\kappa(\theta)}{\phi}, \end{aligned} \tag{3.12}$$

hvor vi har brukt at $\int_S c(y,\phi)e^{\frac{y\theta'-\kappa(\theta')}{\phi}}dy = 1$. Med $\theta' = \theta + s\phi$ finner vi at korresponderende CGF er

$$K(s) = \frac{\kappa(\theta+s\phi)-\kappa(\theta)}{\phi}. \tag{3.13}$$

Fra første og andre kumulant (3.11) finner vi forventning og varians (Dunn, 2018, s. 216)

$$\mathrm{E}[Y] = \frac{d}{d\theta}\kappa(\theta), \text{ og } \mathrm{V}[Y] = \phi\frac{d^2}{d\theta^2}\kappa(\theta). \tag{3.14}$$

3.2.1 Poisson-fordelingen

For å finne forventningen bruker vi resultatet i (3.3) til å identifisere at $\lambda = \exp(\theta)$. Ved å erstatte λ i (3.4) og deretter anvenden (3.14) finner vi

$$\frac{d}{d\theta}\kappa(\theta) = \frac{d^2}{d\theta^2}\kappa(\theta) = \exp(\theta) = \lambda, \tag{3.15}$$

som viser at første og andre kumulant er nøyaktig like. Med $\phi = 1$ finner vi

$$\mathrm{E}[Y] = \mathrm{V}[Y] = \lambda. \tag{3.16}$$

Ekvidispersjon beskriver at spredningen om forventningen er den samme, og begrepet blir brukes når vi diskuterer dispersjon i relasjon med Poisson-fordelingen som følger av resultatet i(3.16) .

3.2.2 Negativ binomialfordeling

For negativ binomisk fordeling med $\phi = 1$ finner vi fra (Hardin, 2007, s. 206) at

$$\mathrm{E}[Y] = \frac{d}{d\theta}\kappa(\theta) = \lambda, \quad \text{og} \quad \mathrm{V}[Y] = \frac{d^2}{d\theta^2}\kappa(\theta) = \lambda + \frac{\lambda^2}{\tau}. \tag{3.17}$$

Fra (3.17) er variansen en kvadrisk funksjon av forventning, og dette blir spesielt viktig når vi skal diskutere dispersjonsmodellens funksjon.

3.3 Generaliserte lineære modeller

I 1972 presenterte J. A. Nelder og R. W. M. Wedderburn sin formulering av GLM (og Wedderburn, 1972). Rammeverket av metoder gjør det mulig å tilpasse lineære modeller når antagelse om at forholdet mellom forventningen for en responsvariabel og et utvalg av forklaringsvariabler er lineært gjennom en transformasjon av forventningen. Siden den gang har det blitt utviklet og introdusert nye utvidelser av denne tilnærmingen til parametriske modeller som tar i bruk antagelser om at responsfordelingen er et medlem av den eksponentielle familien.

3.3.1 Notasjon

Vi bruker vektornotasjon med tykk og liten skrift for å formulere modellene, og lar $\boldsymbol{y}$ er responsvektor slik at

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}. \tag{3.18}$$

Om $\boldsymbol{z_i}^\intercal$ $(1 \leq i \leq N)$ er en vektor med forklaringsvariabler (også kalt prediktorer eller kovariater) korresponderende til respons Y_i slik at $\boldsymbol{z_i}^\intercal \in \boldsymbol{Z}$, så indikerer tykk skrift i kombinasjon med stor bokstav at $\boldsymbol{Z}$ er en matrise med forklaringsvariabler, det vil si

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z_1}^\intercal \\ \vdots \\ \boldsymbol{z_N}^\intercal \end{bmatrix}. \tag{3.19}$$

Innholdet i $\boldsymbol{Z}$ avhenger av sammenhengen vi ønsker å modellere, men vi går for enkelthetsskyld ikke nærmere innpå innholdet i denne nå.

3.3.2 GLM

Parametriske metoder gjør antagelser om den funksjonelle formen av f (James, 2013, s. 21). En enkel antagelse er at f er lineær i $\boldsymbol{Z}$. Den normale lineære modellen gir en basis for generaliserte lineære modeller, og en fullverdig forståelse er kritisk for å forstå GLM. Mange av konseptene i GLM bygger på er hentet fra den normale lineære modellen (De Jong, 2008, s. 42). Parametriske modeller antar at f har en parametrisk form, slik at vi kan predikere $\boldsymbol{y}$ fra et sett med parametere $\boldsymbol{\beta}$ og en designmatrise $\boldsymbol{Z}$.

En lineær modell forsøker å estimere $\boldsymbol{y}$ etter forventningen $\mathrm{E}[\boldsymbol{y}]$, hvor vi videre bruker at $\mathrm{E}[\boldsymbol{y}] = \boldsymbol{\lambda}$. En lineær modell kan formuleres som

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\lambda} = \boldsymbol{Z\beta}. \tag{3.20}$$

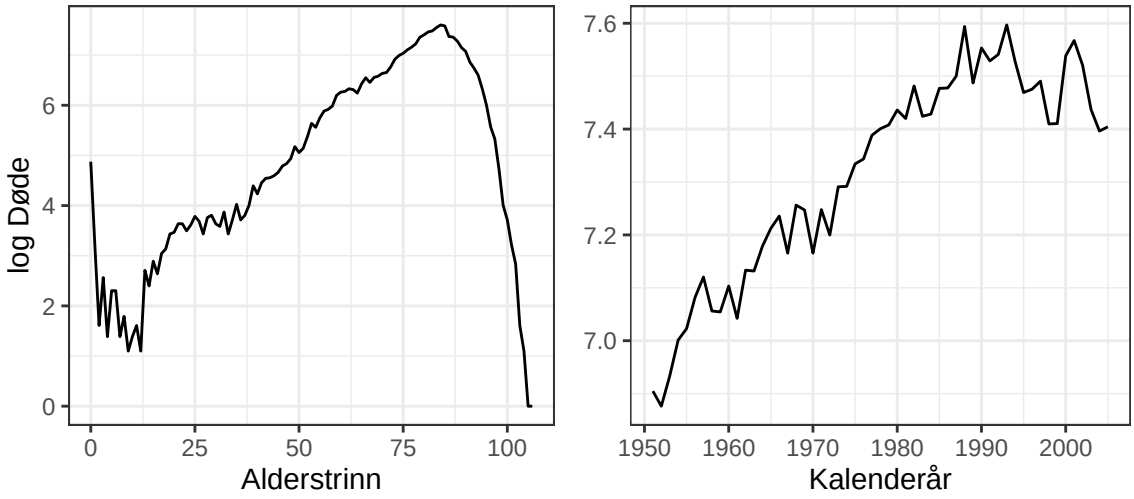
GLM bygger på antagelsene gjort i (3.20) til (De Jong, 2008, s. 35)

$$g(\lambda) = \boldsymbol{Z\beta}. \tag{3.21}$$

Det som skiller (3.20) og (3.21) er funksjonen $g(\cdot)$, som kalles en linkfunksjon, eller transformasjonsfunksjon, og for en GLM må valget av linkfunksjon føre til at $g(\lambda)$ er en monoton og differensierbar funksjon.

Det er viktig å bemerke seg at GLM ikke antar at responsene selv har et lineært forhold med forklaringsvariablene, men at forholdet mellom forventningen lineært forhold eksisterer mellom forventningen og forklaringsvariablene. Det er derfor viktig å utforske forholdene mellom variablene i de dataene man tilpasser modellen med før man velger en linkfunksjon, slik at vi kontrollere at antagelsene vi gjør er rimelige.

For Poisson- og negativ binomiale modeller er en log-link ofte et naturlig valg for g. (1.1) viser at forholdet mellom antall døde og alderstrinn ikke er lineært, mens (3.1) viser at en logaritmen til antall døde har utbedret det ikke-lineære forholdet. Selv om 3.1 antyder at linkfunksjonen bidrar til å utbedre forholdet mellom variablene, og vi kan derfor konkludere med at det er fornuftig å bruke en GLM med log-link.



Figur 3.1: Logaritmen av antall døde som funksjon av alderstrinn og kalenderår.

3.3.3 Dispersjonsmodellering

Blant verktøyene i glmmTMB finner vi dispersjonsmodellen. Dispersjonsmodellen tillater estimering av parametere som ofte blir behandlet som konstanter, og gir oss mulighet til å introdusere et nytt lag av usikkerhet til modellen vår. Om vi antar at $\boldsymbol{y}$ følger en fordeling med dispersjonsparametere $\boldsymbol{\tau}$, og at $\log(\boldsymbol{\tau})$ har et lineært forhold med et sett prediktorer som inngår i designmatrisen $\boldsymbol{Z}^*$, så kan vi modellere dispersjonsparameteren i glmmTMB som

$$\log(\boldsymbol{\tau}) = \boldsymbol{Z}^*\boldsymbol{\beta}^\star, \tag{3.22}$$

hvor $\boldsymbol{\beta}^*$ er en vektor med regresjonsparametere for dispersjon.

For å kunne bruke dispersjonsmodellen så må vi anta at $\boldsymbol{y}$ følger en eksponentiell tetthet med en dispersjonsparameter som kan estimeres. Når vi identifiserte komponentene for Poisson og negativ binomialfordeling brukte vi at $\phi = 1$. Vi kan dermed ikke modellere ϕ i verken av tilfellene, men for negativ binomialfordeling har vi en annen dispersjonsparameter, τ, som kan estimeres da den varierer fritt og inngår i variansen som vist av(3.17).

I denne oppgaven skal vi utforske hvordan vi kan bruke dispersjonsmodellen for å finne en bedre modell, men også for å undersøke hvordan samspillet mellom prediksjoner og varians endrer seg når vi anvender simultan estimering av forventning og dispersjon.

3.3.4 Polynomregresjon

Anta at Y er en tilfeldig variabel med forventning λ og prediktor z.

En enkel GLM for Y kan formuleres som

$$g(\lambda) = \beta_0 + \beta_1 z, \tag{3.23}$$

men dersom forholdet mellom z og $g(\lambda)$ ikke er lineært kan det være nødvendig å tilpasse dataene med en polynomfunksjon(James, 2013, s. 266-267) . Dersom vi tror at forholdet mellom z og $g(\lambda)$ er lineært gjennom en polynomfunksjon av grad a kan vi bytte ut (3.23) med

$$g(\lambda) = \beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3 + ... + \beta_a z^a. \tag{3.24}$$

Ifølge James (2013) er det uvanlig å velge polynomgrader større enn 3 eller 4 da høyere grade kan medføre at kurven blir alt for kompleks. Allikevel vil komplekse data kreve komplekse løsninger. Senere i oppgaven skal vi tilpasse ortogonale polynomer for å modellere forventning og dispersjon.

3.4 Overdispersjon

Overdispersjon utgjør en sentral utfordring i modeller som involverer diskrete responsfamilier (Hardin, 2007, s. 165). Den forekommer når variansen i responsen er større enn antatt, og skyldes ofte positiv korrelasjon mellom responser eller ytterligere variasjon i sannsynlighetene eller frekvensene av responsene. En betydelig konsekvens av overdispersjon er dens evne til å føre til undervurdering av standardfeilen til de estimerte regresjonsparametrene. Dette kan resultere i at en variabel tilsynelatende fremstår som en signifikant prediktor, når den egentlig ikke er det

Ifølge Hardin (2007, s. 165) påvirker overdispersjon utelukkende diskrete modeller, da kontinuerlige modeller har muligheten til å tilpasse dispersjonsparameteren, ϕ. Diskrete modeller mangler denne ekstra parameteren. Selv om kontinuerlige modeller fortsatt forutsetter at variansen er en funksjon av forventningen, gir de muligheten til å skalere forholdet mellom forventning og varians ved hjelp av ϕ. På den annen side har ikke Poisson-modeller denne ekstra parametriske justeringsmuligheten for forholdet mellom variansen og forventningen.

Hardin (2007, s. 176) og Hilbe (2014, s. 37) bruker Pearson-χ^2 statistikken for å måle hvor overdispersert en Poisson-fordeling er. Pearson-χ^2 statistikken, benevnt av Q, er

$$Q = \frac{\chi^2_P}{m - p'}, \tag{3.25}$$

hvor m er antall observasjoner, $p' = p + 1$ parametere, og

$$\chi^2_P = \sum_{i=1}^{m} \frac{(Y_i - \mathrm{E}[Y_i])^2}{\mathrm{V}[Y_i]}. \tag{3.26}$$

χ^2_P er kvadratsummen av Pearson residualer. I kapittel 5 greier vi nøyere ut om Pearson residualene som et verktøy for visualisering og testing av overdispersjon.

Ifølge Hardin (2007, s. 221) indikerer overdispersjon en situasjon der $Q > 1$. Små mengder av overdispersjon er vanligvis av liten bekymring, ettersom helt lik forventning og varians er sjeldent i praksis, men når Pearson-χ^2 er større enn 2 vil det være nødvendig å endre modellen vår (Hardin, 2007, s. 165).

Begrepene 'tilsynelatende' og 'sann' overdispersjon anvendes for å kategorisere de nødvendige endringene for å adressere overdispersjon, slik den er målt av Q. 'Tilsynelatende overdispersjon' omfatter situasjoner der det er mulig å redusere variansen uten å modifisere antagelsene om fordelingen f (Hilbe, 2014, s. 40). Hvis en tilfredsstillende verdi av Q ikke kan oppnås uten å endre f, indikerer det 'reell overdispersjon'.

Begrepet 'tilsynelatende overdispersjon' refererer til situasjoner der modellen viser overdispersjon uten at antagelsene om f er årsaken. På den annen side oppstår 'reell overdispersjon' når overdispersjonen skyldes variasjonen knyttet til antagelsene om f.

Ifølge Hardin (2007, s. 166) skyldes tilsynelatende overdispersjon et av følgende:

1. Modellen utelater viktige prediktorer.
2. Det observerte utvalget inneholder avsidesliggende punkter.

3. Modellen mangler interaksjon.
4. En prediktor må transformeres.
5. Forholdet mellom respons og forklaringsvariabler stemmer

Det vil si at dersom vi møter overdispersjon ved å legge til manglende prediktorer, inkluderer signifikant interaksjon, ekskluderer avsidesliggende observasjoner, og tilfører nødvendige transformasjoner, har vi muligheten til å forbedre modellen når overdispersjonen ikke er reell. Men dersom de øvrige tiltakene ikke gir ønsket effekt, indikerer det at overdispersjonen er reell, og det er dermed nødvendig å endre antagelsene om f.

Når overdispersjon i en Poisson-modell er reell, kan sammensatte Poisson-fordelinger brukes som alternativ (De Jong, 2008, s. 30). Et av disse alternativene er den negative binomialfordelingen. Denne relasjonen kan utledes ved å vise at negativ binomialfordeling oppstår når responsene følger en Poisson-fordeling, hvor forventningene i seg selv følger separate gammafordelinger (Hardin (2007, s. 203), Dunn (2018, s. 399-400), og De Jong (2008, s. 31)).

En negativ binomialfordeling er overdispersert når variansen overstiger $\lambda + \frac{\lambda^2}{\tau}$. Når overdispersjon for en Poisson-modell viser seg å være reell vil en negativ binomial modell kunne gjøre de nødvendige endringene ved å justere standardfeilene da negativ binomialfordeling også egner seg for å modellere telledata, men lar oss justere standardfeil gjennom τ.

Hardin og Hilbe gir oss en enkel definisjon av overdispersjon, men i praksis er det sjelden at vi finner en modell hvor $Q = 1$. I kapittel 5 greier vi nærmere ut om Pearson-χ^2 testen, og sammenligner resultatene fra denne med resultatene fra DHARMa dispersjonstest.

Kapittel 4

Dispersjonsmodellering av svenske dødelighetsdata

Eksemplet *Swedish mortality* fra *Generalized Linear Models for Insurance Data* av De Jong (2008, s. 91-94) illustrerer hvordan negativ binomialfordeling kan forbedre en overdispersert Poisson-modell for svenske dødsfall (menn). Fremgangsmåten involverer endring av parametriske antagelser, kombinert med polynomregresjon. Poisson-modellen blir brukt som utgangspunktet for eksempelet, og den tar i bruk alderstrinn og kalenderår som kategoriske variabler. Etter anvendelse av modifikasjoner argumenterer De Jong (2008, s. 92) for at de har funnet en optimal generalisert lineær modell (GLM) for å predikere dødsfall for svenske menn basert på data fra HMD.

I dette kapittelet vil vi undersøke hvordan dispersjonsmodellen kan være et verdifullt tillegg for å håndtere overdispersjon. I denne sammenhengen vil vi presentere modellene foreslått av De Jong (2008, s. 91-92) for å forutsi svenske dødsfall. Videre vil vi presentere resultatene av tilpasningen av disse modellene ved bruk av svenske populasjonsdata fra perioden 1. januar 1951 til 31. desember 2005.

Deretter vil vi formulere egne forslag til negativ binomiale modeller, der dispersjonsparametrene $\boldsymbol{\tau}$ modelleres ved hjelp av en eller flere prediktorvariabler gjennom en log-lineær transformasjon. De nye forslagene, som innebærer simultan modellering av både $\boldsymbol{\lambda}$ og $\boldsymbol{\tau}$, vil her refereres til som utvidede negativ binomiale modeller.

Vi begynner med å tilpasse fem modeller der forventningen predikeres av kategoriske forklaringsvariabler. Dette innebærer at designmatrisen $\boldsymbol{Z}$ består av indikatorvariabler. De kategoriske modellene inkluderer en Poisson-modell, en negativ binomial modell med konstant dispersjonsparameter, samt tre utvidede negativ binomiale modeller.

Deretter vil vi diskutere seleksjonsprosessen som ble gjennomført for å undersøke hvilke polynomfunksjoner som best egnet seg for en negativ binomial modell, der dispersjonsparameteren holdes konstant, etterfulgt av en utvidet variant.

