NTNU
Det skapende universitet

Phase-type modeller for konkurrerende risikoer

Tilpasning til datasett og identifiserbarhet av modellene

Susanne Hodneland Kjølen

Master i fysikk og matematikk
Innlevert: juni 2015
Hovedveileder: Bo Henry Lindqvist, MATH

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgavetekst

- Sette seg inn i teori om konkurrerende- og semi-konkurrerende risikoer, samt bruk av phase-type modeller i slike situasjoner.
- Foreslå enkle og fleksible phase-type modeller for slike problemer og tilpasse disse til simulerte og reelle datasett.
- Studere identifiserbarhet av modellene og en kort innføring i hvordan kovariater kan tas med.

Sammendrag

Phase-type modeller for konkurrerende risikodata blir tilpasset to datasett. Begge datasettene omhandler pasienter innlagt på en intensivavdeling, og deres tid til de konkurrerende risikoene død eller utskrivelse. Det første datasettet er et standard konkurrerende risikoproblem som i tillegg inneholder informasjon om pasientenes tilstedeværelse av pneumoni ved innleggelse. Det andre datasettet inneholder informasjon om tidspunkt for sykehuservervet pneumoni og er med dette av typen semi-konkurrerende risikoer. Disse to datasettene ligger i R-pakkene henholdsvis `mvna` og `kmi`.

Fire ulike phase-type modeller, av typen coxiske modeller, blir studert og tilpasset simulerte data i tillegg til datasettene beskrevet over. Resultatene fra phase-type modellene blir sammenlignet med ikke-parametriske estimatorer, som Nelson-Aalen og Aalen-Johansen. Dette for å teste modellene og vise at de gir gode resultater. Videre blir identifiserbarhet av modellene diskutert og det blir vist at flere av modellene har mer enn én mulig løsning.

Kovariater har blitt tatt med i de enkleste phase-type modellene. Hvor gode modellene er med kovariater blir testet på tre ulike måter. Først blir phase-type modellene med kovariater sammenlignet med resultatene fra standard Cox proporsjonal hasardregresjon. Videre blir resultatene fra modellene med kovariater sammenlignet med resultatene fra å kjøre separate modeller for hver kovariat, i tilfellet med diskrete kovariater. Til slutt blir de kumulative insidensfunksjonene fra modellene sammenlignet med Aalen-Johansen estimatoren.

Abstract

Phase-type models for competing risks data are fitted to two datasets. Both datasets contains information about patients admitted to an intensive care unit, and their time to the competing risks death or discharge alive. The first dataset is a standard competing risks problem, which additionally contains information about the patient's presence of pneumonia on admission. The second dataset contains, in addition, the time for hospital-acquired pneumonia and are hereby of the type semi-competing risks. These two datasets are found in the R packages `mvna` and `kmi` respectively.

Four different phase-type models of the type Cox distributions, are studied and fitted to simulated data, as well as the datasets described above. Results of these phase-type models are compared with non-parametric estimators, such as Nelson-Aalen and Aalen-Johansen, to test if they give good results. Identifiability of the models are discussed and it is shown that several of the models have more than one possible solution.

Covariates are included in the simplest phase-type models. These are tested in three different ways. First, the phase-type models with covariates are compared with results from standard Cox proportional hazard regression. Secondly, results from the models with covariates, in the case of discrete covariates, are compared to the results from fitting separate models corresponding to each covariate. Finally, the cumulative incidence functions of the models are compared to the Aalen-Johansen estimator.

Forord

Denne masteroppgaven er skrevet ved Norges teknisk-naturvitenskapelige universitet, institutt for matematiske fag, og er en del av min mastergrad i fysikk og matematikk, med fordypning innen industriell matematikk. Arbeidet har blitt utført i 10. semester og har foregått over 20 uker.

Masteroppgaven er en fortsettelse av prosjektoppgaven skrevet i faget TMA4500, som en del av fordypningsprosjektet som gjennomføres i 9. semester.

Gjennom hele prosessen har jeg fått strålende oppfølging av min veileder Bo H. Lindqvist. Jeg vil derfor rette en stor takk til ham for et supert samarbeid.

Innhold

Figurer

Tabeller

Kapittel 1

Innledning

Konkurrerende risikoer (eng. competing risks) forekommer når en enhet er utsatt for svikt av flere gjensidig utelukkende årsaker. Dette er en svært vag definisjon, noe som gjenspeiler fenomenets bredde. En enhet i denne sammenheng kan være alt fra en pasient til en produksjonsmaskin. En svikt gir uttrykk for at enheten ikke lenger er i en tilstand definert som fungerende, og årsakene kan være alt fra død for en pasient til strømbrudd for en maskin.

Et konkret eksempel på en konkurrerende risikosituasjon er pasienter innlagt på en intensivavdeling, der de kan svikte ved å enten dø eller bli utskrevet. Deres levetid kan også være avhengig av ulike faktorer, som alder, kjønn eller tilstedeværelse av tilleggssykdommer. Disse faktorene kan tas med i form av kovariater i modellene.

Den konkurrerende risikometoden blir først og fremst brukt til å finne sannsynlighet for å svikte av en bestemt årsak, også kalt risiko. Innen medisin kan dette være til hjelp når det kommer til å informere pasienter om risiko de kan utsettes for i ulike situasjoner, og også hvilken behandling som bør gis. Mer overordnet kan det også si noe om hvordan helsemidler best skal fordeles, for eksempel hvem som har best nytte av en behandling, og det kan være til hjelp med å forstå langtidsvirkninger av kroniske sykdommer. På omtrent samme måte kan den konkurrerende risikometoden brukes innen industrien. Det å vite risikoen for ulike hendelser, kan være til hjelp i avgjørelsen av hvordan sikkerhetssystemer skal settes opp og hvordan midler skal fordeles.

Det er ulike metoder for å modellere konkurrerende risikoer. En av dem er ved bruk av phase-type modeller. Denne typen modeller har blitt studert av flere, blant annet [Aalen, 1995] og [Bladt, 2005]. Modellene har fått mye oppmerksomhet innen anvendt sannsynlighet. Standard phase-type fordelinger blir konstruert ved å se på en homogen markovprosess med endelig tilstandsrom, der alle tilstandene er transiente, bortsett fra én som er absorberende. Med en startfordeling på de transiente tilstandene har tiden til absorpsjon en phase-type fordeling. For å

tilpasse phase-type modeller til konkurrerende risikoer blir det antatt flere absorberende tilstander i markovprosessen. Dette er tidligere sett på av blant annet [Lindqvist, 2013]. Fordelen med phase-type modellene er at de er intuitive. Det er mulig å sette opp de ulike tilstandene i tilstandsdiagram og se på overgangsrater. Det er også enkelt å finne eksplisitte uttrykk for funksjoner som kumulative insidensfunksjoner og årsaksspesifikke hasardrater.

I prosjektoppgaven [Kjølen, 2014] ble phase-type modeller beskrevet og sammenlignet med tradisjonelle metoder, som Cox [Cox, 1972] og Fine og Gray [Fine and Gray, 1999]. I denne masteroppgaven blir disse phase-type modellene bygget videre på og testet ut på både simulerte og reelle datasett. Da masteroppgaven bygger på prosjektoppgaven, er en del av teorien fra prosjektoppgaven også tatt med her. Dette gjelder avsnitt 4.2 til 4.4.