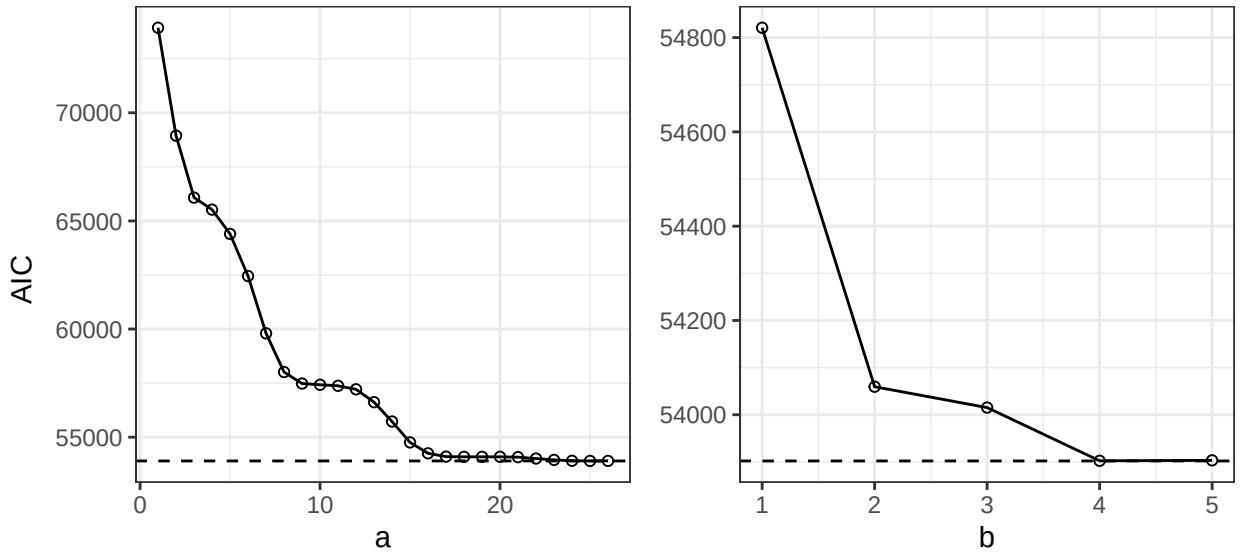
Underveis vil vi presentere resultater for vurderingskriterier, og regresjonsparameterne for dispersjonsmodellen vil bli formidlet. Vi vil også gi kommentarer til prediksjonene basert på svenske populasjonsdata fra 1. januar 2006 til 31. desember 2022 for utvalgte modeller.

Disse modellene danner grunnlaget for å utforske problemstillinger knyttet til identifisering av overdispersjon gjennom visualisering og testing av residualene. Denne analysen vil være sentral i diskusjonen om hvordan modellering av dispersjon kan være et effektivt verktøy for å håndtere overdispersjon i datasettet.

4.1 Modellseleksjon

Tilnærmingen for modellering bygger på algoritmene for parameterseleksjon som er beskrevet av James (2013, s. 205). Disse algoritmene innebærer å tilpasse en modell for hver kombinasjon av tilgjengelige prediktorer i datasettet. Prosessen resulterer i flere modeller, alle tilpasset de samme dataene. For å velge den mest hensiktsmessige modellen, vurderer vi kriterier som vi selv definerer. Det er derfor avgjørende å velge et kriterium som passer til hvordan vi planlegger å anvende resultatene.

Når vi gjør tilpasninger, forblir formelen for forventningen uendret fra modellene beskrevet av De Jong (2008, s. 91-93). Ettersom utvidelsene er nestet i det som er beskrevet i Generalized Linear Models for



Figur 4.1: AIC fra modeller som ble tilpasse under seleksjonsprosessen, som funksjon av polynomgrad a for alderstrinn og b for kalenderår. Til venstre ser vi AIC fra modeller med $b = 4$, og til høyre for modeller med $a = 25$.

Insurance Data, henviser vi kun til formelen for forventningen før vi presenterer modellen for dispersjon.

Seleksjonsprosessen begynner med å vurdere en Poisson-modell, M_0, etterfulgt av en negativ binomisk modell, M_1. Dispersjonsparameteren for M_1 holdes konstant for alle prediksjoner, og utvidelsene, M_2, M_3, og M_4, søker forbedringer ved å modellere dispersjonsparameteren ved å bruke enkle kombinasjoner av alderstrinn og kalenderår som prediktorer.

Deretter tilpasser vi en modell for hver kombinasjon av polynomer av alderstrinn og kalenderår, først for modellen med konstant dispersjon, M_5, og deretter for utvidelsen som inkluderer en polynomfunksjon for dispersjon, M_6. glmmTMB møtte konvergensproblemer rundt grad 29 for alderstrinn og 8 for kalenderår i formelen for forventning i M_5. For konvergensproblemer ble begrensninger satt til grad 10 for alderstrinn og 8 for kalenderår i formelen for dispersjon.

Modellene vil blant annet vurderes i henhold til Akaike Information Criterion (AIC). Ideen bak AIC er å vurdere modellens kompleksitet ved å betrakte hvor tett modellen passer dataene den tilpasset med (Kaas, 2008, s. 248). En modell med mange parametere vil gi oss en veldig god tilpasning til dataene den er tilpasset med, men vil ha færre frihetsgrader som fører at den i praksis kan ha begrenset nytte. Den balanserte tilnærmingen fraråder overtilpassning, og belønner sparsommelighet.

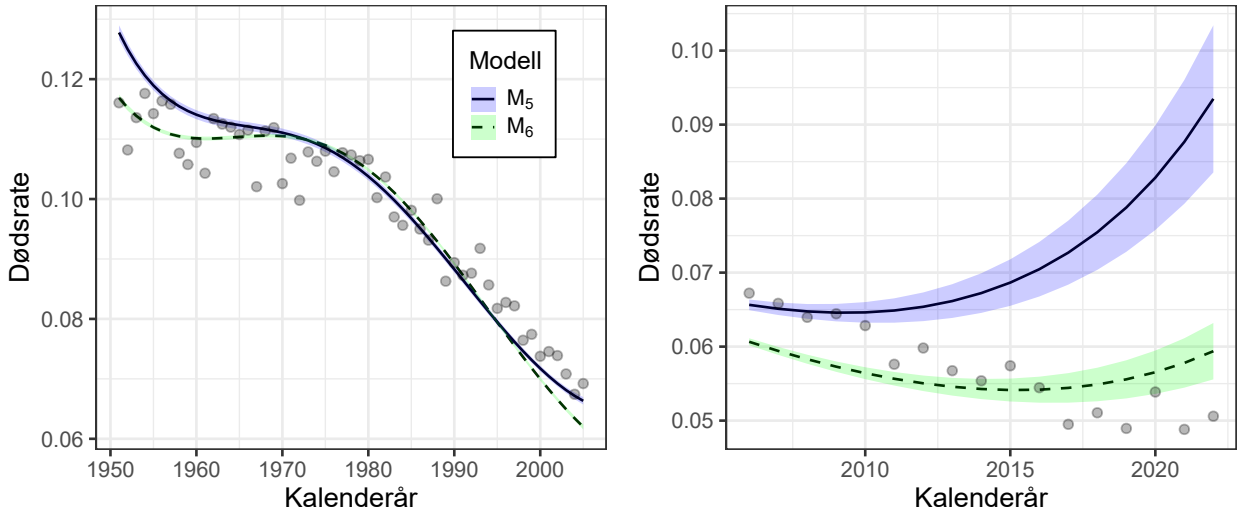
Vi foretrekker derfor AIC fordi en modell som er mer representativ for den underliggende fordelingen av data vil ha bedre evne til å informere oss om datapunkter den ikke har observert, så lenge modellen ikke avviker for mye fra dataene den er tilpasset. Formelen for AIC er

$$AIC = -2\mathcal{L} + 2p', \tag{4.1}$$

hvor $\mathcal{L}$ er log-likelihood og p' er antall parametere.

4.2 Kategoriske modeller

Vi skal nå formulere modeller for antall dødsfall, $\boldsymbol{y}$, med alderstrinn, $\boldsymbol{x}$, og kalenderår, $\boldsymbol{t}$, som prediktorer. I denne seksjonen inkluderes forklaringsvariablene som indikatorvariabler i designmatrisen $\boldsymbol{Z}$. Referansenivåene er satt til alderstrinn 0 og kalenderår 1951. De kategoriske modellene bruker 110 faktornivåer for alderstrinn og 55 for kalenderår. Størrelsen på middelbefolkningen, $\boldsymbol{n}$, brukes som justeringsvariabel. Modellene trenes på data fra 1951 til 2005, noe som innebærer at modelltilpasningene gjøres med $N = 5864$ observasjoner.



(a) M_5 **(b)** M_6

Figur 4.2: Predikert dødsrate per person i middelbefolkningen for 80-åringer fra de polynomiske modellene, inkludert konfidensintervaller. De sorte punktene er observasjoner fra den samme perioden.

M_0 :

Om vi mener at $\boldsymbol{y} \sim \text{Poisson}(\boldsymbol{\lambda})$ kan vi formulere en Poisson-modell som

$$\log(\boldsymbol{\lambda}) = \log(\boldsymbol{n}) + \boldsymbol{Z\beta}. \tag{4.2}$$

M_0 gir oss 109 regresjonsparametere korresponderende til alderstrinn og 54 parametere korresponderende til kalenderår. Indikatornivåene for forklaringsvariablene er alderstrinn 0 og kalenderår 1951. Med en log-likelihood på -30468 blir modellens AIC målt til 61266.

M_1 :

Tidligere har vi nevnt at negativ binomialfordeling egner seg bedre til å modellere data som varierer mer enn forventet av Poisson modellen. Ved å anta at

$$\boldsymbol{y} \sim \text{Negativ binomial}(\boldsymbol{\lambda}, \tau), \tag{4.3}$$

kan vi beholde formelen for predikert forventning slik den er oppgitt i (4.2). $\boldsymbol{Z}$ er også fortsatt den samme. Endring av f har kun konsekvenser for regresjonsparameterne, $\boldsymbol{\beta}$.

Spesielt for glmmTMB estimeres én verdi for τ dersom det ikke oppgis forklaringsvariabler i formelen for dispersjons. τ modelleres dermed gjennom en enkel dispersjonsmodell,

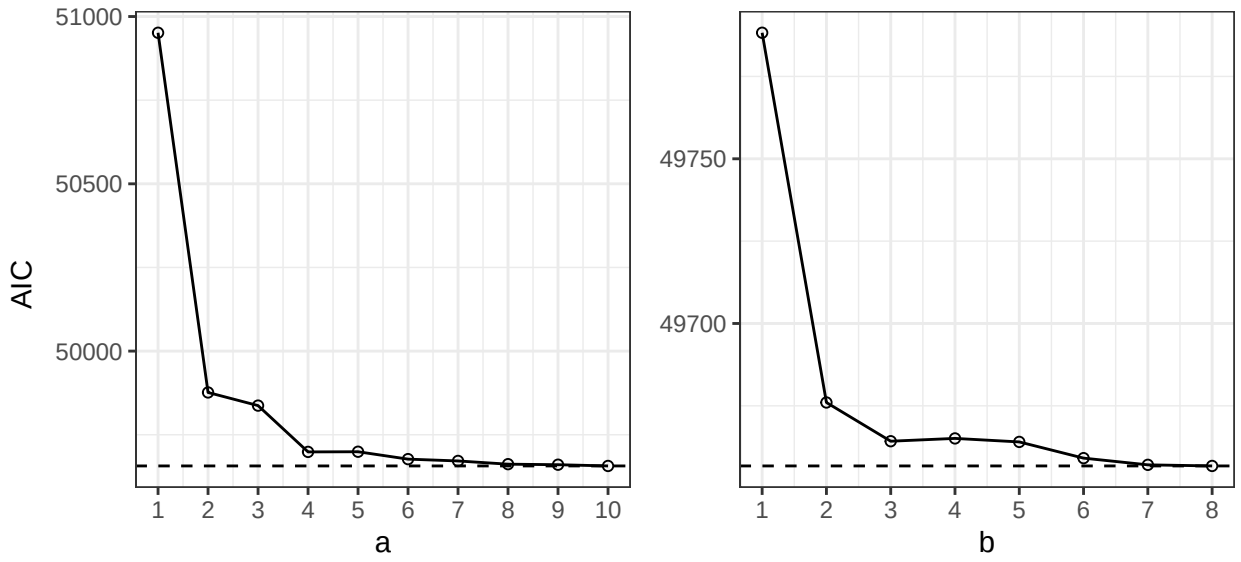
$$\log(\tau) = \beta_0^*, \tag{4.4}$$

hvor β_0^* er skjæringskoeffisienten, og vi tolker τ som maksimalt sannsynlighets estimat (MLE) eller gjennomsnittlig dispersjon for modellert forventning. Estimering av β_0^* medfører at M_1 har 165 parametere, sammenlignet med M_0 som har 164. Log-likelihood for M_1 er -26849, kombinert med 165 parameter gir det en AIC lik 54027. Forandringen av parametrisk antagelsen medfører betydelig reduksjon av AIC sammenlignet med M_0. Med en differanse på 7239 AIC fremgår det tydelig i relasjon med (4.1) at reduksjonen skyldes en større verdi av $\mathcal{L}$.

4.2.1 Utvidede negativ binomiale modeller

Som påpekt av De Jong (2008, s. 91), er M_0 overdispersert, og endringene M_1 gir en form av f som justerer variasjonen gjennom samspillet mellom $\boldsymbol{\lambda}$ og τ. Utvidelsene av M_1 gjør antagelser om at heterogen dispersjonsparameter leder til videre reduksjon av standardfeil. Det innebærer derfor at

$$\boldsymbol{y} \sim \text{Negativ binomial}(\boldsymbol{\lambda}, \boldsymbol{\tau}), \tag{4.5}$$



Figur 4.3: AIC fra seleksjonsprosess som ble gjennomført for å forsøke å finne den optimale polynomiske kurven. Til venstre ser vi AIC fra modeller med dispersjonsmodeller som tilpasset dataene til kurver av grad $b = 8$ til venstre, og til $a = 10$ til høyre.

hvor $\boldsymbol{\tau}$ er en vektor med N dispersjonsparametere. (4.2) inngår på samme måte i de følgende kategoriske modellene, men vi endrer (4.4) til å inkludere prediktorer.

M_2 :

Dersom vi antar at $\log(\boldsymbol{\tau})$ endrer seg som en lineær funksjon av kalenderår, kan en modell formuleres som

$$\log(\boldsymbol{\tau}) = \beta_0^\star + \beta_1^\star \boldsymbol{t}, \tag{4.6}$$

hvor

$$\boldsymbol{t} = \begin{bmatrix} 1951 \\ 1952 \\ \vdots \\ 2005 \end{bmatrix}. \tag{4.7}$$

M_3 :

Alternativt kan vi anta et lineært forhold mellom logaritmen av dispersjonsparameteren og alderstrinn, slik at

$$\log(\boldsymbol{\tau}) = \beta_0^\star + \beta_2^\star \boldsymbol{x}, \tag{4.8}$$

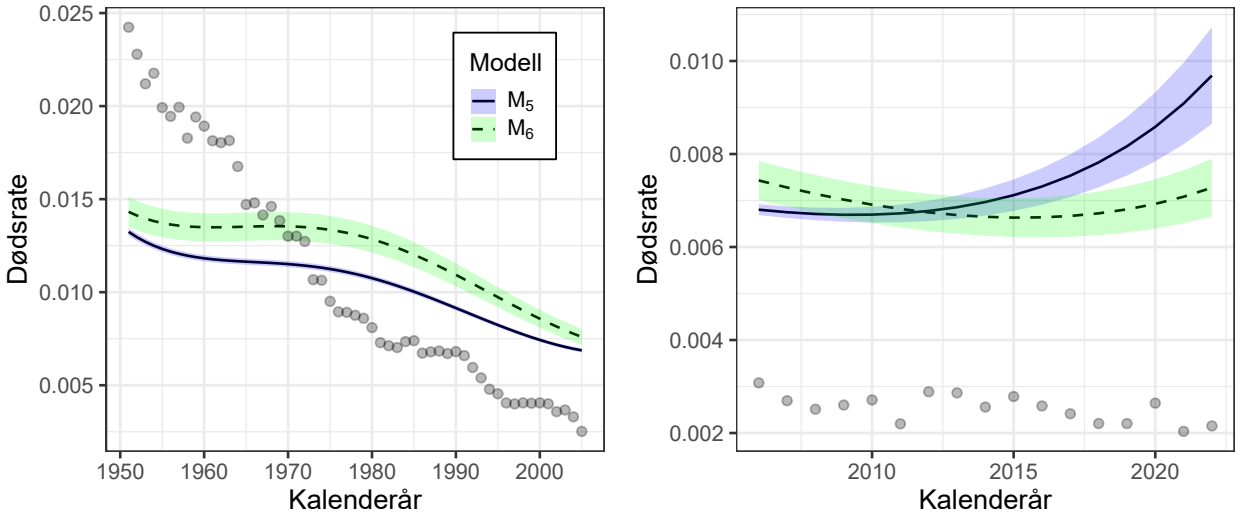
hvor

$$\boldsymbol{x} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 109 \end{bmatrix}. \tag{4.9}$$

M_4 :

Den fulle kombinasjonen av prediktorer inkluderes gir oss

$$\log(\boldsymbol{\tau}) = \beta_0^\star + \beta_1^\star \boldsymbol{t} + \beta_2^\star \boldsymbol{x}. \tag{4.10}$$



(a) M_5 **(b)** M_6

Figur 4.4: Predikert dødsrate for 0-åringer fra polynomiske modeller, med konfidensintervaller. De sorte punktene er observert dødsrate.

Til sammenligning mellom modellene finner vi at M_2 har en AIC på 53837, M_3 har en AIC på 52067, og M_4 har en AIC på 51163. Den betydelige forskjellen mellom M_2 og M_3 indikerer at det utgjør en stor forskjell å estimere dispersjonsparameter avhengig av alderstrinn. Fra M_4 ser vi at alderstrinn kombinert med kalenderår resulterer i det beste alternativet med god margin når prediktorvariablene er oppgitt som faktornivå. Sammenlignet med M_1 ser vi at dispersjonsmodellen har bidratt til en nedgang i AIC på 2864.

4.3 Polynomiske modeller

M_5 :

For å finne en optimal polynomisk modell ble det gjennomført en trinnvis seleksjon av prediktorkombinasjoner. Prosessen involverte tilpasning av polynomiske modeller med kombinasjoner av grader fra 1 til 26 for alderstrinn og 1 til 5 for kalenderår.

4.1 viser AIC for enkelte kombinasjoner av polynomgrader. Vi kan observere at modeller med høyere polynomgrader for prediktorene gir bedre resultater, med lavere AIC, men trenden endret seg etter grad 25 for alderstrinn og grad 5 for kalenderår. Resultatene viser at polynomgradene for den optimale modellen samsvarer med modellen foreslått av De Jong (2008, s. 69), og dette er en modell med polynomer av grad 25 for alderstrinn og 4 for kalenderår. Modellen kan formuleres som

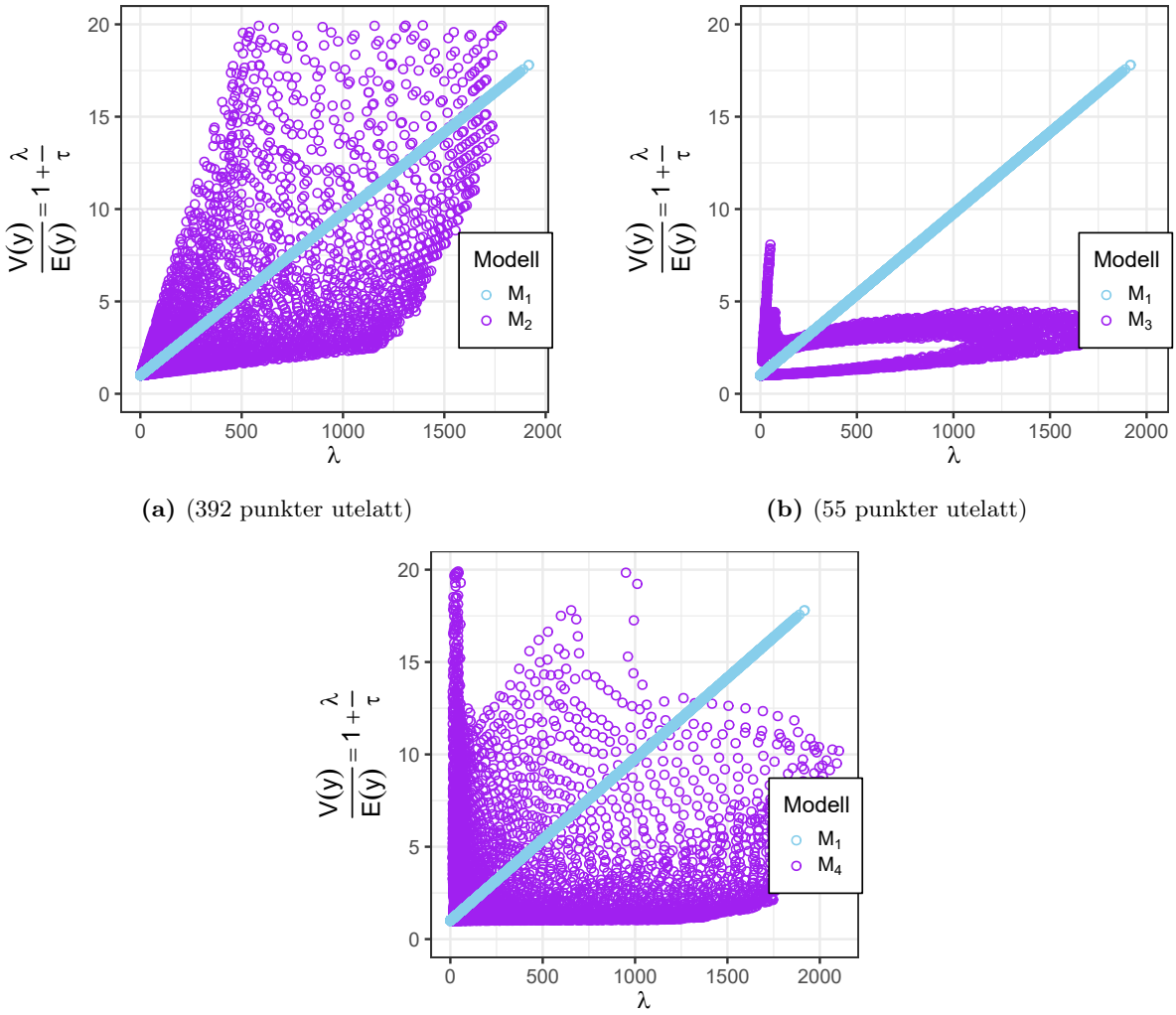
$$\boldsymbol{y} \sim \text{Negativ binomial}\,(\boldsymbol{\lambda}), \quad \log(\boldsymbol{\lambda}) = \log(\boldsymbol{n}) + \beta_0 + \beta_1\boldsymbol{x} + \ldots + \beta_{25}\boldsymbol{x}^{25} + \beta_{(26)}\boldsymbol{t} + \ldots + \beta_{(29)}\boldsymbol{t}^4. \quad (4.11)$$

M_5 har AIC lik 53902 hvor den største andelen av differansen mellom M_5 og M_1 skyldes antall parametere.

M_6 :

For å optimalisere dispersjonsmodellen ved hjelp av polynomregresjon gjennomførte vi seleksjon ved å tilpasse modeller som er nestet fra M_5, men med dispersjonsmodeller sammensatt av hver sin unike kombinasjon av polynomer med grad 1 til 10 for alderstrinn og 1 til 8 for kalenderår. Blant de tilpassede nestede modellene hadde dispersjonsmodellen med polynomgrad 10 for alderstrinn og 8 for kalenderår lavest AIC. Dette var også modellen med flest antall parametere, noe som førte til konvergensproblemer. 4.3 viser tilhørende AIC for alle kombinasjoner av gradene fra den endelige modellen. Modellen følger formuleringen gitt av 4.11, men modellerer også dispersjonsparameteren som

$$\log(\boldsymbol{\tau}) = \beta_0^\star + \beta_1^\star\boldsymbol{x} + \ldots + \beta_{10}^\star\boldsymbol{x}^{10} + \beta_{(11)}^\star\boldsymbol{t} + \ldots + \beta_{(18)}^\star\boldsymbol{t}^8. \quad (4.12)$$



(a) (392 punkter utelatt)

(b) (55 punkter utelatt)

(c) (102 punkter utelatt)

Figur 4.5: Sammenligning av Poisson-overdispersjon modellert av M_2, M_3 og M_4 sammenlignet med M_1.

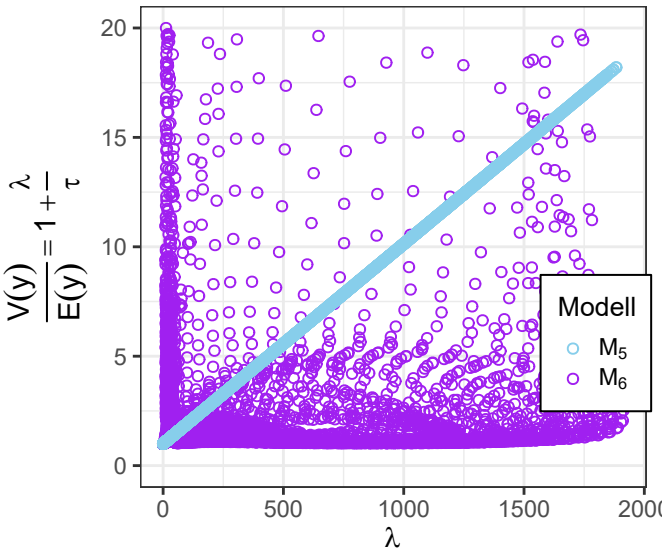
Modellen gir har en loglikelihood på -24779.37 har M_6 en AIC på 49657, og antyder på å være den beste modellen med AIC som seleksjonskriterium. 4.4 viser at M_6 predikerer lavere dødsrate enn M_5. I kontrast til de kategoriske modellene ser ut som dispersjonsmodellen hjelper oss å finne en modell med lavere dødsrate.

Vi har dermed funnet at M_2, M_3, og M_4 måler bedre resultater med utgangspunkt i seleksjonskriteriet, og at M_6 kan erstatte det De Jong (2008) beskriver som optimal modell.

Modellen, med en log-likelihood på -24779.37, har M_6 en AIC på 49657, og antyder å være den beste modellen med AIC som seleksjonskriterium. 4.4 viser at M_6 predikerer lavere dødsrate enn M_5. I kontrast til de kategoriske modellene ser det ut som dispersjonsmodellen hjelper oss å finne en modell med lavere dødsrate.

Vi har dermed funnet at M_2, M_3, og M_4 gir bedre resultater med utgangspunkt i seleksjonskriteriet, og at M_6 kan erstatte det De Jong (2008) beskriver som optimal modell.

For å undersøke fremtidige prognoser fra M_5 og M_6, kan vi bruke data fra 2006 til 2022 for å måle prediksjonsstyrken. 4.2b og 4.4b viser henholdsvis den predikerte dødsraten for 80- og 0-åringer som funksjon av kalenderår. De sorte sirklene viser den observerte dødsraten for samme periode. Vi kan fastslå at kurvenes helninger antyder høyere dødssannsynlighet for M_5. Dette vil vise seg å utgjøre forskjell når vi estimerer forventet dødelighet fra modellene i kapittel 56.



Figur 4.6: Sammenligning av dispersjon mellom M_5 og M_6 (204 punkter utelatt).

4.4 Dispersjon

Dispersjonsmodellen lar dispersjonsparameteren variere på tvers av observasjoner i stedet for å anta at den er konstant for alle observasjoner i modellen. Dette gir oss mulighet til å identifisere kildene til dispersjonen ved å se på parameterestimatene for de kategoriske modellene i 4.1.

Når vi tolker overdispersjon, kan det være enklere å forstå effekten av dispersjonsparameteren ved å se på hvordan interaksjonen mellom $\boldsymbol{\lambda}$ og $\boldsymbol{\tau}$ påvirker $\mathrm{V}[\boldsymbol{y}]$. Vi fant tidligere ut at formelen for variansen i den negative binomiske modellen var

$$\mathrm{V}[\boldsymbol{y}] = \boldsymbol{\lambda} + \frac{\boldsymbol{\lambda}^2}{\boldsymbol{\tau}}. \tag{4.13}$$

For M_2 avtar $\boldsymbol{\tau}$ med kalenderår, slik at $\mathrm{V}[\boldsymbol{y}]$ øker med kalenderår. Derimot avtar dispersjonsparameteren for alderstrinn i både M_3. Dermed kan vi tolke at variansen for M_4 vil avta med alderstrinn, men øker med høyere kalenderår. Den største forskjellen ser vi at oppstår mellom skjæringsparameterene, som absolutt er avgjørende for estimert prediksjonsvarians.

Vi har tidligere omtalt Poisson-fordelingen som et grensetilfelle av negativ binomisk fordeling med $\frac{1}{\boldsymbol{\tau}} = 0$. Fordelen med å bruke negativ binomisk fordeling var at vi kunne estimere en dispersjonsparameter, og vi har nå brukt dispersjonsmodellen til å vekte dispersjonen for å forklare variasjonen ytterligere enn hva en parameter klarer.

Forholdet mellom forventning og varians i de negativt binomiale modellene er gitt ved

$$\frac{\mathrm{V}[\boldsymbol{y}]}{\mathrm{E}[\boldsymbol{y}]} = 1 + \frac{\boldsymbol{\lambda}}{\boldsymbol{\tau}}. \tag{4.14}$$

4.5 viser hvordan dispersjonen om forventningen blir vektet forskjellig blant de kategorielle modellene i forhold til M_1. Vi ser at forholdet endres slik at mindre prediksjoner varierer mer, og større prediksjoner varierer mindre i M_4. 4.6 viser at trendene for dispersjonen i M_6 er det samme som i M_5. Spredningen av det proporsjonelle forholdet fra (4.14) viser seg meget kompleks for M_4 i4.5c og M_6 i 4.6.

	β_0^*	β_1^*(Kalenderår)	β_2^*(Alderstrinn)
M_1	114.115	0	0
M_2	127.45	−0.062	0
M_3	1.935	0	0.059
M_4	191.425	−0.096	0.071

Tabell 4.1: Parameter estimater for dispersjon fra de kategoriske modellene.

Kapittel 5

Residualanalyse

I kapittel 4 gjorde vi endringer i våre antakelser om f ved å gå fra en Poisson-tilnærming til en negativ binomial modell. Deretter justerte vi dispersjonsparameteren avhengig av prediktorvariablene. I den initielle vurderingen av modellene brukte vi AIC som utgangspunkt.

I dette kapittelet vil vi utforske metoder for å teste og avdekke overdispersjon. (Hartig, 2022a) presenterer en simuleringsbasert tilnærming for å identifisere overdispersjon. Før vi går nærmere inn på dette, vil vi undersøke Pearson-residualer og presentere det teoretiske grunnlaget for residualer. Det finnes flere typer residualer, men i denne sammenhengen vil vi fokusere på tre typer: responsresidualer, Pearson-residualer og kvantil-residualer.

Alle residualer har sitt utgangspunkt i responsresidualene. Responsresidualer kan formuleres som

$$\boldsymbol{r} = \boldsymbol{y} - \mathrm{E}[\boldsymbol{y}], \tag{5.1}$$

hvor $\boldsymbol{y}$ er en responsvektor med forventning $\mathrm{E}[\boldsymbol{y}]$ (Dunn, 2018, s. 298). Respons residualene oppgir avstanden mellom utfall og prediksjon, og er ifølge Dunn (2018, s. 298) ikke spesielt nyttige når vi analyserer antagelser av GLM.

5.1 Pearson residualer

Et Pearson residual er ifølge Dunn (2018, s. 299) og Hilbe (2014, s. 78)

$$r_i^{(P)} = \frac{Y_i - \mathrm{E}[Y_i]}{\sqrt{\mathrm{V}[Y_i]}}. \tag{5.2}$$

Pearson residualene standardiserer responsresidualet ved å bruke kvadratroten av estimert varians som estimator for standardavviket.

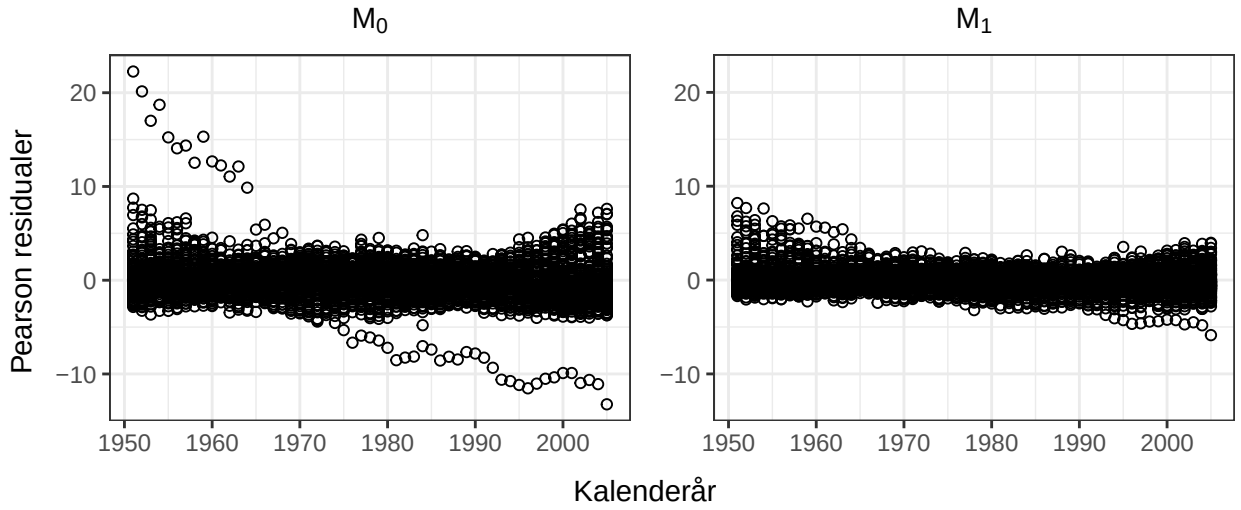
Pearson residualene for en Poisson-modell er (Dunn, 2018, s. 300)

$$\boldsymbol{r}^{(P)} = \frac{\boldsymbol{y} - \boldsymbol{\lambda}}{\sqrt{\boldsymbol{\lambda}}}. \tag{5.3}$$

De standardiserte residualene er omtrent normalfordelte, selv om de vanligvis ikke er uavhengige (Dobson, 2018, s. 24). Et intuitivt argument er at dersom $r_i^{(P)} \sim \text{Normal}(0, 1)$ så er $(r_i^{(P)})^2 \sim \chi^2(1)$ slik at (Dobson,

	AIC
M_0	61265
M_1	54027
M_2	53837
M_3	52067
M_4	51163
M_5	53902
M_6	49657

Tabell 5.1: AIC for modellene som ble tilpasset i kapittel 4.



Figur 5.1: Pearson residualer som funksjon av kalenderår for M_0 og M_1

2018, s. 25)

$$\sum_{i=1}^{N}\left(r_i^{(P)}\right)^2 = \sum_{i=1}^{N}\left(\frac{Y_i-\lambda_i}{\sqrt{\lambda_i}}\right)^2 \sim \chi^2(N-p'), \tag{5.4}$$

hvor N er antall observasjoner, og p' er antall parametere.

Ifølge Dobson (2018, s. 200) har vi relasjonen

$$\chi^2 = \sum_{i=1}^{N}\left(r_i^{(P)}\right)^2, \tag{5.5}$$

hvor χ^2 er kji-kvadrat test-statistikken, som kan brukes for å teste passform (Goodness of fit) av telledata, og formulerer χ^2 som (Dobson, 2018, s. 25)

$$\chi^2 = \sum_{i=1}^{N}\frac{(o_i-e_i)^2}{e_i} \sim \chi^2(N-p'), \tag{5.6}$$

hvor o_i er observert antall med forventning e_i.

(5.6) er ofte satt i sammenheng med testing av homogenitet i krysstabeller (Devore, 2012, s. 749), men egner seg til å teste passform av GLM om man setter χ^2 i sammenheng maksimal sannsynlighets estimering(Dobson, 2018, s. 162-163).