Hovedfokuset i denne oppgaven ligger først og fremst på selve modellene. Det er ikke et mål å tolke datasettene i seg selv, eller finne ut noe om disse. Målet er å tilpasse en modell som gir de samme resultatene som man kan få fra ikke-parametriske estimatorer som Nelson-Aalen estimatoren og Aalen-Johansen estimatoren. Datasettene er her til hjelp for å utvikle en bra phase-type modell, ikke omvendt. Som verktøy i hele denne oppgaven blir programmet R brukt [R Core Team, 2014].

Det blir sett på to datasett i denne oppgaven. Begge omhandler pneumoni, også kjent som lungebetennelse. Det ene sier noe om pasienter som har pneumoni når de blir innlagt, mens det andre inneholder informasjon om pasienter som får pneumoni i løpet av et opphold på intensivavdelingen. De kan se ganske like ut, men er rimelig ulikt bygd opp. Det første er et standard konkurrerende risikoproblem, med to konkurrerende risikoer. Det andre er et semi-konkurrerende risikoproblem.

Et viktig aspekt som har meldt seg underveis i dette arbeidet er identifiserbarhetsproblemer. Dette går ut på hvorvidt det finnes ett og bare ett sett med parametere som gir den samme fordelingsfunksjonen, altså identiske kumulative insidensfunksjoner for en phase-type modell.

Oppbygningen av denne oppgaven er forsøkt gjort så oversiktlig som mulig. Kapittel 2 er en kort beskrivelse av den viktigste litteraturen bak oppgaven. Kapittel 3 er en beskrivelse av de to datasettene som blir analysert og litt om hvordan de er oppbygd. Videre følger kapittel 4 med en samling av viktig grunnleggende teori. Kapittel 5 er en oversikt over de fire phase-type modellene som blir tilpasset i denne oppgaven. Dette kapittelet blir det referert til mange ganger og er ment som en oversikt man kan gå tilbake til og slå opp i. I kapittel 6 blir modellene tilpasset

til simulerte data. Her dukker det opp identifiserbarhetsproblemer som blir sett nærmere på i kapittel 7. Kapittel 8 og 9 er tilpasning til henholdsvis datasett 1 og 2. Kapittel 10 tar for seg kovariater, og kapittel 11 er en oppsummering av foreslått videre arbeid, før en samlet konklusjon i kapittel 12. For oversiktens skyld er det meste av diskusjonen tatt underveis i de respektive kapitlene, slik at konklusjonen er ment som et helhetsbilde.

Kapittel 2

Litteratur

Utgangspunktet for denne masteroppgaven er boken *Competing Risks and Multistate models with R*, [Beyersmann, 2012]. Den tar for seg konkurrerende risikoer og flertilstandsmodeller. Flere datasett blir analysert og to av dem blir brukt i denne oppgaven. En beskrivelse av disse er gitt i kapittel 3. I boken blir det fokusert på hvordan datasettene kan bli analysert i R. Metodene som blir brukt er ikke-parametrisk estimering og Cox proporsjonale hasardregresjon. I denne masteroppgaven blir deler av den samme analysen gjort ved bruk av phase-type modeller. På den måten kan resultater sammenlignes og det er mulig å få en indikasjon på hvor gode phase-type modellene er.

Motivasjonen til bruk av phase-type modeller kommer fra veileder Bo H. Lindvinst [Lindqvist, 2013], og ble også skrevet om i prosjektoppgaven [Kjølen, 2014]. En annen som har jobbet med phase-type modeller er Eric V. Slud og hans artikkel [Slud and Suntornchost, 2014] har vært en bakgrunnsfaktor.

Kapittel 3

Datasett

I dette kapittelet blir det gitt en oversikt og beskrivelse av de to datasettene som phase-type modeller blir tilpasset. Informasjonen rundt studiene som datasettene kommer fra, er hentet fra boken [Beyersmann, 2012], kort beskrevet i kapittel 2.

3.1 Datasett 1: Pneumoni ved innleggelse på intensivavdelingen

Dette datasettet er en del av R-pakken `mvna`, og heter `sir.adm`. Det inneholder et tilfeldig utvalg på 747 pasienter fra kohortstudien SIR 3 (Spread of nosocomial Infections and Resistant pathogens) ved Charité universitetssykehus i Berlin. En kohortstudie er en studie som, over tid, tar for seg personer som har opplevd de samme viktige hendelsene. Her er denne hendelsen det å ha vært innlagt på intensivavdelingen ved Charité universitetssykehus. Datasettet inneholder informasjon om pneumonistatus ved innleggelse, altså om en pasient har pneumoni ved innleggelse eller ikke. Det inneholder også tid tilbrakt på intensivavdelingen og utfallet av intensivbehandlingen, som kan være død eller utskrivelse. I tillegg inneholder det informasjon om pasientenes alder og kjønn. Pneumoni er en alvorlig infeksjon, og er av den grunn forventet å gi forlenget intensivbehandling og økt dødelighet.

De ulike kolonnene i datasettet kan oppsummeres som følger:

`sir.adm$id` Id-nummer til pasient

`sir.adm$pneu` 1 hvis pasienten har pneumoni ved innleggelse, 0 hvis ikke

`sir.adm$status` Status ved endt observasjon. 1 for utskrevet, 2 for død, 0 for sensurert

sir.adm$time Lengden på sykehusoppholdet

sir.adm$age Alder

sir.adm$sex Kjønn, F for kvinne eller M for mann

I undersøkelsen var det 97 pasienter med pneumoni ved innleggelse. Av dem var det 21 som døde, 68 som ble utskrevet og 8 som ble sensurert. Det var totalt 650 pasienter som ikke hadde pneumoni, og av dem døde 55, 589 ble utskrevet og 6 ble sensurert.

Dette datasettet er av typen konkurrerende risikoer. Det er fordi både tid til svikt og type svikt, utskrevet eller død, blir observert.

3.2 Datasett 2: Sykehuservervet pneumoni

Dette datasettet er hentet fra R-pakken **kmi**, og heter `icu.pneu`. Det inneholder et tilfeldig utvalg på 1313 pasienter fra den samme studien som datasett 1. Forskjellen fra det andre datasettet er at det nå blir sett på pneumoni fått på intensivavdelingen, mens det tidligere ble sett på pneumoni ved innleggelse.

Sykehuservervede infeksjoner er et stort problem for helsevesenet. De leder til økt sykelighet, dødelighet og øker lengden på sykehusopphold. Lengden på sykehusopphold er ofte brukt til å tallfeste kostnader for helsevesenet. Ekstra helsekostnader forbundet med sykehuservervede infeksjoner blir brukt i nytte-kostnadsanalyser av smitteverntiltak som for eksempel isolasjonsrom.