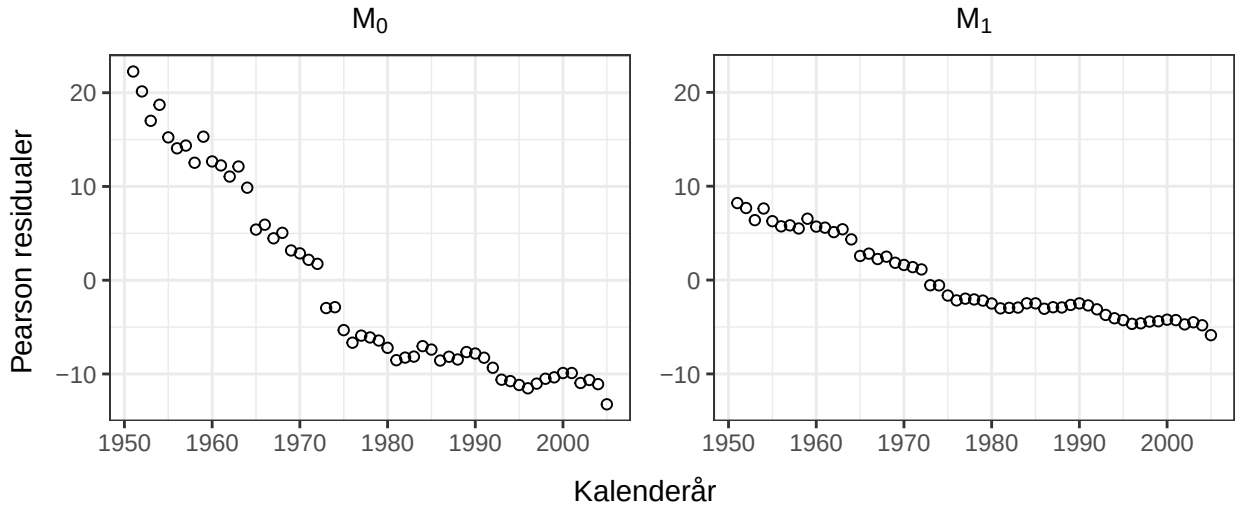
For en god modell antar vi at de standardiserte residualene vil være omtrent standardnormalfordelte, det vil være svært gjenkjennelig på et residualplot. I den perfekte modellen vil avvik fra denne antagelsen vise seg som standardiserte ureduserbare feil $\frac{\boldsymbol{\sigma}_\epsilon}{\sqrt{\mathrm{V}[\boldsymbol{y}]}}$.

For en negativ binomial modell er Pearson residualene

$$\boldsymbol{r}^{(P)} = \frac{\boldsymbol{y}-\boldsymbol{\lambda}}{\sqrt{\boldsymbol{\lambda}\left(1-\frac{\boldsymbol{\lambda}}{\boldsymbol{\tau}}\right)}}. \tag{5.7}$$

I figur 5.1 fremstilles Pearson-residualene som funksjon av kalenderår for M_0 og M_1. Det er tydelig i figuren at det eksisterer betydelig korrelasjon mellom nærliggende residualer, et fenomen beskrevet av James (2013, s. 94) som sporing. Sporing oppstår når nærliggende residualer er korrelerte, og dette manifesterer seg som synlige 'spor' når residualene vises som en funksjon av kalenderår. For modell M_0 er det klart synlig et spor som strekker seg gjennom tid. Korrelasjonsmønsteret er derimot ikke like tydelig for residualene fra M_1

Ved en nærmere undersøkelse av residualene avdekket vi at korrelasjonsmønsteret skyldes prediksjoner for alderstrinn 0. En isolert sammenligning for dette alderstrinnet er vist i figur 5.2. Det er tydelig at spredningen er betydelig redusert hos M_0. Ifølge ligning (5.2) kan vi anta at spredningen oppstår når avstanden mellom $\boldsymbol{y}$ og $\boldsymbol{\lambda}$ er vesentlig større enn standardavviket. Ved å nøye inspeksjon av de vertikale aksene i 5.2 legger vi



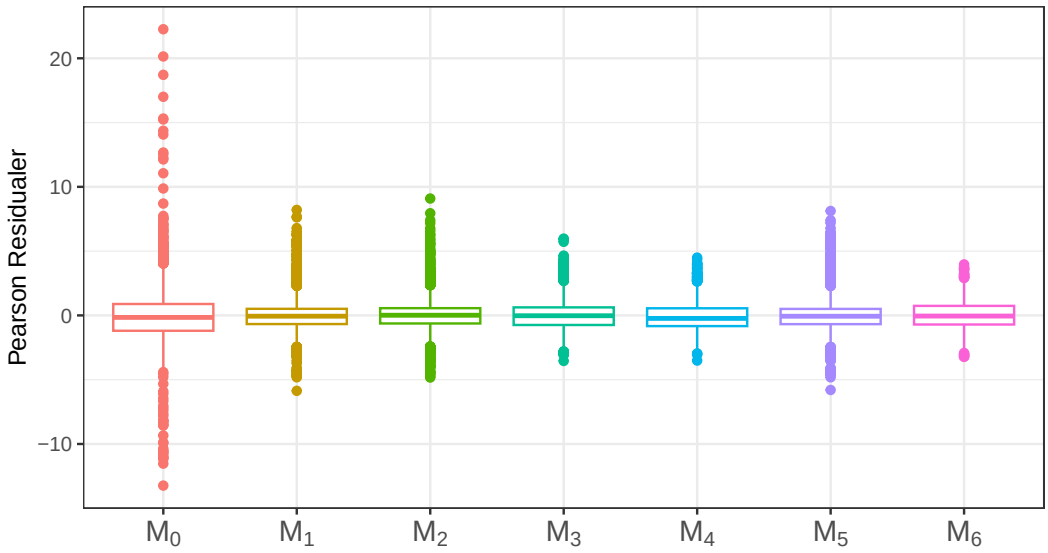
Figur 5.2: Pearson residualene for alderstrinn 0 for M_0 og M_1 preges av *sporing*.

også merke til at residualene strekker seg i overvekt mot høyere verdier. Dette indikerer klart at spredningen blant residualene avviker fra det som ifølge Dobson (2018, s. 25) skulle være en standard normalfordeling. Denne overvekten gir grunn til å mistenke at M_0 er overdispersert.

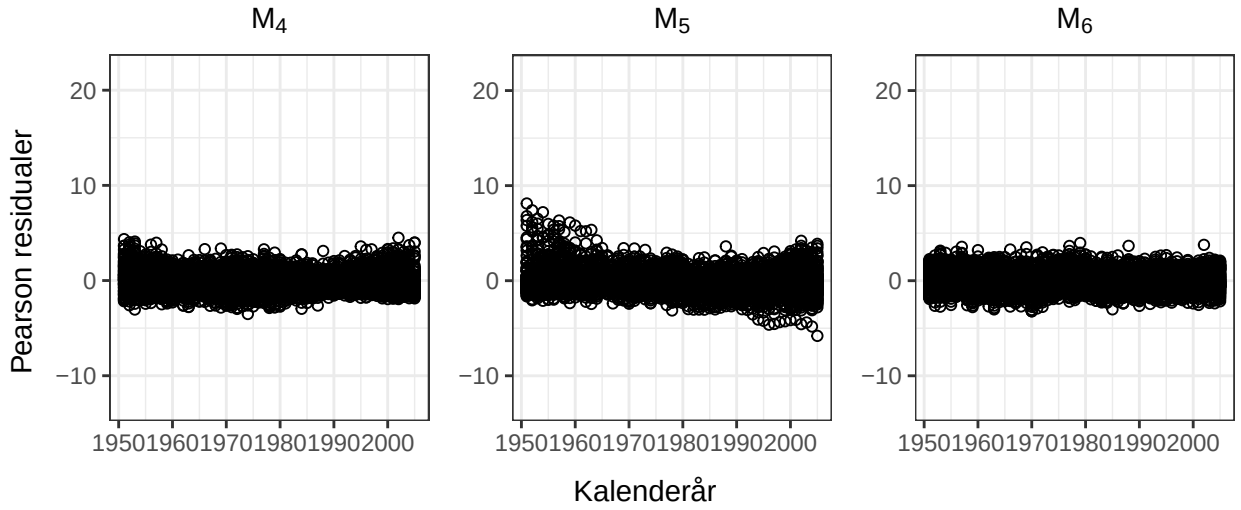
Dermed kan vi argumentere med hensyn til residualene for M_0 at antagelsene som modellen bygger på, er feil. Ved å vurdere at M_1 reduserer spredningen, antar vi at en negativ binomialfordeling forklarer avstanden mellom $\boldsymbol{y}$ og $\boldsymbol{\lambda}$ i større grad. Allikevel er spredningen fortsatt større enn forventet.

5.3 gir en oppsummering av spredningen blant residualene for hver av modellene. Figuren muliggjør en enkel sammenligning av spredningen blant residualene for hver modell. Det er tydelig at spredningen blant residualene fra modell M_0 er betydelig større enn de øvrige modellene, mens spredningen er minst for M_4 og M_6. Det er verdt å merke seg at det er en tydelig sammenheng mellom spredningen i 5.3 og resultatene for AIC i 5.1.

I 5.4 er residualene for modell M_4 presentert, og det er tydelig at de er mer jevnt fordelt sammenlignet med M_0 og M_1 i 5.1. Antydningen om at residualene for M_4 er mer normaliserte gir et sterkt argument for å foretrekke M_4 over M_1. 5.4 viser også residualene fra modell M_5 og M_6. Spredningen av residualene fra modell M_5 ligner i stor grad på det vi observerer blant residualene fra modell M_1. Både M_1 og M_5 har konstant dispersjonsparameter, og det kan være rimelig å anta at transformasjonen av prediktorvariablene



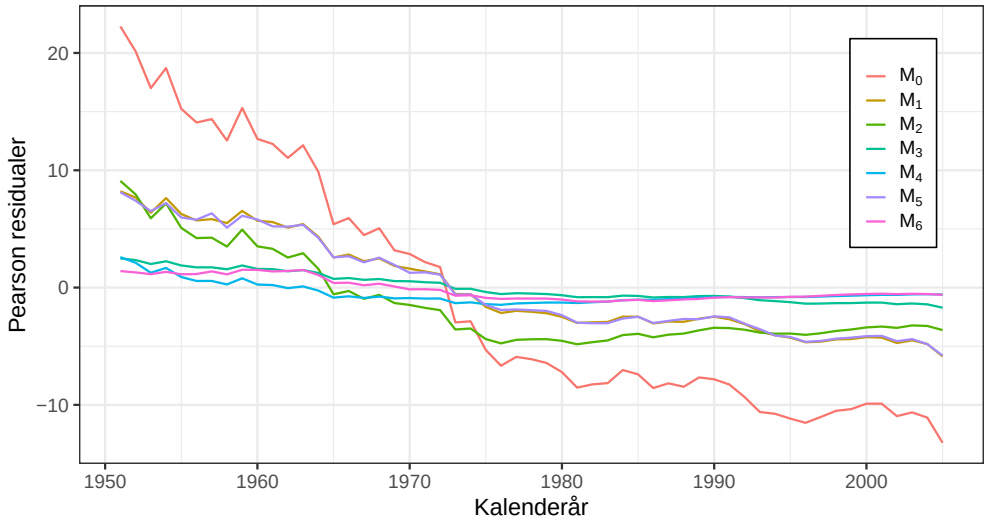
Figur 5.3: Sammenligning av standardiserte residualer mellom hver av modellene.



Figur 5.4: Pearson residualer for M_4, M_5 og M_6.

ikke bidrar til betydelige endringer i residualene. Derimot viser M_6 en markant endring som ligner en normal spredning.

Vi avslutter undersøkelsen av Pearson-residualer ved å vurdere 5.5, som viser Pearson-residualene for alderstrinn 0 isolert sett for hver modell som funksjon av kalenderår. Forskjellene mellom parametriske antagelser blir tydelige ved sammenligning av M_0 og M_1. Vi observerer at endringer i prediktorvariabler hadde begrenset effekt på spredningen for M_1 og M_5, da residualene delvis overlapper. Ved å anvende en dispersjonsmodell for å modellere spredning ser vi at antagelsene om normalitet er bedre oppfylt ved å sammenligne resultatene for M_1 i 5.1 med M_4 i 5.4, samt M_5 med M_6 i 5.4.



Figur 5.5: Linjeplot med residualer for alderstrinn 0 fra modellene som funksjon av kalenderår.

5.2 Simulerte residualer

Det sier seg selv at å argumentere for nivået av dispersjon ved 1 av 110 alderstrinn ikke gir oss rettferdig helhetsvurdering, men det vil bli mye arbeid dersom vi skal utlede samsvarende vurderinger for hver av de resterende 109 alderstrinnene. Vi forsøkte å gjøre oss meninger om spredning i 5.1 og 5.4, men å bedømme dispersjon på bakgrunn av spredning langs den vertikale aksen er utfordrende med overlappende residualer, spesielt om det forekommer like mønstre. Vi søker derfor en løsning som gir oss en helhetsvurdering av residualene slik at vi enkelt kan vurdere om modellen er overdispersert på tvers av alle observasjonene.

DHARMa (Hartig, 2022a) gir en simulerings-basert tilnærming for å lage tolkbare og nedskalerte kvantilresidualer for GLM. Kvantilresidualene standardiseres til verdier mellom 0 og 1, og vil dermed kunne tolkes som residualer fra en lineær regresjon. Pakken tilbyr også flere måter for å teste feilspesifiseringsproblemer som over- eller underdispersjon, null-inflasjon og temporal autokorrelasjon. Kombinert har pakken som formål å gi brukeren et sett med verktøy som pålitelig tester og visualiserer residualer til enklere modellvurdering.

Pakken gir oss tre metoder for å teste overdispersjon. I dette kapittelet skal vi bruke to av dem. Den første testen er kjent som Pearson χ^2 test, og den andre kan beskrives som en ikke-parametrisk dispersjonstest, men omtales her som DHARMa dispersjonstest. En utfordring med DHARMa er at fremgangsmåten for utregninger ikke er tydelig dokumentert, og i dette kapittelet skal vi forsøke å formulere hva DHARMa gjør samtidig som vi referer til kildene som gav grunn til formuleringen.

5.2.1 Parametrisk bootstrap

DHARMa bruker parametrisk bootstrap for å simulere responser. Parametrisk bootstrap lar oss generere tilfeldige variabler fra fordelingen som antas av parametriske modeller. Ved å generere flere tilfeldige utvalg under identiske forhold kan vi reflektere selve prosessen som genererte det observerte utvalget (Rizzo, 2019, s. 153). De genererte utvalgene kan dermed brukes til å avdekke usikkerhet om fordelingen antatt av modellen vår.

Vi lar nå $\boldsymbol{y}_{obs}$ være en vektor med observerte responser,

$$\boldsymbol{y}_{obs} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \tag{5.8}$$

Vi antar at $\boldsymbol{y}_{obs}$ er generert fra fordelingen til $\boldsymbol{y}$. Det vil si at $\boldsymbol{y} \sim f(\boldsymbol{\Theta})$ hvor

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1^\intercal \\ \vdots \\ \boldsymbol{\theta}_N^\intercal \end{bmatrix}. \tag{5.9}$$

En parametrisk modell, som for eksempel GLM, antar at f er kjent og estimerer parametere $\boldsymbol{\theta}_i$ for $(1 \leq i \leq N)$ som inngår i $\hat{\boldsymbol{\Theta}}$. Parametrisk bootstrap anser y_i som et utfall fra fordelingen til den stokastiske variabelen $Y_i \sim f(\boldsymbol{\theta}_i)$, og generer B replikater $\boldsymbol{y}_i^{(b)}$ $(1 \leq b \leq B)$ fra $\hat{Y}_i \sim f(\hat{\boldsymbol{\theta}}_i)$ for enhver i. De $B \times N$ responsene utgjør B bootstrap utvalg,

$$\boldsymbol{y^{(b)}} = \begin{bmatrix} y_1^{(b)} \\ \vdots \\ y_N^{(b)} \end{bmatrix}. \tag{5.10}$$

Det fins flere fremgangsmåter, men prinsippet som gir grunnlag for parametrisk bootstrap er det samme. Dersom $f(\hat{\boldsymbol{\Theta}})$ er et godt estimat av $f(\boldsymbol{\Theta})$ regner vi med at $\boldsymbol{y}_{obs}$ deler egenskaper med $\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(B)}$. Med andre ord behandler vi bootstrap utvalgene som en populasjon, og undersøker om det er rimelig å anta at utvalget har oppstått tilfeldig fra populasjonen.

Fra hver av modellene presentert i kapittel 4, genererte vi $B = 10\,000$ utvalg, som betyr at vi simulerte $B \times N = 58\,640\,000$ responser fra hver av modellene. Senere i kapittelet skal vi formidle hvordan de simulerte responsene ble brukt til å konstruere kvantilresidualer, og til å teste om antagelsene om dispersjon var tilstrekkelige. For å komme frem til hvordan DHARMa konstruerte simulerte residualer brukte vi kommentarer fra Florian Hartige (utvikleren som står bak DHARMa) som kilde(Hartig, 2022b). For å konkludere om hvordan kvantilresidualene ble regnet ut måtte vi undersøke hvordan residualer ble skalert ved å sammenligne egen kode med github dokumentasjonen (Hartig, 2022c) da det ikke ble gitt tydelige

føringer fra
R-dokumentasjonen (Hartig, 2022a). For å komme frem til hvordan dispersjonstesten ble utført måtte vi rekonstruere testen ved hjelp av github kode (Hartig, 2022d), og vi fant at enkelte føringer angående testen var feil i R-dokumentasjonen.

5.2.2 Kvantilresidualer

DHARMa bruker de simulerte responsene for å skalere residualer mellom 0 og 1. Et kvantilresidual er konstruert ved å konstruere en empirisk kumulativ fordelingsfunksjon (ECDF) $F_{\hat{Y}_i}(y) = P(\hat{Y}_i \leq y)$ av bootstrap replikatene som er generert fra $\hat{Y}_i$ andelen av punkt som havner innenfor intervallet $(-\infty, y]$(Rizzo, 2019, s. 36). For et ordnet utvalg $y_i^{(1)} \leq y_i^{(2)} \leq ... \leq y_i^{(B)}$ med ECDF $F_{\hat{Y}_i}$ er den kumulative sannsynligheten $F_{\hat{Y}_i}(y)$, det vil si

$$F_{\hat{Y}_i}(y) = \begin{cases} 0 & y < y_i^{(b)}, \\ \frac{b}{B} & y_i^{(b)} \leq y < y_i^{(b+1)}, b = 1, ..., B-1, \\ 1 & y_i^{(B)} \leq y. \end{cases} \tag{5.11}$$

Vi lar nå y_i være en observert respons og $y_i^{(1)} \leq y_i^{(2)} \leq ... \leq y_i^{(b-1)} \leq y_i^{(b)} \leq ... \leq y_i^{(B)}$ være et ordnet utvalg av bootstrap replikater generert fra $f(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$. Dersom $y_i^{(b-1)} < y_i \leq y_i^{(b)}$, så følger det at det korresponderende kvantilresidualet for observasjon i er

$$r_i^{(q)} = U, \tag{5.12}$$

hvor

$$U \sim \text{Uniform}\left(F_{\hat{Y}_i}\left(y_i^{(b-1)}\right), F_{\hat{Y}_i}\left(y_i^{(b)}\right)\right). \tag{5.13}$$

Når vi undersøkte hvordan residualene var regnet ut var det ikke mulig å gjenskape de samme residualene som blir konstruert av DHARMa da U er en tilfeldig variabel. Ved konstruksjon av den empiriske kumulative fordelingsfunksjonen oppdaget vi at resultatet $F_{\hat{Y}_i}\left(y_i^{(b-1)}\right) < r_i^{(q)} \leq F_{\hat{Y}_i}\left(y_i^{(b)}\right)$ samsvarte med github dokumentasjonen.

Ifølge Dunn (2018, s. 306) viser Pearson- og kvantilresidualer en eksakt normalfordeling når responsene følger en normalfordeling. Det er verdt å merke seg at Pearson-residualer ikke alltid følger en normalfordeling, spesielt for diskrete eksponentielle familier. Som et alternativ anbefaler han bruk av kvantilresidualer for diskrete responsfamilier, da dette kan bidra til å unngå forstyrrende mønstre.

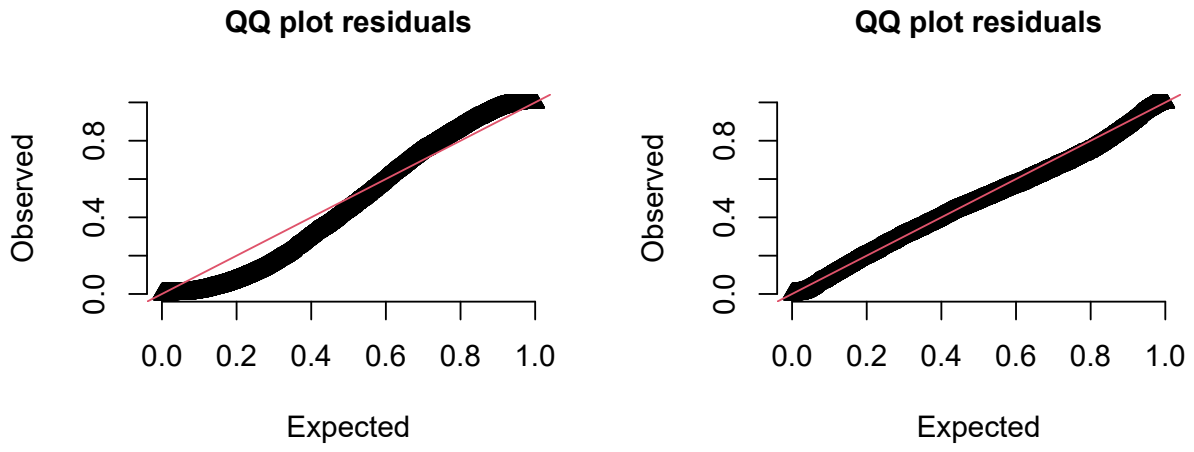
Hvert kvantilresidual er basert på hvor verdiene fra det observerte utvalget ligger i forhold til medianen blant de tilsvarende bootstrap-residualene. En verdi av $r_i^{(q)}$ mellom 0.5 og 1 indikerer at den observerte verdien er større enn medianverdien av de simulerte verdiene fra fordelingen antatt av modellen. På den annen side er medianverdien større når $r_i^{(q)}$ ligger mellom 0 og 0.5. Når $r_i^{(q)} = 1$ eller $r_i^{(q)} = 0$, antyder residualet at y_i er en avsidesliggende observasjon.

5.3 QQ-plot

Et QQ-plot sammenligner kvantilresidualene mot teoretiske kvantiler. DHARMa sammenligner det ordnede utvalget av kvantilresidualer med teoretiske kvantiler fra en standard uniform fordeling (Hartig, 2022a, s. 26). De teoretiske kvantilene representerer den kumulative fordelingsfunksjonen for den forventede fordelingen av residualene. Ifølge Hartig (2022e) kan det uniforme QQ-plottet brukes til å oppdage alle avvik fra den forventede fordelingen, noe som igjen kan antyde overdispersjon.

5.6a viser kvantilresidualene for M_0. Den røde linjen representerer de teoretiske kvantilene for en standard uniform fordeling, mens kvantilresidualene er oppsamlingen av svarte punkter som former en tykk graf. Ifølge Hartig (2022e) antyder S-formen at modellen er overdispersert da den tydelig avviker sterkt fra forventet fordeling i øvre og nedre kvantil. I figur 5.6 og 5.8 er det tydelig at punktene legger seg om den røde linjen i større grad enn det vi ser i 5.6a.

Analysen av kvantilresidualene indikerer tydelig at modell M_0 er overdispersert, og det er merkbart at hver av de negativt binomiale modellene konvergerer nærmere den forventede fordelingen. Imidlertid kan det oppleves som utfordrende å skille mellom residualplotene. I situasjoner hvor vi skal avgjøre hvilke av modell som er minst overdispersert blant et utvalg nestede modellen, kan det være vanskelig å være fullstendig sikker i egen vurdering basert på residualplotene alene.



(a) M_0 **(b)** M_1

Figur 5.6: QQ-plot for M_0 og M_1

5.4 Testing av dispersjon

For å avgjøre hvor dispersert de ulike modellen er skal vi teste for dispersjon ved å sammenligne residualvarians blant det observerte og de simulerte utvalgene, og ved å beregne summen av pearson residualene fra de observerte residualene. Pearson-χ^2-testen er godt kjent fra krysstabeller (Devore, 2012, s. 744-751), mens den ikke-parametriske dispersjonstesten er især for DHARMa (Hartig, 2022a, s. 43).

5.4.1 Pearson-χ^2 test

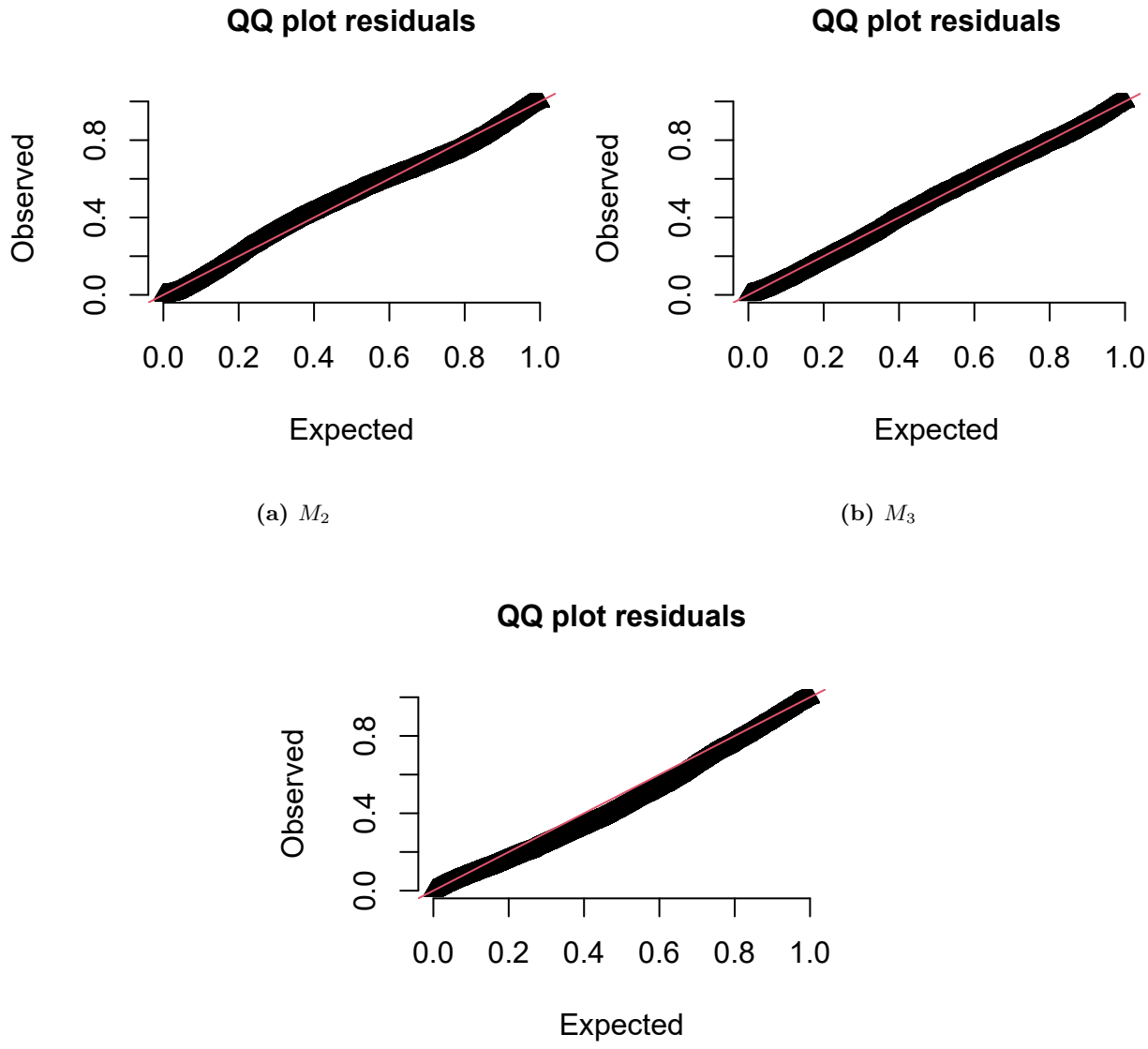
Vi har tidligere beslrevet hvordan vi, i henhold til Hardin (2007, s. 167), kan bruke Pearson χ^2 statistikken i (3.25), her referert til som Q, for å vurdere nivået av dispersjon.

I følge Hardin (2007, s. 165) indikerer en verdi av $Q = 2$ at modellen må justeres, mens alle verdier av $Q > 1$ tyder på overdispersjon. Hilbe (2014, s. 37) påpeker at enhver verdi av $Q > 1$ indikerer overdispersjon i en Poisson-modell uansett verdien, og argumenterer for at det kan være mer hensiktsmessig å bruke en negativ binomialfordeling. Dette skyldes det ekstra leddet i variansen, representert ved $\frac{\lambda^2}{\tau}$ (Hilbe, 2014, s.). Med andre ord gir negativ binomiale modeller en tilnærming til å modellere data der variansen overstiger forventningen eller når observert varians overstiger forventningen (Hilbe, 2014, s. 133). Ideelt sett vil en godt tilpasset negativ binomial modell resultere i $Q = 1$.

	Q	χ^2	Frihetsgrader	P
M_0	3.76	21407	5700	0
M_1	1.35	7708	5699	0
M_2	1.33	7550	5698	0
M_3	1.17	6682	5698	0
M_4	1.09	6205	5697	0
M_5	1.32	7722	5833	0
M_6	1.05	6104	5815	0

Tabell 5.2: Pearson-χ^2 test

I tabell 5.2 observerer vi at justering av dispersjonsparameteren betydelig reduserer graden av overdispersjon. Vurdering resultatene for Q i de ulike modellene legger viser at inkluderingen av en parameter for alderstrinn og kalenderår fører til en betydelig reduksjon av Q fra 1.35 for M_1 til 1.09 for M_4. En nedgang på 0.26 i Q antyder at introduksjonen av en dispersjonsmodell bidrar vesentlig til å minimere overdispersjonen. Mellom modellene M_5 og M_6 var differansen 0.28, men denne forskjellen kom på bekostning av tolkningen av regresjonsparameterne.



(a) M_2

(b) M_3

(c) M_4

Figur 5.7: QQ-plot for kategoriske modeller.

5.4.2 DHARMa dispersjonstest

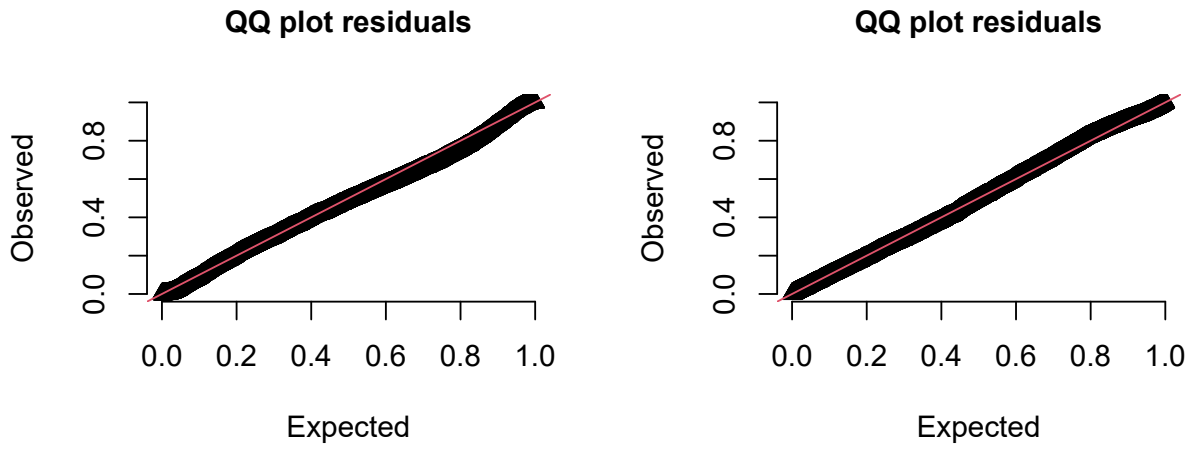
DHARMa muliggjør også en enkel test for dispersjon ved hjelp av en ikke-parametrisk tilnærming som sammenligner variasjonen i observerte med simulerte responsresidualer. Dokumentasjonen for denne testen kan oppfattes som utydelig i prosedyren, og den teoretiske begrunnelsen bak er ikke klart presentert.

For å formulere resultatene fra den ikke-parametriske dispersjonstesten ble det forsøkt flere tilnærminger for å beregne testresultatene. Vi observerte at DHARMa brukte observerte og simulerte responsresidualer for å utføre testen. Resultatene fra dispersjonstesten inkluderer en rapport om dispersjonstesten og en tilhørende P-verdi som er vises i tabell 5.3.

For å kunne formulere funnene tydelig lar vi responsresidualene for det observerte og simulerte utvalget henholdsvis være

$$\boldsymbol{r} = \boldsymbol{y}_{obs} - \hat{\boldsymbol{\lambda}} \quad \text{og} \quad \boldsymbol{r}^{(\boldsymbol{b})} = \boldsymbol{y}^{(\boldsymbol{b})} - \hat{\boldsymbol{\lambda}}, \tag{5.14}$$

hvor $\hat{\boldsymbol{\lambda}}$ er en vektor med prediksjoner fra modellen som residualene er generert fra, mens $\boldsymbol{y}_{obs}$ og $\boldsymbol{y}^{(\boldsymbol{b})}$ er henholdsvis formulert av (5.8) og (5.10).



(a) M_5 **(b)** M_6

Figur 5.8: QQ-plot for polynomiske modeller.

DHARMa beregner utvalgsvariansen s_{obs} fra de observerte residualene som

$$s_{obs} = \sqrt{\sum_{i=1}^{N} \frac{(r_i - \bar{r})^2}{N-1}}, \quad \text{hvor} \quad \bar{r} = \sum_{i=1}^{N} \frac{r_i}{N}, \tag{5.15}$$

og fra de simulerte residualene

$$s_{sim} = \sqrt{\sum_{i=1}^{N}\sum_{b=1}^{B} \frac{(r_i^{(b)} - \bar{r}_{sim})^2}{BN-1}}, \quad \text{hvor} \quad \bar{r}_{sim} = \sum_{i=1}^{N}\sum_{b=1}^{B} \frac{r_i^{(b)}}{BN}. \tag{5.16}$$

Dispersjonstesten rapporterer kvotienten av observert over simulert residualvarians, her benevnt som F_{test}, i lag med en P-verdi (Hartig, 2022d). Dermed kunne vi verifisere at

$$F_{test} = \frac{s_{obs}^2}{s_{sim}^2}, \tag{5.17}$$

med 5 desimalers sikkerhet.