I datasettet er hver pasient representert ved enten én eller to rader. Pasienter som får pneumoni under oppholdet har to rader. Ved innleggelse er det, naturlig nok, ingen som har sykehuservervet pneumoni. Den første raden representerer infeksjonsfri periode. Da er infeksjonsstatus `icu.pneu$pneu` lik 0. De som får pneumoni får en ekstra rad som representerer tid med pneumoni til utskrivning, død eller sensurering, da er `icu.pneu$pneu` lik 1.

Kolonnene i datasettet er oppsummert slik:

icu.pneu$start Start på tidsintervall

icu.pneu$stop Slutt på tidsintervall

`icu.pneu$status` 1 hvis observert slutthendelse ved tid `icu.pneu$stop`, 0 ellers. Det vil si at for pasienter med to rader, er denne alltid 0 i første rad.

`icu.pneu$event` Resultat av sykehusopphold. 3 for død, 2 for utskrevet. Denne har ingen betydning dersom `icu.pneu$status==0`.

`icu.pneu$pneu` 0 hvis ikke pneumoni, altså i første rad, 1 hvis pneumoni, altså i andre rad (hvis to rader).

`icu.pneu$age` Alder

`icu.pneu$sex` Kjønn, F for kvinne eller M for mann

I dette datasettet fikk totalt 108 pasienter sykehuservervet pneumoni. Av disse ble 82 utskrevet, 21 døde og 5 ble sensurert. Blant de som ikke fikk pneumoni døde 126 pasienter, 1063 ble utskrevet og 16 ble sensurert.

Dette datasettet er et semi-konkurrerende risikoproblem. Det vil si at det har en ikke-terminerende hendelse som er det å få pneumoni, og to terminerende hendelser som er død og utskrivelse. De terminerende hendelsene sensurerer hverandre og den ikke-terminerende hendelsen, men ikke motsatt. Mer om dette i teoridelen avsnitt 4.6.

Kapittel 4

Teori

I dette kapittelet blir det tatt opp teori som ligger til grunn for resten av oppgaven. Det blir først gitt en introduksjon til levetidsdata generelt. Deretter blir konkurrerende risikoer introdusert. Videre kommer en kort beskrivelse av phase-type fordelinger, før det blir vist hvordan disse kan brukes innen konkurrerende risikoer. Deretter blir det sett på hvordan kovariater kan bli tatt med i modellene, og videre en introduksjon til semi-konkurrerende risikoer. Etter dette følger en oversikt over bruk av likelihoodfunksjonen. Til slutt blir de to ikke-parametriske estimatorene Nelson-Aalen og Aalen-Johansen definert. Disse blir brukt som sammenligningsgrunnlag for store deler av oppgaven.

4.1 Levetidsdata

Innen levetidsanalyse er det viktig å ha en klar, entydig definisjon av tidsorigo, altså der levetidsmålingen starter. En kort beskrivelse av dette finnes i [Kalbfleisch and Prentice, 2011]. Hva som er tidsorigo er helt avhengig av hva som skal analyseres og hva som skal undersøkes. Det kan for eksempel være at tid representerer alder og at tidsorigo er fødsel av individet. I andre sammenhenger kan tidsorigo representere en spesiell begivenhet, som for eksempel innleggelse på sykehus, som er tilfellet for dataene i kapittel 3.

Det er viktig å ha en klar definisjon av hva som forårsaker en svikt. Uten dette kan hele studiegrunnlaget bli feil, og dermed analysen bli meningsløs og uten troverdighet. Dette blir det ikke gått nærmere inn på i denne oppgaven, men antas å være i orden i datasettene.

Levetidsdata inneholder ofte noen enheter som ikke svikter under observasjonstiden. Dataene til slike enheter sies å være høyresensurerte, bare kalt sensurerte

heretter. Det finnes også andre typer sensurering, som venstretrunkering, men det er ikke relevant i denne oppgaven. Det kan være flere årsaker til sensurering. Det kan for eksempel være at enheten overlevde studieperioden uten svikt, eller at enheten forlot studien av andre grunner enn den definerte svikten. Sensurering er uavhengig hvis sviktratene til enhetene under risiko ved hver tid $t > 0$ er de samme som de hadde vært hvis det ikke var noen sensurering. Anta at hasardraten ved tid t, uten sensurering, for en gruppe enheter med kovariat x er

$$\lambda(t; x) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h) | x, T \geq t)}{h},$$

der T er en tilfeldig variabel som representerer tid til svikt. Anta at innen denne gruppen blir enheter sensurert i henhold til en bestemt mekanisme. For gruppen av enheter som er under risiko for svikt ved tid $t > 0$, er sensureringsmekanismen uavhengig hvis hasardraten er $\lambda(t; x)$. Altså kreves det at

$$\lim_{h \to 0} \frac{P(T \in [t, t+h) | x, T \geq t)}{h} = \lim_{h \to 0} \frac{P(T \in [t, t+h) | x, T \geq t, Y(t) = 1)}{h},$$

der $Y(t) = 1$ indikerer at enheten fortsatt er i studien ved tid t. Hvis sensureringen er uavhengig tilfører en enhet som er sensurert $P(T > t; x) = F(t; x)$ til likelihood-funksjonen, se avsnitt 4.7. Altså sier informasjonen om at enheten er sensurert bare at sviktiden er større enn t.

4.2 Konkurrerende risikoer

La fortsatt T være en tilfeldig variabel som representerer tid til svikt for en enhet. Det antas at når en svikt forekommer er det på grunn av én av k ulike årsaker indeksert ved $j \in \{1, 2, \ldots, k\}$. La C være en tilfeldig variabel som representerer årsaken til svikten. Konkurrerende risikodata er dermed av typen (T, C). En nærmere beskrivelse av konkurrerende risikoer i gitt i [Crowder, 2001].

Simultanfordelingen til paret (T, C) er fullstendig spesifisert ved de årsaksspesifikke kumulative fordelingsfunksjonene

$$F_j(t) = P(T \leq t, C = j), \tag{4.2.1}$$

der $t \geq 0$ og $j \in \{1, 2, \ldots, k\}$. Innen medisin blir denne funksjonen ofte kalt kumulativ insidensfunksjon (eng: cumulative incidens function), og denne termen

vil bli brukt i resten av denne oppgaven. De kumulative insidensfunksjonene gir andel enheter som undersøkes ved tid t og som har sviktet av årsak j, tatt hensyn til at svikten kunne vært som følge av andre årsaker.

Videre er de årsaksspesifikke tetthetsfunksjonene gitt ved

$$f_j(t) = F_j'(t).$$

Marginalfordelingen til T er gitt ved summen av de kumulative insidensfunksjonene for alle mulige risikoer, slik at

$$F(t) = P(T \leq t) = \sum_{j=1}^{k} F_j(t).$$

Marginalfordelingen til C er gitt ved

$$\pi_j = P(C = j) = F_j(\infty).$$

Fordelingen til (T, C) kan alternativt bli representert ved årsaksspesifikk hasardrate $\lambda_j(t)$. Dette er momentan risiko for svikt av en spesifisert årsak j, gitt at enheten fortsatt virker etter tid t [Prentice et al., 1978].

$$\lambda_j(t) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h), C = j | T \geq t)}{h} = \frac{f_j(t)}{1 - F(t)}. \tag{4.2.2}$$

4.3 Phase-type fordelinger

En phase-type fordeling er en sannsynlighetsfordeling konstruert som en konvolusjon av eksponentialfunksjoner. Den er et resultat av en eller flere sammenhengende Poissonprosesser som forekommer i rekke, eller faser. Fordelingen kan bli representert ved en tilfeldig variabel som beskriver tid til absorpsjon for en markovprosess. Da representerer hver tilstand i markovprosessen en av fasene. For en fullstendig beskrivelse av phase-type fordelinger se [Neuts, 1981].