P-verdien for testen ble beregnet ved å sammenligne s_{obs}^2 med de empiriske variansene ${s_{sim}}^{(b)}$ med hvert av de simulerte utvalgene. For å finne ut nøyaktig hvordan beregnet vi ${s_{sim}}^{(b)}$ fra

$${s_{sim}}^{(b)} = \sqrt{\sum_{i=1}^{N} \frac{(r_i^{(b)} - \bar{r}^{(b)})^2}{N-1}}, \quad \text{hvor} \quad \bar{r}^{(b)} = \sum_{i=1}^{N} \frac{r_i^{(b)}}{N}. \tag{5.18}$$

For hver b regnet vi ut forholdet mellom det simulerte utvalget over standardavviket til de simulerte responsene,

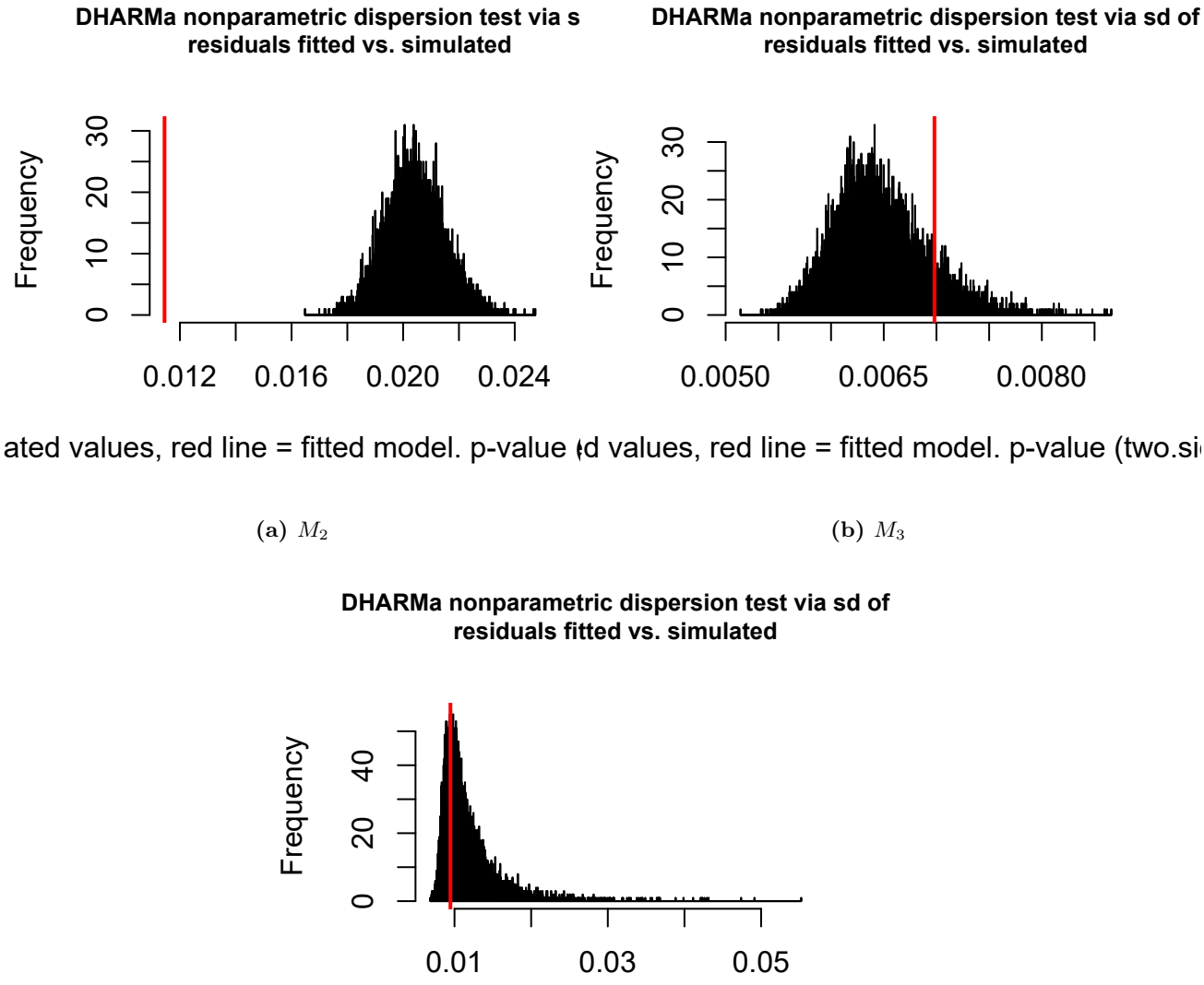
$$F_{sim} = \frac{{s_{sim}^2}^{(b)}}{\sigma_{sim}^2}, \tag{5.19}$$

hvor

$$\sigma_{sim} = \sqrt{\sum_{i=1}^{N}\sum_{b=1}^{B} \frac{(y_i^{(b)} - \bar{y}^{(b)})^2}{BN-1}}, \quad \text{med} \quad \bar{r}_{sim} = \sum_{i=1}^{N}\sum_{b=1}^{B} \frac{y_i^{(b)}}{BN}. \tag{5.20}$$

Ved å la

$$F_{obs} = \frac{s_{obs}^2}{\sigma_{sim}^2}. \tag{5.21}$$



(a) M_2 **(b)** M_3

(c) M_4

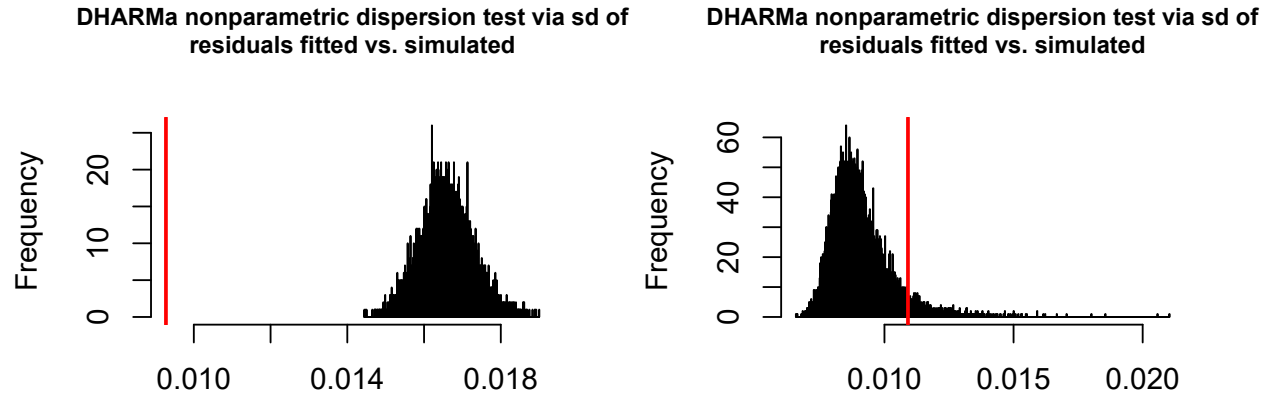
Figur 5.9: DHARMa dispersjonstest for kategoriske modeller.

være den korresponderende mengden til F_{sim} for det observerte utvalget kunne vi avdekke at P-verdien for testen var beregnet som andelen av utvalg som oppfylte ett av følgende kriterier:

$$P = \begin{cases} P(F_{obs} \geq F_{sim}) \\ P(F_{obs} \leq F_{sim}) \\ 2P(F_{obs} \leq F_{sim}) \end{cases} . \tag{5.22}$$

Nøyaktig hvilke betingelser P-verdien ble beregnet i henhold til har vi ikke formulert her. Hvilke av alternativene som ble valgt opplevdes uklart med tanke på at vi verken spesifiserte en ensidig eller tosidig test.

Ved førsteinntrykket syntes dispersjonstesten å ligne en vanlig F-test, som beskrevet av Devore (2012, s. 175). Formålet med en F-test er å sammenligne variansen mellom to utvalg, enten med like eller ulike størrelser. DHARMas dispersjonstest bygger på dette konseptet, men i stedet for å direkte sammenligne variansene mellom observerte og simulerte responser, tester vi om det er sannsynlig at residualvariansen til det observerte utvalget kan stamme fra populasjonen av residualvarianser generert fra fordelingen antatt av den



(a) M_5 **(b)** M_6

Figur 5.10: DHARMa dispersjonstest for polynomiske modeller.

parametriske modellen. Dette blir tydelig da P-verdien fra testen gir informasjon om hvor residualvariansen i det observerte utvalget plasserer seg i forhold til de simulerte utvalgene.

I 5.10 og 5.9 ser vi en sammenligning av dispersjonen blant observerte responser og simulerte responser fra de nestede modellene. For M_2, M_3, og M_4 anvendes indikatorvariabler som prediktorvariabler, mens M_5 og M_6 bruker ortogonale polynomer som prediktorvariabler. Histogrammene illustrerer dermed avstanden mellom den underliggende populasjonen av residualvariansen, beskrevet av de simulerte bootstrap-utvalgene basert på modellen, og den observerte residualvariansen. De sorte søylene i histogrammene representerer frekvensen av observerte verdier av F_{sim}. Den røde linjen indikerer hvor F_{obs} befinner seg. Histogrammet viser hvor observert residualvarians befinner seg i forhold til den simulerte populasjonen av residualsvarianser. Når F_{obs} befinner seg utenfor populasjonen, indikerer det liten eller ingen sannsynlighet for å observere s^2_{obs} fra fordelingen antatt av modellen vi har simulert responser fra. Dette tolker vi som at det befinner seg et ekstra lag med variasjon i dataene som ikke er beskrevet av $f(\hat{\boldsymbol{\Theta}})$.

5.9 indikerer at M_1 viser en betydelig avvikelse i testen, da den røde linjen som representerer den observerte residualvariansen, er markant høyere enn det som genereres fra populasjonen. Denne trenden gjentar seg også i (5.10) for for M_5. Imidlertid observerer vi videre at både M_4 og M_6 viser en klar forbedring gjennom dispersjonsmodellering, da den observerte residualvariansen ligger innenfor rammene til den simulerte residualvariansen. Vi kan dermed anta at modellene som benytter dispersjonsmodellering har estimert parametere som hører til fordelinger med større sannsynlighet for å være den underliggende fordelingen til dataene våre.

	F_{test}	P
M_0	4.31	0.00
M_1	0.56	0.00
M_2	0.56	0.00
M_3	1.08	0.23
M_4	0.83	0.57
M_5	0.56	0.00
M_6	1.21	0.12

Tabell 5.3: DHARMa dispersjonstest

Etter avrunding til 2 desimaler fant vi at resultatene for M_1, M_2, og M_5, vist i tabell 5.3, var like. Fra forholdet beskrevet av (5.17), tolker vi $F_{test} = 0.56$ som om $\boldsymbol{r}$ har lavere varians sammenlignet med $\boldsymbol{r^{(1)}}, ..., \boldsymbol{r^{(B)}}$. Fra (5.22) ser vi at en mulig oppfatning av $P = 0$ er at det ikke er sannsynlig å observere residualvariansen fra fordelingen som er antatt av modellen vår. Derimot indikerer resultatene fra M_3, M_4

og M_6 at residualvarians er bedre forklart av modellene som følge av at utfallet for F_{test} antyder at observert residualvarians er mer forenlig med simulert residualvarians. $P > 0$ indikerer at det er sannsynlig å observere residualvariansen indikert av modellen vår.

Om vi tilfører samme tolkning av F_{test} som for Q, er det mulig å konkludere med at homogen dispersjonsparameter gir oss underdisperserte modeller. Mens kalenderår på egenhånd ikke viser tegn til utbedring av underdispersjon, fant vi at alderstrinn alene, eller i kombinasjon med kalenderår, resulterte i modeller som forklarte variabiliteten i data i større grad. Parametrisk bootstrap har gitt oss mulighet til å generere et stort antall responser som har gitt oss et omfattende innblikk i variabiliteten av populasjonen som blir anslått av modellene våre. Undersøkelsen vår antyder at DHARMa dispersjonstest kan tilføre evidens som dokumenterer effektiviteten av simultan modellering med glmmTMB.

Kapittel 6

Overlevelsesanalyse

6.1 Overlevelsesanalyse

La X være en overlevelsestid og la $f(x)$ være sannsynlighets tetthetsfunksjonen. Sannsynligheten for at en hendelse inntreffer før en spesifikk tid x er gitt av den kumulative fordelingsfunksjonen

$$F(x) = \mathbb{P}(X < x) = \int_0^x f(s)\, ds. \tag{6.1}$$

Overlevelsesfunksjon er sannsynligheten for å være i live etter tid x, og er gitt av

$$S(x) = \mathbb{P}(X \geq x) = 1 - F(x). \tag{6.2}$$

Hasard funksjonen er sannsynligheten for å dø mellom tid x og $x + \delta x$, gitt at subjektet fortsatt er i live ved tid x,

$$\begin{aligned} h(x) &= \lim_{\delta x \to 0} \frac{\mathbb{P}(x \leq X < (x + \delta x) | X > x)}{\delta x} \\ &= \lim_{\delta x \to 0} \frac{F(x + \delta x) - F(x)}{\delta x} \left(\frac{1}{S(x)} \right) \\ &= \frac{f(x)}{S(x)} \\ &= -\frac{d}{dx} \log S(x) \end{aligned} \tag{6.3}$$

Ved å se nærmere på forholdet mellom $h(x)$ og $S(x)$ finner vi fra (6.3) at

$$S(x) = \exp[-H(x)] \tag{6.4}$$

og

$$H(x) = \int_0^x h(x)\, du, \tag{6.5}$$

eller

$$H(x) = -\log S(x). \tag{6.6}$$

$H(x)$ omtalest som den kumulative hasardfunksjonen (Dobson, 2018, s. 25-26).

6.2 Kaplan-Meier

Kaplan-Meier overlevelsesanalyse gir en ikke-parametrisk metode for å oppsummere overlevelsessannsynligheter (May, 2009). Tilnærmingen gir en grafisk representasjon av fremtidig dødelighet.

I neste seksjon skal vi greie ut om hvordan vi kan estimere livtabeller fra de svenske populasjons dataene. Dermed vil vi bruke tabellen korresponderende til 2005 til å konstruere Kaplan-Meier kurver. Kurvene tar utgangspunkt i at vi først estimerer livtabellen, deretter kan vi estimere hvor sannsynlig det er at en person ved alder x er i live om n år gjennom den kumulative hasardfunksjonen i (6.5). Livtabellen gir oss en enkel metode for å konstruere overlevelseskurver som tar utgangspunkt i den kumulative hasardfunksjonen mellom alder x og $x + s$ for enhver $s \leq n$.

Det som vil være interessant for oss er å undersøke hvilke konsekvenser vi finner av å estimere overlevelseskurver fra modellene våre. En sammenligning av Kaplan-Meier estimater fra prediksjoner og observasjoner vil gi innsikt i konsekvensene når vi anvender estimatene i forsikringssammenheng.

6.3 Livtabell

En livtabell, ofte kalt dødelighetstabell, gjør det enkel å analysere dødelighet for en vilkårlig populasjon basert på historiske data. Tabellen bruker et konstruert årskull som består av 100 000 personer, som deretter blir eksponert for dødeligheten i den gitte perioden som tabellen representerer. Det mest interessante aspektet ved dette er å observere hvordan det fiktive årskullet gradvis taper medlemmer. Dette oppnås ved å registrere antall individer som er i live og antall som har gått bort ved hvert alderstrinn. Enda mer fascinerende er muligheten til å beregne forventet levealder for dette årskullet.

Det er viktig å være klar over at en dødelighetstabell representerer en hypotetisk situasjon. I virkeligheten vil ingen årskull oppleve nøyaktig den samme dødeligheten som vises i tabellen. For å finne den faktiske forventede levealderen for et årskull, må man benytte seg av data fra faktiske kohorter. Imidlertid krever dette at man venter i over hundre år til alle individene i årskullet er borte.

En dødelighetstabell tar utgangspunkt i dødssannsynligheten q_x som er sannsynligheten for at en person ved alder x dør før alder $x+1$ (Foss, 1995). Prosessen starter med opprettelsen av en hypotetisk kohort l, som ved alder 0 består av l_0 personer. I denne oppgaven konstruerer vi hypotetiske kohorter med $l_0 = 100\ 000$ personer. Denne kohorten blir deretter utsatt for dødeligheten knyttet til ulike alderstrinn i henhold til den perioden tabellen er beregnet for.

Dette kan også uttrykkes matematisk som

$$l_x = l_{x-1}(1 - q_{x-1}),$$

hvor $(1 - q_{x-1})$ er overlevelsessannsynligheten ved alder $x - 1$.

Antall personer som dør i alderen mellom x og $x + 1$ år i den hypotetiske kohorten er

$$d_x = l_x q_x.$$

For å beregne forventet gjenstående levetid for en person ved alder x antar vi at dødfallene fordeler seg jevnt over året slik at gjennomsnittet samtidig levende ved alder x er

$$L_x = \frac{l_x + l_{x+1}}{2}. \tag{6.7}$$

Videre bruker vi summen av gjenstående leveår i alderstrinn x,

$$T_x = \sum_{i=x}^{\infty} L_x, \tag{6.8}$$

til å beregne forventet gjenstående levetid ved alderstrinn x som

$$e_x = \frac{T_x}{l_x}. \tag{6.9}$$

Fremgangsmåten for beregninger av livtabeller er generelt ganske lik, men det fins flere beregningsmetoder for q_x.

6.3.1 Dødsintensitet, dødssannsynlighet og overlevelsessannsynlighet

I vårt tilfelle skal vi predikere dødsintensiteten μ for hvert alderstrinn i hvert kalenderår. Vi antar at dødsintensiteten er konstant gjennom året og ulik mellom alderstrinn. Vi antar at populasjonen n for hvert alderstrinn dør med lik rate μ og at λ betegner korresponderende totalt antall døde ved slutten av året. Antagelsene gjør at vi kan estimere μ fra λ slik at

$$\mu = \frac{\lambda}{n} = \exp[\boldsymbol{z}^\intercal \boldsymbol{\beta}], \tag{6.10}$$

og på samme måte er den ett-årige dødintensiteten ved alderstrinn x ved kalenderår t

$$\mu_x = \int_x^{x+1} \mu(s)\, ds = \frac{\lambda_x}{n_x}. \tag{6.11}$$

Vi lar nå $h(x) = \mu(x)$ slik at sannsynligheten for at x åring dør mellom alder x og $x+1$ er

$$q_x = 1 - \exp\left[-\mu_x\right]. \tag{6.12}$$

Vi kan også bruke dødsintensiteten fra (6.11) eller dødssannsynligheten fra (6.12) til å estimere korresponderende overlevelsessannsynligheter som

$$\begin{aligned} p_x &= 1 - q_x \\ &= \exp\left[-\mu_x\right]. \end{aligned} \tag{6.13}$$

For ordensskyld påpeker vi at p_x er den komplementære hendelsen til q_x.

6.3.2 Fremskrivelser

Overlevelseskurver kurver

Når vi har funnet overlevelsessannsynlighetene for et vilkårlig år, kan vi bruke dem i tråd med (6.4) og (6.5) til å estimere Kaplan-Meier kurver for alder x som

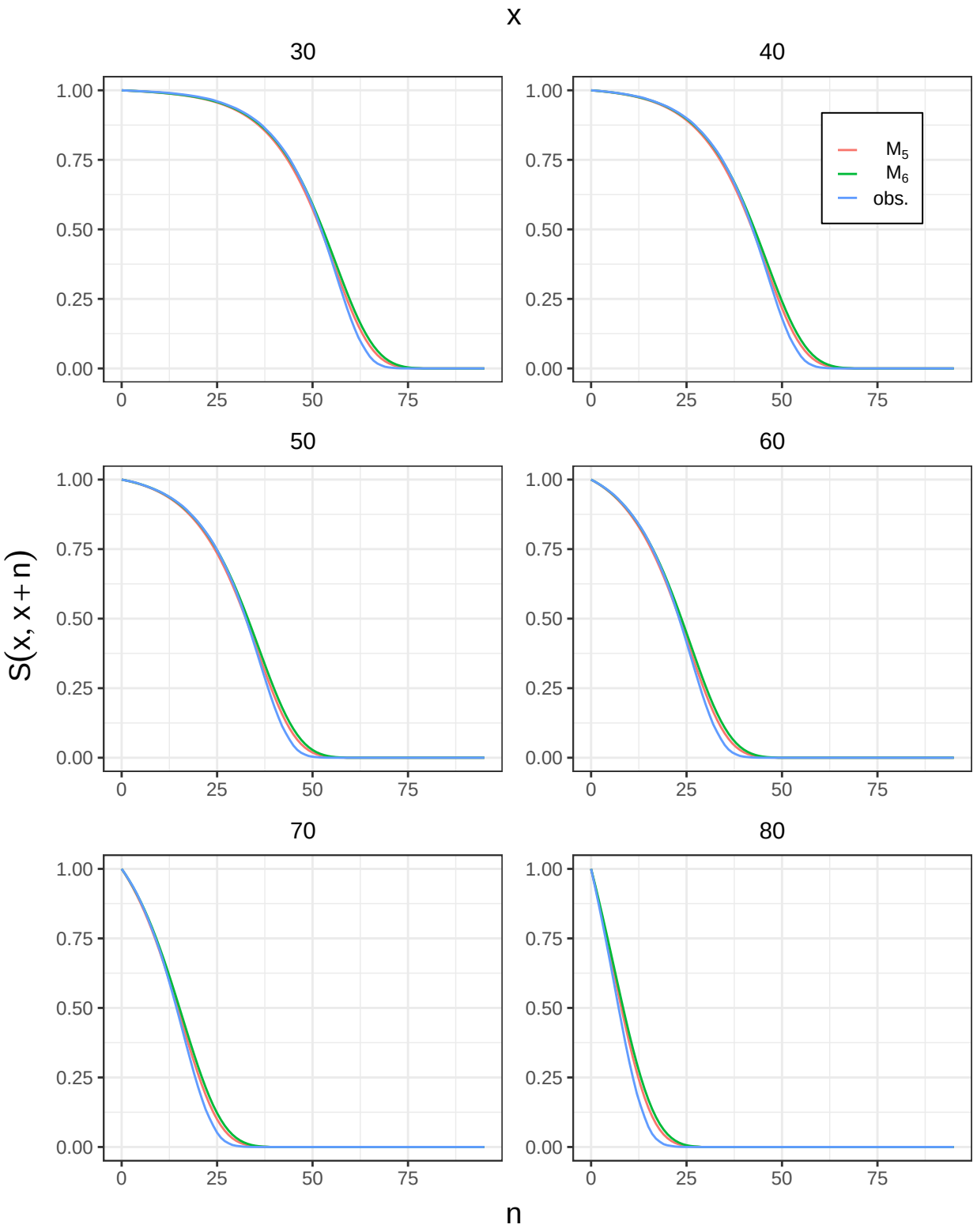
$$\begin{aligned} {}_np_x &= \exp\left[-\int_0^n \mu_x\, ds\right] \\ &= \prod_{s=0}^{n} \exp\left[-\mu_{x+s}\right] \\ &= \prod_{s=0}^{n} p_{x+s} \\ &= S(x, x+n). \end{aligned} \tag{6.14}$$

6.1 viser Kaplan-Meier kurver konstruert i henhold til (6.14) med sannsynligheter basert på predikerte livtabeller fra M_6 og M_5 til sammenligning med tilsvarende tabeller for observasjoner. Kurvene viser hvor sannsynlig det er at en x-åring fortsatt lever innen n år, og er konstruert med data fra 2005. Overlevelseskurvene viser at prediksjonen for M_6 gir oss lengre levetider.

Forventet levealder

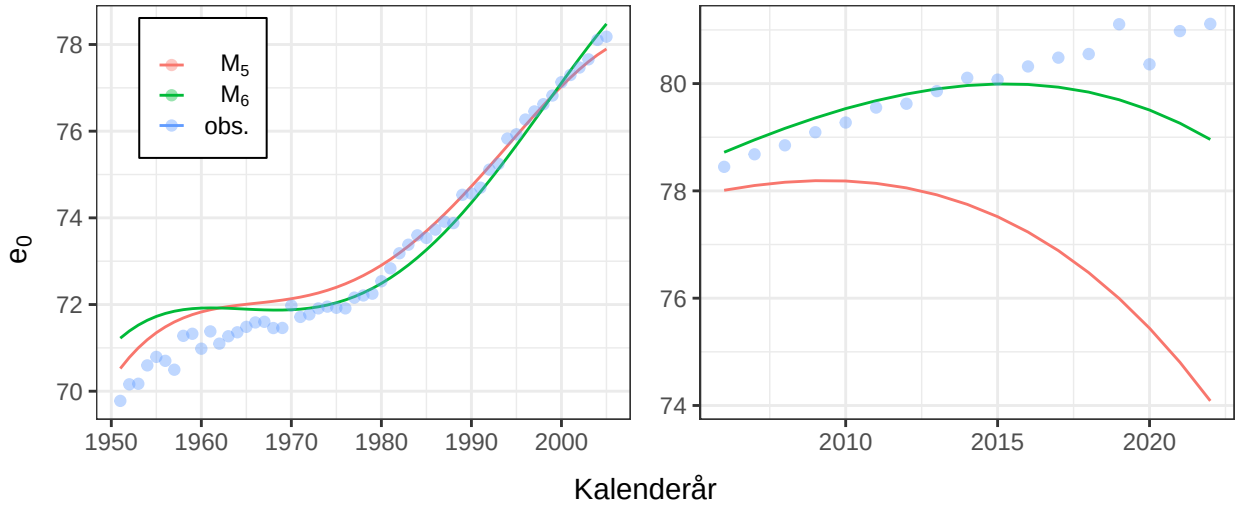
Forventet gjenstående levetid ved fødsel er det gjennomsnittlige antallet år som en nyfødt kan forvente å leve, hvis han eller hun skulle gjennom livet og bli utsatt for de kjønns- og aldersspesifikke dødsratene som er gjeldende på tidspunktet for fødselen, for et bestemt år, i et gitt land, territorium eller geografisk område (WHO). Forventet gjenstående levealder ved fødsel omtales ofte som forventet levealder, og ettersom e_x er tilnærmet lineært avtagende med alder gjenspeiler e_0 det generelle dødelighets nivået i en befolkning da den oppsummerer dødelighets mønsteret som gjelder på tvers av alle alderstrinn.

For å få et innblikk i hvordan forventet levealder vil utvikle seg gjennom perioden 1951-2005 estimerer vi livtabeller fra denne perioden, og regner ut forventet dødelighet som gitt av (6.9). For å undersøke om livtabellene som er basert på prediksjoner fra modellene våre egner seg til estimering av forventet levealder sammenligner vi prediksjoner for 2006 til 2022 med observerte verdier fra denne perioden. 6.2 viser at forventet levetid som er beregnet fra prediksjoner fra M_6 treffer mye bedre enn M_5.



Figur 6.1: Kumulative overlevelsessannsynligheter fra livtabeller beregnet ved å bruke prediksjoner fra M_5 og M_6 sammenlignes med livtabeller basert på observerte verdier.

Feilestimering av risiko kan ha store konsekvenser, spesielt rettet mot pensjon hvor man også må være lenge i forkant for å avdekke risiko tilknyttet pensjonsavtaler. Her har vi brukt to tilnærminger for å estimere dødelighet, og vi ser at dispersjonsmodellering bidrar stort til å oppnå en risiko som gir oss mer treffsikre estimater når vi anvender prediksjonene til estimering av livtabeller som vist av forventet dødelighet i 6.2.



Figur 6.2: Forventet levealder.

Kapittel 7

Diskusjon

Vi har gjennomført en grundig undersøkelse av hvordan modellering av dispersjonsparameteren som en log-lineær funksjon av prediktorvariabler gir betydelige fordeler for tilpasning av overdisperserte populasjonsdata fra HMD. Våre funn støtter konklusjonen om at negativ binomialfordelingen er spesielt egnet for å håndtere den variasjonen som Poisson-fordelingen ikke klarer å fange opp. Vi har også nøye vurdert konsekvensene av situasjoner der dataene viser større variasjon enn det modellen vår forventer.

Resultatene indikerer klart at antakelsen om samme funksjonelle form for heterogen dispersjonsparameter, predikert av glmmTMB, har gitt oss en effektiv tilnærming for å forbedre en modell med klare tegn på overdispersjon. Vår tilnærming til dispersjonsmodellering har ikke bare korrigert modellens prediksjonsstyrke, men også forbedret estimeringen av dispersjonen. Dette har resultert i en reduksjon av den uforklarte variasjonen. Ved nøye utvelgelse av det beste settet med prediktorvariabler i dispersjonsmodellen, har vi optimalisert ytelsen ytterligere. Det er også verdt å merke seg at bruken av polynomer i dispersjonsmodellen har vist seg å ha en lignende effekt som i modellen for forventningen. Dette har resultert i bedre prediksjoner, men samtidig kommet med en kostnad i form av tolkningen av regresjonsparameterne.

Selv om det ikke er overraskende at vi oppnår bedre prediksjoner gjennom dispersjonsmodellen, da dette ofte er konsekvensen av å tilføre mer informasjon til modellen, er det viktig å erkjenne at det sjelden eksisterer en universelt overlegen modell. I maskinlæring og statistikk er det viktig å være bevisst på at ulike modeller har sine styrker og svakheter, og at nye teknologier stadig utvider våre muligheter til å modellere komplekse fenomener. Identifiseringen av overtilpasning i modellen vår er utfordrende, spesielt når vi har tilgjengelig data for hele den svenske befolkningen de siste 72 årene. Vi kan argumentere for at det er hensiktsmessig å utnytte all tilgjengelig informasjon gitt konteksten, men samtidig må vi være forsiktige med å generalisere resultatene. Å predikere dødelighet i populasjoner basert på en modell tilpasset svenske data kan potensielt føre til problemer med generalisering, spesielt når vi møter populasjoner som er betydelig forskjellige fra den svenske. Dette understreker betydningen av å avveie informasjonsrikdom mot behovet for generaliserbarhet i våre modelltilpasninger.

Opprinnelig var intensjonen å tilpasse ulike modeller og deretter vurdere tolkbarheten av dispersjonsmodellen, samt den informasjonen den kunne gi oss om overdispersjon. Imidlertid oppstod det utfordringer underveis når det gjaldt identifisering og vurdering av modeller med en innbygd hierarkisk struktur. Det ble også komplisert å bedømme i hvilken grad vår modell adresserte overdispersjonen som Poisson-fordelingen ikke klarte å forklare. Det ble derfor nødvendig å finne metoder for å kvantifisere overdispersjonen. Vi begynte med å utforske Q som et tidligere dokumentert verktøy for å vurdere om en modell er overdispersert, men vi oppdaget at Q alene ikke ga tilstrekkelige kriterier for å bedømme overdispersjon annet enn når $Q > 1$. I vår søken etter alternative verktøy for å avdekke overdispersjon, fremsto DHARMa dispersjonstesten som en lovende løsning for å avgjøre om vi hadde forbedret modellen i tilstrekkelig grad til å kunne stole på at tilstrekkelig mengde variasjon var forklart. Ved å sammenligne testresultatene ser vi at denne simuleringsbaserte metoden gjør det enkelt å vurdere sannsynligheten for å observere samme mengde variasjon i et nytt utvalg fra modellfordelingen.

7.1 Videre arbeid

The Human Mortality Database tilbyr tilgjengelig data fra flere land enn bare Sverige. Dette gir oss muligheten til å undersøke om dispersjonsmodellering har en lignende effekt blant ulike populasjoner. Videre gir tilgjengeligheten av data fra andre land også rom for formulering av alternative modeller, da ulike kurver kan passe bedre til forskjellige populasjoner. En mulig tilnærming er å kombinere modellering av forventning med kategoriske variabler, samtidig som dispersjonsparametrene blir predikert ved bruk av polynomregresjon. Det er ingen spesiell grunn til at polynomregresjon ikke skulle ha samme effekt når forventning er predikert av kategoriske variabler. I vår studie har vi kun gjort prediksjoner på test data fra de polynomiske modellene, da det er enkelt å gjøre sterke prediksjoner fra generaliserte lineære modeller med kontinuerlige variabler. Imidlertid er det også muligheter for å finne metoder som gjør det mulig å bruke regresjonsparameterne for å gjøre prediksjoner fra de kategoriske modellene.

Generelt sett tilbyr DHARMa en rekke verktøy for modellvurdering, selv om vårt hovedfokus primært har vært på verktøy som tar sikte på å håndtere problemstillinger knyttet til sann overdispersjon. Dermed står det flere verktøy til rådighet i programvaren som kan benyttes for å utføre modellvurderinger i tilfelle av nestede eller hierarkiske modeller. I denne studien har vi konsentrert oss om enkle modeller, men det ville også vært interessant å undersøke om dispersjonstestene kunne bidra til å fastslå om introduksjonen av interaksjoner og tilfeldige effekter kunne forklare variasjonen i dataene.

Videre gir glmmTMB også muligheten til å predikere dispersjonsparametere for kontinuerlige fordelinger tilhørende den eksponentielle familien, inkludert Conway-Maxwell-Poisson-fordelingen. Denne fordelingen er velegnet for å modellere diskrete data og inneholder en parameter som tillater håndtering av både over- og underdispersjon. Et videre arbeid kan være å undersøke om Conway-Maxwell-Poisson fordelingen kan utbedre overdispersjon i anledning av populasjonsdata.

Bibliografi

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2): 378–400, 2017. doi: 10.32614/RJ-2017-066.

Roger L Casella, George og Berger. *Statistical inference*. Cengage Learning, second edition edition, 2002.

Gillian Z De Jong, Piet og Heller. *Generalized linear models for insurance data*. Cambridge University Press, 2008.

Kenneth N mfl. Devore, Jay L og Berk. *Modern mathematical statistics with applications*, volume 285. Springer, second edition edition, 2012.

Adrian G Dobson, Annette J og Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 4th ed. edition, 2018.

Gordon K mfl. Dunn, Peter K og Smyth. *Generalized linear models with examples in R*, volume 53. Springer, 2018.

Benedikte Evensen. glmmtmb for telledata, 2018.

Aslaug Hurlen Foss. Definisjhoner og beregningsmetoder for dødelighetstabell. *SSB Notater*, 1995.

Joseph M Hardin, James W og Hilbe. *Generalized linear models and extensions*. Stata press, 2007.

Florian Hartig. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2022a. URL https://CRAN.R-project.org/package=DHARMa. Hentet 03.11.2023 (DHARMa).

Florian Hartig. Cross validated: simulated residuals, how are they created?, 2022b. URL https://stats.stackexchange.com/questions/570727/simulated-dharma-residuals-how-are-they-created. Hentet 03.11.2023 (DHARMa).

Florian Hartig. github dokumentasjon: help.r, 2022c. URL https://github.com/florianhartig/DHARMa/blob/master/DHARMa/R/helper.R. Hentet 03.11.2023 (DHARMa).

Florian Hartig. github dokumentasjon: testdispersion, 2022d. URL https://github.com/florianhartig/DHARMa/blob/master/DHARMa/R/tests.R. Hentet 04.11.2023 (DHARMa).

Florian Hartig. Vignette: Residual diagnostics for hierarchical (multi-level / mixed) regression models, 2022e. URL https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html. Hentet 03.11.2023 (DHARMa).

Joseph M. Hilbe. *Modeling Count Data*. Cambridge University Press, 2014.

HMD. The human mortality database, 2022. URL https://www.mortality.org/Data/DataAvailability. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Hentet 20.07.2023.

Gareth mfl. James. *An introduction to statistical learning*, volume 112. Springer, 2013.

Rob mfl. Kaas. *Modern actuarial risk theory: using R*, volume 128. Springer Science & Business Media, 2008.

Håvard Kolve. Age-estimation of harp seals in r, 2022.

Warren L. May. *Kaplan-Meier Survival Analysis*, pages 1590–1593. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-540-47648-1. doi: 10.1007/978-3-540-47648-1_3196. URL https://doi.org/10.1007/978-3-540-47648-1_3196.

Nelder og Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344614.

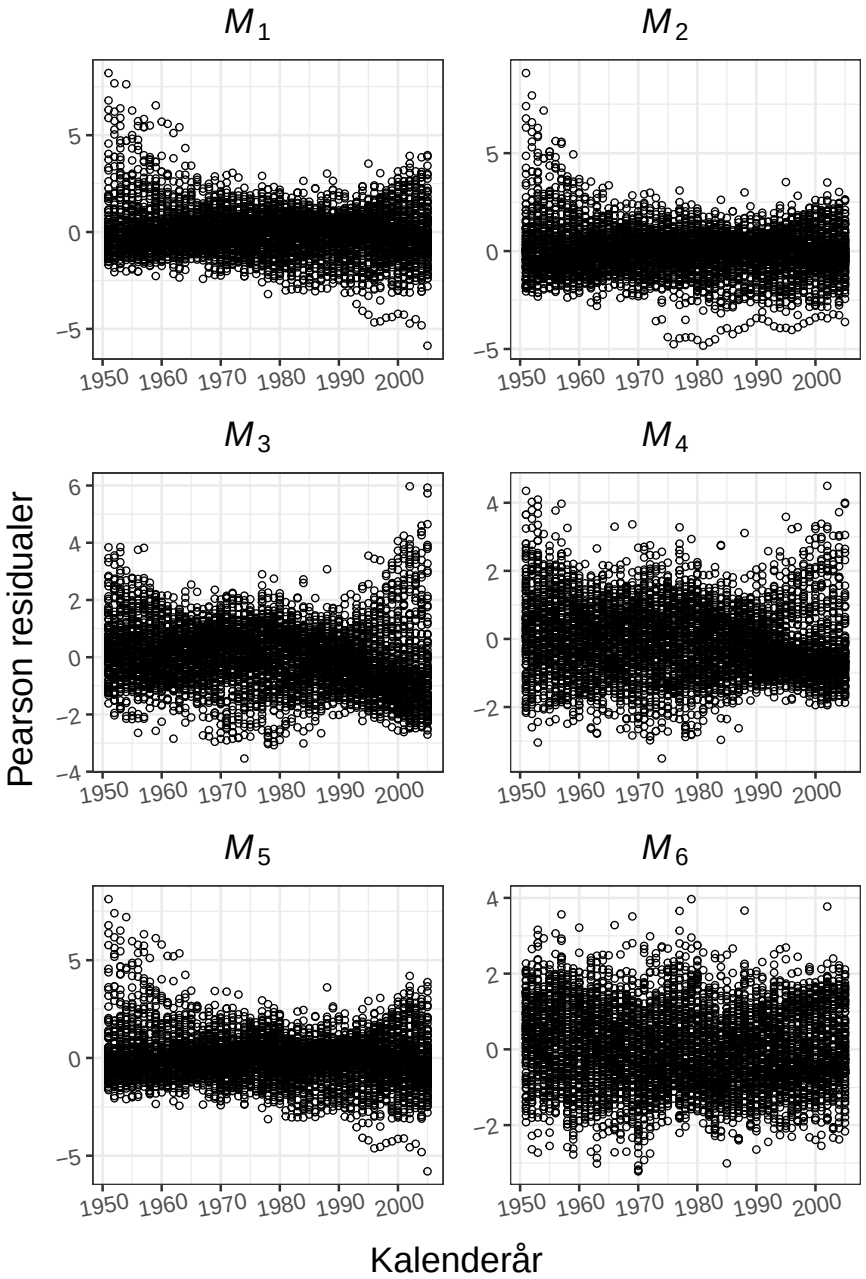
Maria L Rizzo. *Statistical computing with R*. CRC Press, 2019.