La $\{X(t)\}_{t \geq 0}$ være en kontinuerlig-tid markovprosess, med endelig tilstandsrom $S = \{1, 2, \ldots, q, q+1\}$, der tilstandene $1, \ldots, q$ er transiente og tilstanden $q+1$ er

absorberende. For en mer inngående definisjon av markovprosesser se [Ross, 2010]. Prosessen $\{X(t)\}_{t\geq 0}$ har intensitetsmatrise på formen

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q} & \mathbf{L} \\ \mathbf{0}_1 & \mathbf{0}_2 \end{bmatrix},$$

der $\mathbf{Q}$ er en $q \times q$ matrise og $\mathbf{L}$ er en $q \times 1$ vektor. Her representerer $\mathbf{Q}$ intensitetene mellom de transiente tilstandene, mens $\mathbf{L}$ representerer intensitetene fra de transiente til den absorberende tilstanden. Videre er det kjent fra [Ross, 2010] at den tilhørende overgangsmatrisen kan finnes ved

$$\mathbf{P}(t) = e^{\mathbf{A}t} = \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbf{A}^i, \tag{4.3.1}$$

der $P_{ij}(t)$ er sannsynligheten for å gå fra tilstand i til j ved tid t.

Initialfordelingen til $X(t) = X_t$ er gitt ved

$$p_i = P(X_0 = i), \quad i = 1, \ldots, q, \tag{4.3.2}$$

for de transiente tilstandene. For den absorberende tilstanden gjelder

$$P(X_0 = q + 1) = 0.$$

Dette betyr at det ikke er mulig å starte i den absorberende tilstanden.

La T representere tid til absorpsjon i tilstand $q + 1$, definert som

$$T = \inf\{t \geq 0 | X_t = q + 1\}.$$

Da har T en phase-type fordeling, skrevet som $T \sim PH(\mathbf{p}, \mathbf{Q})$.

For å finne fordelingsfunksjonen til T blir det betinget på initialtilstand og hvilken tilstand absorpsjonen skjer fra [Bladt, 2005].

$$\begin{aligned} f(s)ds &= P\left(T \in [s, s+ds)\right) \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} P(T \in [s, s+ds)|X(s)=j, X_0=i)P(X(s)=j|X_0=i)P(X_0=i) \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} L_j P_{ij}^s p_i ds \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} p_i e_{ij}^{\mathbf{Q}s} L_j ds \\ &= \mathbf{p}e^{\mathbf{Q}s}\mathbf{L}ds \end{aligned}$$

Dette gir at dersom $T \sim PH(\mathbf{p}, \mathbf{Q})$ så er tettheten $f(t)$ gitt ved

$$f(t) = \mathbf{p}e^{\mathbf{Q}t}\mathbf{L}.$$

Den kumulative fordelingsfunksjonen, $F(t)$, kan finnes ved integrering, men en enklere metode er å se at $1-F(t)$ er sannsynligheten for at markovprosessen enda ikke har blitt absorbert ved tid t. Dette er det samme som at prosessen er i en av de transiente tilstandene, noe som gir

$$\begin{aligned} 1-F(t) &= 1-P(T \leq t) \\ &= P(T > t) \\ &= P(X(t) \in \{1, 2, \ldots, q\}) \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} P(X(t)=j|X_0=i)P(X_0=i) \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} P_{ij}^t p_i \\ &= \sum_{i=1}^{q}\sum_{j=1}^{q} p_i e_{ij}^{Qt} \\ &= \mathbf{p}e^{\mathbf{Q}t}\mathbf{1}_q, \end{aligned}$$

der $\mathbf{1}_q$ er en $1 \times q$ vektor av bare enere. Den kumulative fordelingsfunksjonen er dermed

$$F(t) = 1 - \mathbf{p}e^{\mathbf{Q}t}\mathbf{1}_q. \tag{4.3.3}$$

Det finnes flere ulike typer phase-type fordelinger. For en oversikt, se [Aalen, 1995]. En av de vanligste er coxisk fordeling. Det er en modell med irreversible overganger. Den er asyklisk og slik at for hver tilstand kan man enten gå til neste tilstand eller direkte til en av de absorberende tilstandene. Det er denne typen modeller det blir sett nærmere på i resten av oppgaven.

4.4 Phase-type modellering for konkurrerende risikoer

For å tilpasse en phase-type modell til det konkurrerende risikoproblemet, blir det nå sett på tilfellet der en markovprosess har flere absorberende tilstander, si k stykker. Dette gir intensitetsmatrisen

$$\mathbf{A} = \begin{bmatrix} \mathbf{Q} & \mathbf{L} \\ \mathbf{0}_1 & \mathbf{0}_2 \end{bmatrix}, \tag{4.4.1}$$

der $\mathbf{Q}$ er som før og $\mathbf{L}$ er en $q \times k$ matrise som gir intensitetene fra de transiente tilstandene til de absorberende. Overgangsmatrisen kan fra (4.3.1) skrives som

$$\mathbf{P}(t) = \begin{bmatrix} e^{\mathbf{Q}t} & \mathbf{Q}^{-1}(e^{\mathbf{Q}t} - \mathbf{I})\mathbf{L} \\ \mathbf{0}_1 & \mathbf{I} \end{bmatrix}.$$

Det som er interessant er de kumulative insidensfunksjonene fra (4.2.1). Tiden til absorpsjon, også kalt hendelsestid, defineres som

$$T = \inf\{t \geq 0 | X_t = q + j\}, \quad j = 1 \dots k.$$

For å finne et funksjonsuttrykk er det lett å se at dersom tiden til absorpsjon, i en bestemt tilstand j, er mindre enn en gitt tid t, så er prosessen i tilstand j ved tid t, fordi det ikke er mulig å forlate en absorberende tilstand. Dermed kan det skrives

$$\begin{aligned} F_j(t) &= P(T \leq t, C = j) \\ &= P(X(t) = j) \\ &= \sum_{i=1}^{K} P(X(t) = j | X(0) = i) P(X(0) = i) \\ &= \sum_{i=1}^{K} p_i P_{ij} \\ &= \mathbf{p}\mathbf{Q}^{-1}(e^{\mathbf{Q}t} - \mathbf{I})\mathbf{L}v_j, \end{aligned} \tag{4.4.2}$$

der v_j er en vektor av størrelse $k \times 1$, med 1 på plass j og 0 ellers. Vektoren $\mathbf{p}$ er initialfordelingen som i (4.3.2). Den årsaksspesifikke tetthetsfunksjonen $f_j(t)$ blir

$$\begin{aligned} f_j(t) &= F_j'(t) \\ &= \frac{d}{dt}\mathbf{p}\mathbf{Q}^{-1}(e^{\mathbf{Q}t} - \mathbf{I})\mathbf{L}v_j \\ &= \mathbf{p}\mathbf{Q}^{-1}\frac{de^{\mathbf{Q}t}}{dt}\mathbf{L}v_j \\ &= \mathbf{p}\mathrm{e}^{\mathbf{Q}t}\mathbf{L}\mathbf{v_j}. \end{aligned} \tag{4.4.3}$$

Den kumulative fordelingsfunksjonen blir

$$\begin{aligned} F(t) &= \sum_{j=1}^{k} F_j(t) \\ &= \sum_{j=1}^{k} \mathbf{p}\mathbf{Q}^{-1}(e^{\mathbf{Q}t} - \mathbf{I})\mathbf{L}v_j \\ &= \mathbf{p}\mathbf{Q}^{-1}e^{\mathbf{Q}t}\mathbf{L}\mathbf{1}_k - \mathbf{p}\mathbf{Q}^{-1}\mathbf{L}\mathbf{1}_k \\ &= -\mathbf{p}e^{\mathbf{Q}t}\mathbf{1}_q + \mathbf{p}\mathbf{1}_q \\ &= 1 - \mathbf{p}e^{\mathbf{Q}t}\mathbf{1}_q, \end{aligned} \tag{4.4.4}$$

der det er brukt at $\mathbf{Q}$ og $\mathbf{L}$ må oppfylle $\mathbf{L}\mathbf{1}_k = -\mathbf{Q}\mathbf{1}_q$ på grunn av egenskapene til en intensitetsmatrise. Funksjonen (4.4.4) er på samme form som i tilfellet med bare én absorberende tilstand i (4.3.3).

De årsaksspesifikke hasardratene (4.2.2) kan nå skrives som

$$\lambda_j(t) = \lim_{h \to 0^+} \frac{P(T \in [t, t+h), C = j | T \geq t)}{h} = \frac{f_j(t)}{1 - F(t)} = \frac{\mathbf{p} \mathrm{e}^{\mathbf{Q}t} \mathbf{L} \mathbf{v}_j}{\mathbf{p} e^{\mathbf{Q}t} \mathbf{1}}, \quad (4.4.5)$$

og de kumulative årsaksspesifikke hasardratene kan finnes ved

$$\Lambda_j(t) = \int_0^t \lambda_j(t) dt. \quad (4.4.6)$$

4.5 Konkurrerende risikodata med kovariater

I levetidsanalyse er det ofte tilgjengelig annen informasjon om enheten eller systemet enn bare levetiden. I datasettene i kapittel 3 er det med informasjon om pasientenes alder og kjønn. Disse variablene kalles kovariater. En introduksjon til levetidsdata med kovariater er gitt i [Ansell and Phillips, 1994].

Anta at det for hver observasjon (T, C), i en konkurrerende risikosituasjon, er en tilhørende kovariatvektor x. Det er ulike måter disse kovariatvektorene kan tas med i modellene. En standard metode er Cox proporsjonal hasardregresjon [Cox, 1972], fra nå av bare kalt Cox-regresjon, der hver årsaksspesifikk hasardrate er assosiert med en kovariatvektor på formen

$$\lambda_j(t; x) = \lambda_0(t) e^{\beta_j x}.$$

Mer om dette i kapittel 10. I phase-type modellene kan kovariater tas med på de ulike overgangsratene med ulike β-vektorer. På denne måten kan en kovariat påvirke ulikt ettersom hvilken tilstand man befinner seg i.

4.6 Semi-konkurrerende risikoer

Semi-konkurrerende risikoer refererer til tilfellet der en hendelsestid kan bli sensurert av en annen hendelsestid, men ikke omvendt. Dette problemet ble først introdusert av [Fine et al., 2001]. Problemet oppstår ofte i forbindelse med studier av kroniske sykdommer og kliniske studier som involverer både terminerende og

ikke-terminerende hendelser. Datasett 2 er et semi-konkurrerende risikoproblem som beskrevet i avsnitt 3.2. Her vil det å få sykehuservervet pneumoni tilsvare en ikke-terminerende hendelse, mens død og utskrivelse vil være terminerende hendelser. En terminerende hendelse kan sensurere en ikke-terminerende hendelse dersom denne skjer først. Det vil si at det er mulig å dø eller bli utskrevet uten å få pneumoni. I motsatt tilfelle sensurerer ikke de ikke-terminerende hendelsene de terminerende. Selv om en pasient får pneumoni kan fortsatt tid til død eller utskrivelse observeres etter dette. Det er nettopp her forskjellen fra vanlig konkurrerende risikoer ligger.

En oversikt over det semi-konkurrerende risikoproblemet kan finnes i artikkelen [Peng et al., 2008]. La T_1 være tid til den ikke-terminerende hendelsen og T_2 tid til den terminerende hendelsen. La videre C være en uavhengig sensureringstid for T_1 og T_2, og $T = T_1 \wedge T_2$, der $\wedge$ er minimumoperatoren. I semi-konkurrerende risikosituasjoner blir det observert data av typen $\{T \wedge C, \delta = I_{\{T_1 \leq T_2 \wedge C\}}\}, Y = T_2 \wedge C, \xi = I_{\{T_2 \leq C\}}$. Kumulativ insidensfunksjon for den ikke-terminerende hendelsen kan uttrykkes som

$$F_1(t) = P(T_1 \leq t, T_2 > T_1).$$

En estimator for denne funksjonen er fra [Peng et al., 2008]

$$\hat{F}_1(t) = \int_0^t \hat{S}_T(u^-)\hat{\Lambda}_1(u).$$

Her er

$$\hat{S}_T(t) = \prod_{T_i \wedge C_i \leq t} \left[1 - \frac{d\bar{N}_1(t) + d\bar{N}_2(t)}{\bar{Y}(T_i \wedge C_i)}\right]$$

og

$$\hat{\Lambda}_1(t) = \int_0^t \frac{d\bar{N}_1(s)}{\bar{Y}(s)} = \sum_{i=1}^{n} \frac{I_{\{T_i \wedge C_i \leq t, \delta_i = 1\}}}{\sum_{l=1}^{n} I_{\{T_i \wedge C_i \leq X_l\}}},$$

der $\bar{N}_j(t) = \sum_{i=1}^{n} N_{j,i}(t)$, $N_{1,i}(t) = I_{\{T_i \wedge C_i \leq t, \delta_i = 1\}}$, $N_{2,i}(t) = I_{\{T_i \wedge C_i \leq t, \delta_i = 0, \xi_i = 1\}}$, $\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)$ og $Y_i(t) = I_{\{T_i \wedge C_i \geq t\}}$.

4.7 Likelihoodfunksjonen

I boken [Meeker and Escobar, 2014], blir ideen bak likelihoodinferens beskrevet som å tilpasse modeller til data ved å finne modellparameterkombinasjoner som gjør sannsynligheten for dataene stor. Modellparameterkombinasjoner med relativt høy sannsynlighet er mer pleusible enn kombinasjoner med lav sannsynlighet. Likelihoodmetoder er generelle og fleksible metoder for å tilpasse modeller til data. Metodene kan brukes på både parametriske og ikke-parametriske modeller med sensurerte data. Likelihoodmetoder fungerer best på store utvalg, altså store datamengder.