WHO. World health organization: Life expectancy at birth (years). URL https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years). Hentet 18.11.2023.

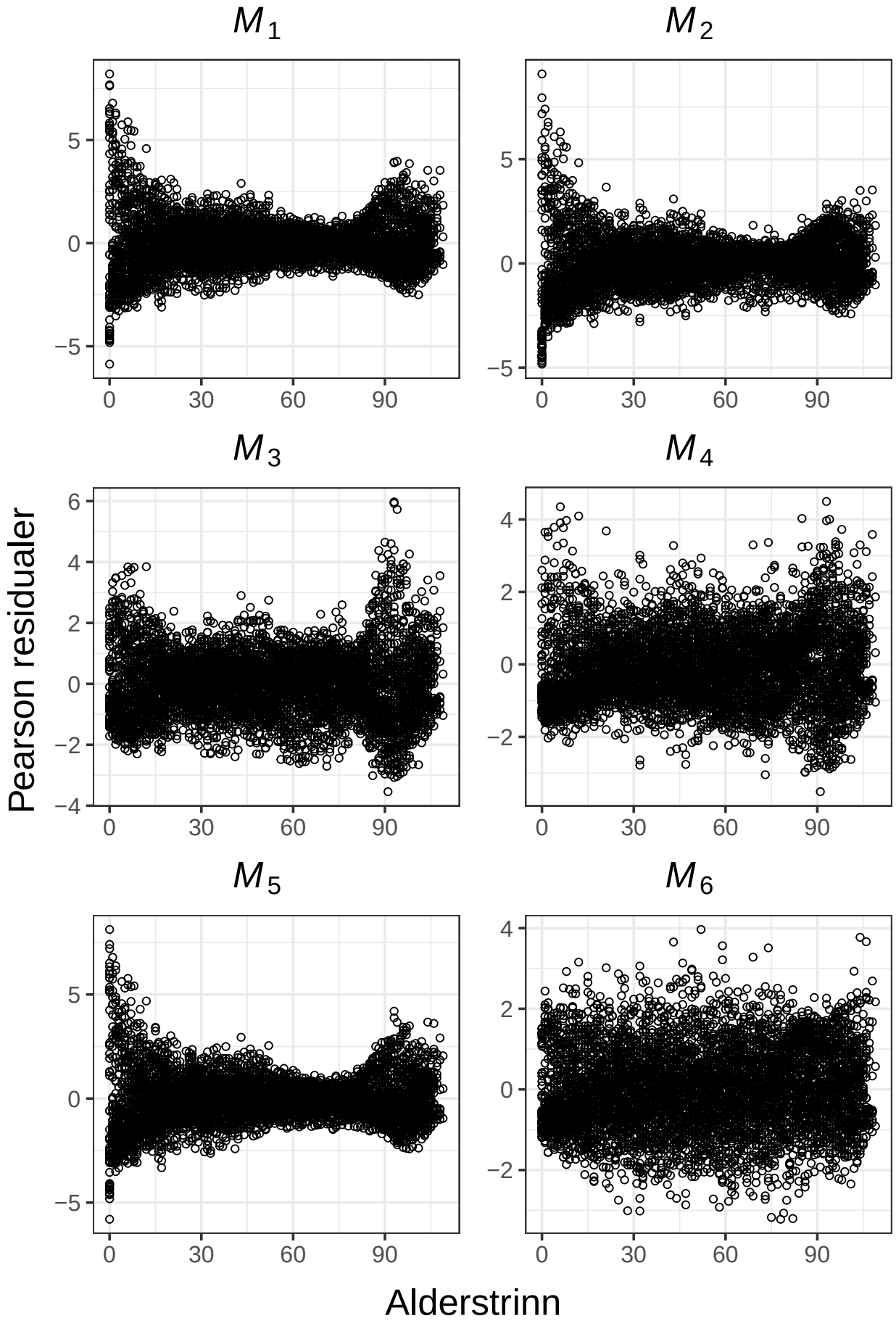
Tillegg A

Figurer

A.1 Pearson residualer

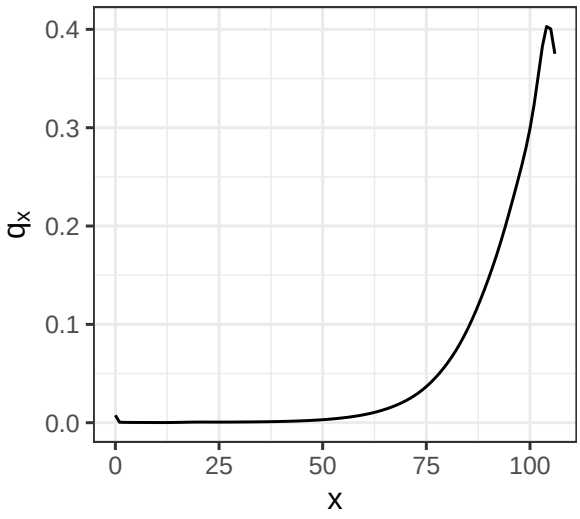


Figur A.1: Pearson residualer som funksjon av kalenderår.

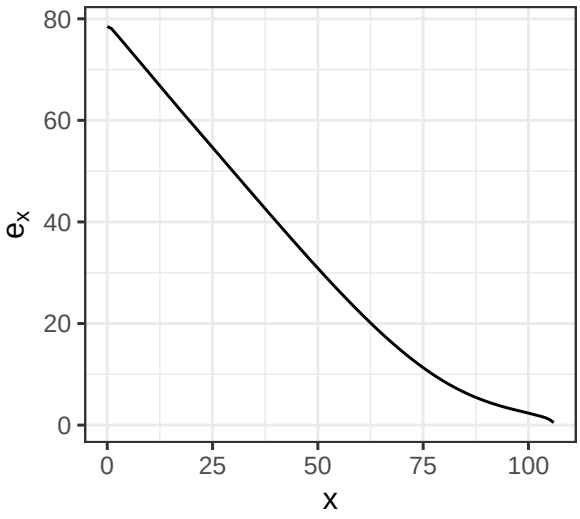


Figur A.2: Pearson residualer som funksjon av alderstrinn.

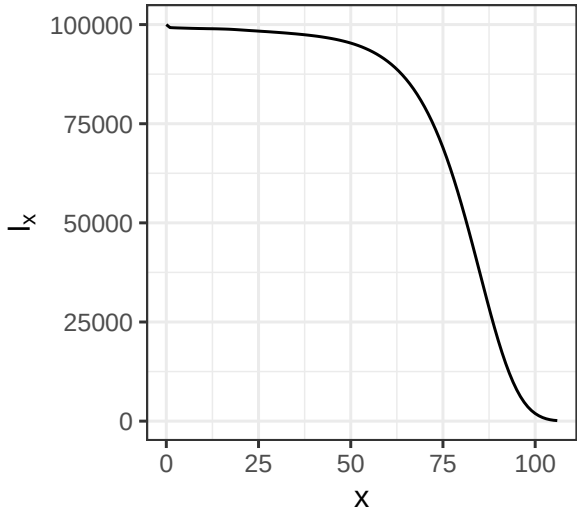
A.2 Livtabell



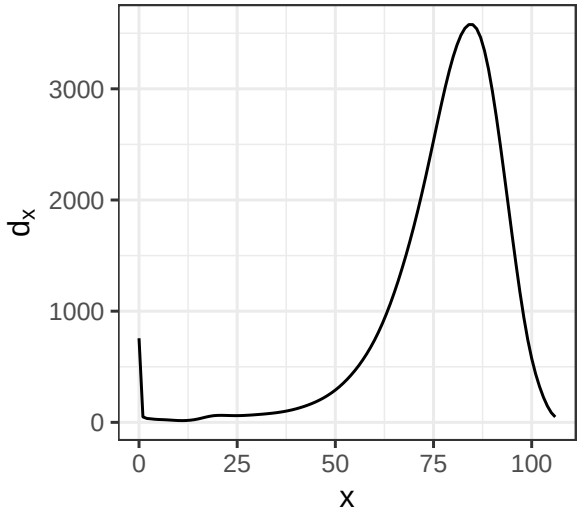
(a) Dødssannsynlighet



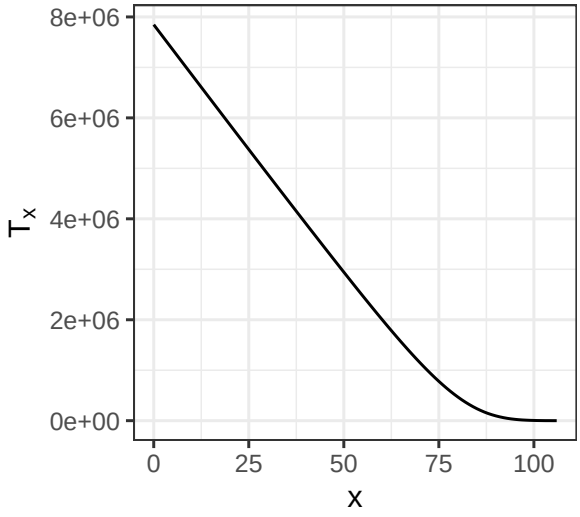
(b) Forventet gjenstående levetid



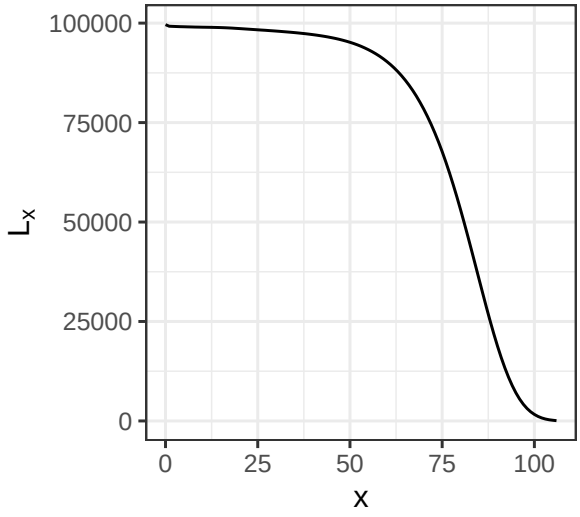
(c) I live



(d) Døde



(e) Sum samtidig levende



(f) Samtidig levende

Figur A.3: Variabler fra livtabell. Livtabellen ble konstruert fra prediksjoner for 2005 fra M_6.

Tillegg B

Kode

B.1 Modeller

```
library(glmmTMB)
itLimit <- glmmTMBControl(optCtrl = list(iter.max = 1e5,
                                         eval.max = 1e5))
# Data
SVE_exposures <- read.delim("data/SVE/Exposures_1x1.txt", sep="")
SVE_deaths <- read.delim("data/SVE/Deaths_1x1.txt", sep="")

# ---- combine_data_full() ----
combine_data_full <- function(Exposures_1x1, Deaths_1x1) {
  data_temp <- data.frame(
    t =  Deaths_1x1$Year,
    x = as.numeric(Exposures_1x1$Age),
    d_m = Deaths_1x1$Male,
    n_m = Exposures_1x1$Male,
    d_k = Deaths_1x1$Female,
    n_k = Exposures_1x1$Female
  )
  data_temp <- na.omit(data_temp)
  df <- data_temp[data_temp$n_m > 0,]
  df <- df[df$n_k > 0,]
}
# ---- combine_data_le_2005() ----
combine_data_le_2005 <- function(Exposures_1x1, Deaths_1x1) {
  data_temp <- data.frame(
    t =  Deaths_1x1$Year,
    x = as.numeric(Exposures_1x1$Age),
    d_m = Deaths_1x1$Male,
    n_m = Exposures_1x1$Male,
    d_k = Deaths_1x1$Female,
    n_k = Exposures_1x1$Female
  )
  data_temp <- na.omit(data_temp)
  df <- data_temp[data_temp$n_m > 0,]
  df <- df[df$n_k > 0,]
  df2005 <- df[df$t <= 2005, ]
}


# ---- nominal_mod_M() ----
nominal_mod_M <- function(df2005) {
  list(
    M0 = glmmTMB(
      data = df2005,
```

```
      formula = d_m ~ factor(x) + factor(t),
      family = poisson,
      offset = log(n_m),
      control = itLimit
    ),
    M1 = glmmTMB(
      data = df2005,
      formula = d_m ~ factor(x) + factor(t),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ 1,
      control = itLimit
    ),
    M2 = glmmTMB(
      data = df2005,
      formula = d_m ~ factor(x) + factor(t),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ t,
      control = itLimit
    ),
    M3 = glmmTMB(
      data = df2005,
      formula = d_m ~ factor(x) + factor(t),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ x,
      control = itLimit
    ),
    M4 = glmmTMB(
      data = df2005,
      formula = d_m ~ factor(x) + factor(t),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ x + t,
      control = itLimit
    )
  )
}
# ---- polynomial_model_M() ----
polynomial_mod_M <- function(df2005) {
  list(
    M5 = glmmTMB(
      data = df2005,
      formula = d_m ~ poly(x, 25) + poly(t, 4),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ 1,
      control = itLimit
    ),
    M6 = glmmTMB(
      data = df2005,
      formula = d_m ~ poly(x, 25) + poly(t, 4),
      family = nbinom2,
      offset = log(n_m),
      dispformula = ~ poly(x, 10) + poly(t, 8),
      control = itLimit
    )
  )
}
# Data 2006-2022
SVE_data_full <- combine_data_full(SVE_exposures, SVE_deaths) # SVE = Sverige
```

```
# Data 1951-2005
SVE_data <- combine_data_le_2005(SVE_exposures, SVE_deaths)
# - Modelle: menn, nominell -
M_names <- c("M0", "M1", "M2", "M3", "M4")
SVE_M <- nominal_mod_M(SVE_data)
# - Modelle: menn, polynom -
M_names_polynomial <- c("M5", "M6")
SVE_M_polynomial <- polynomial_mod_M(SVE_data)
```

B.2 Livtabell

```r
# -- my_mu() --
# Prediksjoner for doedsintensitet.
my_mu <- function(mod, data, method) {
  if (method == 1){
    lambda = predict(mod, type = "response", newdata = data[,c('x','t','n_m')])
  } else if (method == 0){
    lambda = predict(mod, type = "response")
  } else {
    stop("Select method = 0 or method = 1.")
  }
  N    = data$n_m
  mu   = lambda / N
  return(mu)
}

# -- my_q() --
# Doedssannsynlighet / hasard
my_q <- function(mu) {
  q = 1 - exp(-mu)
  return(q)
}

# -- l.at.t() --
# I live
l.at.t <- function(q_data, I){
  l <- c()
  l[1] = 1e5
  for (i in 2:I) {
    l[i] = l[i - 1] * (1 - q_data$q[i - 1])
  }
  return(l)
}

# -- d.at.t() --
# Doede
d.at.t <- function(t_data, I, l){
  d <- c()
  d = l*t_data$q
  return(d)
}

# -- L.at.t() --
# Samtidig levende
L.at.t <- function(l, I){
  L <- c()
  J <- I-1
  for (i in 1:J) {
    L[i] = (l[i] + l[i + 1]) / 2
  }
  L[I] = l[I] / 2
  return(L)
}

# -- T.at.t() --
# Sum samtidig levende
T.at.t <- function(L, I){
  T. <- c()
  for (i in 1:(I)) {
    T.[i] = sum(L[i:I])
  }
```

```
61   return(T.)
62 }
63
64 # -- e.at.t() --
65 # Forventet gjenstaaende levetid
66 e.at.t <- function(T., l){
67   e = T./l
68   return(e)
69 }
70
71 # -- by_t() --
72 by_t <- function(t_data) {
73   I  = length(t_data$x)
74   l  = l.at.t(t_data, I)
75   d  = d.at.t(t_data, I, l)
76   L  = L.at.t(l , I)
77   T. = T.at.t(L , I)
78   e  = e.at.t(T., l)
79   table_t <- cbind(t_data,
80                    e,
81                    l,
82                    d,
83                    T.,
84                    L)
85   return(table_t)
86 }
87
88 # -- my_l() --
89 my_l <- function(q_data) {
90   t = unique(q_data$t)
91   l_by_t  <- data.frame()
92   for (i in t) {
93     l_by_t <- rbind(l_by_t, by_t(q_data[q_data$t == i, ]))
94   }
95   return(l_by_t)
96 }
97
98 # -- my_lifetable() --
99 # Predikerer doedsintensitet fra et objekt og returnerer doedsintensitet med
100 # data.frame().
101 # obs    = 0: livtabell fra prediksjoner.
102 # obs    = 1: livtabell fra observasjoner.
103 # method = 0: prediksjoner paa tilpasset data.
104 # method = 1: prediksjoner paa ny data.
105 my_lifetable <- function(mod, data, method, obs) {
106   if (obs == 0) {
107     mu = my_mu(mod, data, method)
108     q  = my_q(mu)
109     rate_data <- data.frame(t = data$t,
110                             x = data$x,
111                             q)
112     l  = my_l(rate_data)
113     l  = cbind(l, mu)
114     colnames(l) <- c('t','x','q','e','l','d','T','L', 'mu')
115
116   } else if (obs == 1){
117     q = data$d_m / data$n_m
118     rate_data <- data.frame(t = data$t,
119                             x = data$x,
120                             q)
121     l  = my_l(rate_data)
122     colnames(l) <- c('t','x','q','e','l','d','T','L')
```

```
123   } else {
124     stop('Select obs = 0 or obs = 1.')
125   }
126   return(1)
127 }
```

B.3 Overlevelseskurver

```r
# -- surv_cast() --
# Fremskrivelser av overlevelsessannsynligheter
surv_cast <- function(table, s) {
  n    = length(table$q)
  Px   = c(1 - table$q, rep(0, n + 1))
  Px.s <-matrix(data = 1,
                nrow = s + 1,
                ncol = n + 1)

  for (i in 0:n + 1) {
    Px.s[1:s + 1, i] = cumprod(Px[i:(i + s - 1)])
  }
  return(Px.s)
}

# -- NP_forecast() --
# Ikke-parametriske fremskrivinger
# Vi fremskriver i S aar fra kalenderaar t til T
NP_forecast <- function(table, T.){
  t = max(table$t)
  S = length(t:T.)
  s = 0:S

  cum.p = surv_cast(table[table$t==t,], S)
  return(cum.p)
}
```