Likelihoodfunksjonen er enten lik eller tilnærmet lik proporsjonal med sannsynligheten til dataene. Formen på likelihoodfunksjonen vil avhenge av ulike faktorer som

- Den antatte sannsynlighetsmodellen
- Formen på tilgjengelige data, som for eksempel sensurerte eller intervall sensurerte
- Målet med studien

Den totale likelihooden kan skrives som en simultansannsynlighet for dataene. Ved å anta n uavhengige observasjoner, er utvalgslikelihooden

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; DATA) = \prod_{i=1}^{n} \mathcal{L}_i(\theta, data_i) = \prod_{i=1}^{n} f(data_i; \theta),$$

der $\mathcal{L}_i(\theta; data_i)$ er sannsynligheten for observasjon i, $data_i$ er dataene for observasjon i og θ er en vektor av parameterne som skal estimeres. Dette gjøres ved å finne de parameterne θ som maksimerer likelihoodfunksjonen $\mathcal{L}(\theta)$.

Sensurerte data, se avsnitt 4.1, gir en nedre grense for sviktttiden. Denne typen data bidrar derfor til likelihoodfunksjonen med

$$\mathcal{L}_i(\theta) = \int_{t_i}^{\infty} f(t)dt = F(\infty) - F(t_i) = 1 - F(t_i),$$

der $f(t)$ er tetthetsfunksjonen.

Deles observasjonene inn i to disjunkte sett, $\mathcal{U}$ for usensurerte data og $\mathcal{S}$ for sen-

surerte data, kan den totale likelihooden i dette tilfellet skrives som

$$\mathcal{L}(\theta) = \left\{ \prod_{i \in \mathcal{U}} f(t_i; \theta) \right\} \left\{ \prod_{i \in \mathcal{S}} S(t_i; \theta) \right\}, \tag{4.7.1}$$

der $S(t_i; \theta)$ er overlevelsesfunksjonen.

4.7.1 Profil-likelihood

Anta at de ukjente parameterne i likelihoodfunksjonen er θ, og at disse kan partisjoneres slik at $\theta' = (\alpha', \beta')$. Det er nødvendig å estimere både α og β, men interessen ligger bare i å teste α og konstruere konfidensintervall for denne. Til dette brukes profil-likelihood. Det går ut på å sette α, og maksimere likelihooden med hensyn på β for ulike α. Med andre ord finnes

$$\hat{\alpha}_T = \arg\max_{\alpha} \mathcal{L}_\alpha(\hat{\beta}_\alpha) = \arg\max_{\alpha} \mathcal{L}_T(\alpha, \hat{\beta}_\alpha).$$

For å tegne profil-likelihood, settes α til diskrete verdier i et intervall. Likelihooden maksimeres for de andre parameterne og plottes som funksjon av α. På denne måten kan det sjekkes om likelihooden maksimerer riktig dersom α er en usikker parameter.

4.8 Ikke-parametriske estimatorer

Alle resultater fra phase-type modellene i denne oppgaven sammenlignes med ikke-parametriske estimatorer. De kumulative årsaksspesifikke hasardratene sammenlignes med Nelson-Aalen estimatoren gitt ved

$$\hat{\Lambda}_j(t) = \sum_{T_i \wedge C_i \leq t} \frac{d\bar{N}_j(T_i \wedge C_i)}{\bar{Y}(T_i \wedge C_i)},$$

for årsak j, der notasjonen er som i avsnitt 4.6. De kumulative insidensfunksjonene blir sammenlignet med Aalen-Johansen estimatoren gitt ved

$$\hat{P}(T \leq t, X_T = j) = \sum_{T_i \wedge C_i \leq t} \hat{P}(T > (T_i \wedge C_i)-) \frac{d\bar{N}_j(T_i \wedge C_i)}{\bar{Y}(T_i \wedge C_i)},$$

der

$$\hat{P}(T > t-) = \prod_{T_i \wedge C_i < t} \left(1 - \frac{d\bar{N}_j(T_i \wedge C_i)}{\bar{Y}(T_i \wedge C_i)}\right).$$

For en mer detaljert beskrivelse av disse og hvordan de kan finnes ved hjelp av R, se [Beyersmann, 2012].

Kapittel 5

Modellbeskrivelse

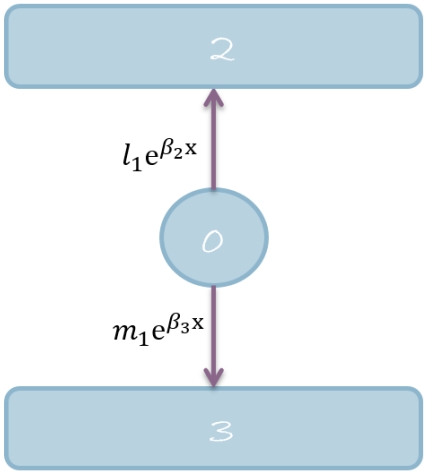
I resten av denne oppgaven blir det sett på fire phase-type modeller av typen coxiske fordelinger beskrevet i avsnitt 4.3, med kovariater på de to første, som beskrevet i avsnitt 4.5.

De to første modellene beskrevet her er hentet fra prosjektoppgaven [Kjølen, 2014]. Nummereringen av tilstandene er valgt å beholde for oversiktlighetens skyld. Dette forklarer den noe ulogiske nummereringen av tilstandene for modell 3 og 4. Den første modellen består kun av tre tilstander der to av dem er absorberende. Modell 2 har en ekstra transient tilstand og det er også mulig å gå fra denne til de absorberende tilstandene. De to siste modellene har henholdsvis tre og fire transiente tilstander, men fortsatt bare mulighet for å gå til de absorberende tilstandene fra to av dem.

Det har blitt testet mange flere modellvarianter enn de som er vist i dette kapittelet. Her er det valgt et representativt utvalg for å understreke interessante poeng. Det finnes variasjoner i det uendelige og det er mye mer som kan studeres ved disse. Modellene representert her er valgt med tanke på at de samme modellene skal kunne brukes på ulike måter og for begge datasettene. Det har derfor blitt lagt vekt på fleksibilitet.

5.1 Modell 1: Tre tilstander, der én er transient

Den første phase-type modellen består av tre tilstander, der to av dem er absorberende. Det fører til at systemet har to overgangsrater a_{0j}, $j = 2, 3$. Fra tilstand 0 til 2 er $a_{02} = l_1 e^{\beta_2 x}$ og fra tilstand 0 til 3 er $a_{03} = m_1 e^{\beta_3 x}$. Variabelen x er en vektor av kovariater, l_1 og m_1 er konstanter, β_2 og β_3 er vektorer av kovariatkoeffisienter. Tilstandsdiagram for tilhørende markovformulering er vist i figur 5.1.



Figur 5.1: Tilstandsdiagram for markovformulering tilhørende modell 1. Tilstand 0 er transient og tilstand 2 og 3 er absorberende. Overgangsratene er gitt på de to overgangene mellom transient og absorberende tilstander. Videre er x en vektor av kovariater, β_2 og β_3 er vektorer av kovariatkoeffisienter og l_1 og m_1 er konstanter.

Intensitetsmatrisen til dette systemet blir

$$\mathbf{A}(\mathbf{x}) = \begin{array}{c} \\ 0 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 2 & 3 \end{array} \\ \begin{pmatrix} -(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}) & l_1 e^{\beta_2 x} & m_1 e^{\beta_3 x} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array},$$

slik at fra (4.4.1) er

$$\mathbf{Q}(\mathbf{x}) = -(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}) \qquad \text{og} \qquad \mathbf{L}(\mathbf{x}) = [l_1 e^{\beta_2 x}, m_1 e^{\beta_3 x}].$$

De årsaksspesifikke tetthetsfunksjonene for årsak 2 og 3 er fra (4.4.3) gitt som

$$\begin{aligned} f_2(t;x) &= l_1 e^{\beta_2 x} e^{(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x})t} \\ f_3(t;x) &= m_1 e^{\beta_3 x} e^{(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x})t} \end{aligned} \tag{5.1.1}$$

Uttrykk for årsaksspesifikk hasardrate finnes ved bruk av (4.4.5). Den gir at hasardratene til tilstand 2 og 3 blir

$$\begin{aligned}\lambda_2(t;x) &= l_1 e^{\beta_2 x} \\ \lambda_3(t;x) &= m_1 e^{\beta_3 x}\end{aligned} \tag{5.1.2}$$

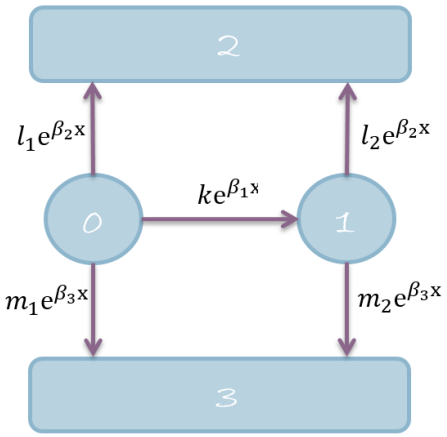
og de kumulative insidensfunksjonene fås fra (4.4.2) og blir for årsak 2 og 3

$$\begin{aligned}F_2(t;x) &= \frac{\left(1 - e^{-t\left(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}\right)}\right) l_1 e^{\beta_2 x}}{l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}} \\ F_3(t;x) &= \frac{\left(1 - e^{-t\left(l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}\right)}\right) m_1 e^{\beta_3 x}}{l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}}.\end{aligned} \tag{5.1.3}$$

Denne modellen blir tilpasset begge datasettene beskrevet i kapittel 3. Det mest interessante med den, er studien av kovariater i kapittel 10. I den sammenheng er denne modellen hensiktsmessig, siden den kan sammenlignes med Cox-regresjon.

5.2 Modell 2: Fire tilstander, der to er transiente

Modell 2 er en utvidelse av modell 1 i den forstand at det blir lagt til en ekstra transient tilstand, kalt tilstand 1. Det er mulig å gå fra tilstand 0 til 1 og fra tilstand 1 til de absorberende. Ellers er den lik som modell 1. Kovariater er lagt på alle overgangene. Tilstandsdiagram for denne modellen er vist i figur 5.2.



Figur 5.2: Tilstandsdiagram for markovformulering tilhørende modell 2. Tilstand 0 og 1 er transiente og tilstand 2 og 3 er absorberende. Videre er x en vektor av kovariater, β_j, $j = 1, 2, 3$, er vektorer av kovariatkoeffisenter og k, l_1, l_2, m_1, m_2 er konstanter.

Innen medisin kan de transiente tilstandene i en slik modell svare til ulike stadier i et sykdomsforløp, mens de absorberende er ulike avslutninger på et sykehusopphold, som død og utskrivning. Fordelen med en slik modell kontra den med én transient tilstand, er at det nå er mulig å skille absorpsjonsratene fra de ulike transiente tilstandene, dersom disse har fått ulike betydninger. Altså kan det bli tatt hensyn til at det for eksempel er større sjanse for å dø dersom sykdommen har forverret seg.

For denne modellen blir intensitetsmatrisen og de ulike funksjonene store og uoversiktlige. Det blir derfor innført at

$$a(x) = ke^{\beta_1 x} + l_1 e^{\beta_2 x} + m_1 e^{\beta_3 x}$$
$$b(x) = l_2 e^{\beta_2 x} + m_2 e^{\beta_3 x}.$$

Valget av disse er ikke tilfeldig. Dette er de negative av diagonalen i $\mathbf{Q}$-matrisen (5.2.1), og dermed også egenverdiene til systemet. Derfor går disse uttrykkene igjen og det blir også vist at disse har en spesiell betydning i forbindelse med identifiserbarhet i kapittel 7.

Intensitetsmatrisen til denne modellen kan da skrives som

$$\mathbf{A}(\mathbf{x}) = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} -a(x) & ke^{\beta_1 x} & l_1 e^{\beta_2 x} & m_1 e^{\beta_3 x} \\ 0 & -b(x) & l_2 e^{\beta_2 x} & m_2 e^{\beta_3 x} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

slik at fra (4.4.1) er

$$\mathbf{Q}(\mathbf{x}) = \begin{bmatrix} -a(x) & ke^{\beta_1 x} \\ 0 & -b(x) \end{bmatrix} \tag{5.2.1}$$

og

$$\mathbf{L}(\mathbf{x}) = \begin{bmatrix} l_1 e^{\beta_2 x} & m_1 e^{\beta_3 x} \\ l_2 e^{\beta_2 x} & m_2 e^{\beta_3 x} \end{bmatrix}.$$

Den årsaksspesifikke tetthetsfunksjonen for årsak 2 er fra (4.4.3)

$$f_2(t) = \left[l_1 - \frac{ke^{\beta_1 x} l_2}{a(x) - b(x)} \right] e^{-a(x)t + \beta_2 x} + \frac{ke^{\beta_1 x} l_2}{a(x) - b(x)} e^{-b(x)t + \beta_2 x}. \tag{5.2.2}$$

Som tidligere blir de årsaksspesifikke hasardratene funnet fra (4.4.5). For årsak 2 kan den skrives som

$$\lambda_2(t; x) = \frac{l_1 e^{\beta_2 x}(a(x) - b(x))e^{-a(x)t} + ke^{\beta_1 x}\left(e^{-b(x)t} - e^{-a(x)t}\right) l_2 e^{\beta_2 x}}{(a(x) - b(x))e^{-a(x)t} + ke^{\beta_1 x}\left(e^{-b(x)t} - e^{-a(x)t}\right)}. \tag{5.2.3}$$

De kumulative insidensfunksjonene blir funnet fra (4.4.2) og er for årsak 2

$$\begin{aligned} F_2(t; x) = & \frac{\left(1 - e^{-a(x)t}\right) l_1 e^{\beta_2 x} - \frac{ke^{\beta_1 x}\left(e^{-b(x)t} - e^{-a(x)t}\right) l_2 e^{\beta_2 x}}{a(x) - b(x)}}{a(x)} \\ & - \frac{ke^{\beta_1 x}\left(e^{-b(x)} - 1\right) l_2\, e^{\beta_2\, x}}{a(x)b(x)} \end{aligned} \tag{5.2.4}$$

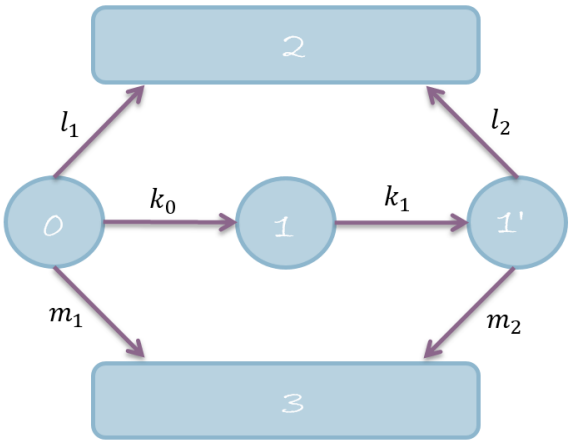
De tilsvarende funksjonene for årsak 3 er som for årsak 2, bare med m_1 i stedet for l_1 og m_2 i stedet for l_2.

Denne modellen er svært fleksibel og blir brukt i alle sammenhenger beskrevet i denne oppgaven.

5.3 Modell 3: Fem tilstander, der tre er transiente

Modell 3 har 5 tilstander, der tre av dem er transiente. Det er som før mulig å gå fra tilstand 0 til tilstand 1, 2 og 3. Fra tilstand 1 er det kun mulig å gå til tilstand 1', og fra 1' til tilstand 2 og 3. Tilstandsdiagram for modell 3 er vist i figur 5.3.

Kovariater er ikke tatt med i denne modellen. Det kunne vært tatt med kovariater på samme måte som for de forrige modellene, men dette er valgt å ikke gjøre. Hovedfokuset her ligger i å finne enkle intuitive modeller. Det er likevel forsøkt å bruke denne modellen med kovariater, men da det ikke gav noen spesielle resultater eller tilførte noe nytt i forhold til modell 2, velges det å ikke fokusere på dette. Det er heller i forbindelse med identifiserbarhet denne modellen er interessant.



Figur 5.3: Tilstandsdiagram for markovformulering tilhørende modell 3. Tilstand 0, 1 og 1' er transiente, mens tilstand 2 og 3 er absorberende. Videre er k_0, k_1, l_1, l_2, m_1 og m_2 konstanter, og det er ikke tatt med kovariater i denne modellen.

Definerer igjen to variabler tilsvarende to av de negative av egenverdiene til **Q**-matrisen (5.3.1),

$$a = k_0 + l_1 + m_1$$
$$b = l_2 + m_2.$$

Intensitetsmatrisen kan da skrives som

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 1' \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 1' & 2 & 3 \end{array} \\ \begin{pmatrix} -a & k_0 & 0 & l_1 & m_1 \\ 0 & -k_1 & k_1 & 0 & 0 \\ 0 & 0 & -b & l_2 & m_2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

slik at fra (4.4.1) er

$$\mathbf{Q} = \begin{bmatrix} -a & k_0 & 0 \\ 0 & -k_1 & k_1 \\ 0 & 0 & -b \end{bmatrix} \tag{5.3.1}$$

og

$$\mathbf{L} = \begin{bmatrix} l_1 & m_1 \\ 0 & 0 \\ l_2 & m_2 \end{bmatrix}.$$

Den årsaksspesifikke tetthetsfunksjonen for årsak 2 er, fra (4.4.3), for denne modellen,

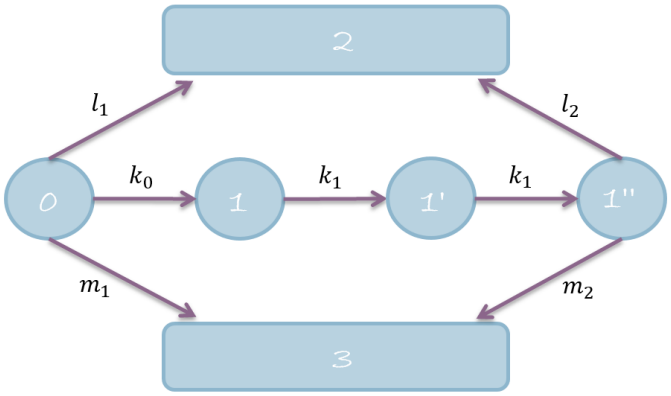
$$\begin{aligned} f_2(t) = &\left[\frac{k_0 k_1 l_2}{(b-k_1)(a-k_1)}\right] e^{-k_1 t} + \left[l_1 + \frac{k_0 k_1 l_2}{(a-b)(a-k_1)}\right] e^{-at} \\ &+ \left[-\frac{k_0 k_1 l_2}{(b-k_1)(a-b)}\right] e^{-bt}, \end{aligned} \tag{5.3.2}$$

og tilsvarende for årsak 3 med m_1 i stedet for l_1 og m_2 i stedet for l_2.

Det er ikke av interesse å skrive ut uttrykkene for de andre funksjonene for denne modellen, da de blir svært store. De kan likevel finnes som for de andre modellene om ønskelig. Dette gjelder også for modell 4.

5.4 Modell 4: Seks tilstander, der fire er transiente

Denne modellen har enda en transient tilstand, som vist i figur 5.4. Total rate ut fra tilstand 1 og 1' er lik. Det gjør at overgangstiden fra 1 til 1" blir gammafordelt.



Figur 5.4: Tilstandsdiagram for markovformulering tilhørende modell 4. Tilstandene 0, 1, 1' og 1" er transiente, mens tilstand 2 og 3 er absorberende. Overgangsratene ut fra tilstand 1 og 1' er like. Videre er k_0, k_1, l_1, l_2, m_1 og m_2 konstanter, og det er ikke tatt med kovariater i denne modellen.

Intensitetsmatrisen er gitt som

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 1' \\ 1'' \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 1' & 2 & 3 & \end{array} \\ \begin{pmatrix} -a & k_0 & 0 & 0 & l_1 & m_1 \\ 0 & -k_1 & k_1 & 0 & 0 & 0 \\ 0 & 0 & -k_1 & k_1 & 0 & 0 \\ 0 & 0 & 0 & -b & l_2 & m_2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

slik at fra (4.4.1) er

$$\mathbf{Q} = \begin{bmatrix} -a & k_0 & 0 & 0 \\ 0 & -k_1 & k_1 & 0 \\ 0 & 0 & -k_1 & k_1 \\ 0 & 0 & 0 & -b \end{bmatrix} \qquad \text{og} \qquad \mathbf{L} = \begin{bmatrix} l_1 & m_1 \\ 0 & 0 \\ 0 & 0 \\ l_2 & m_2 \end{bmatrix}.$$

Matrisen $\mathbf{Q}$ har to like verdier på diagonalen, som tilsvarer at den ene egenverdien er av multiplisitet 2. Av den grunn blir et ledd i de årsaksspesifikke tetthetsfunksjonene multiplisert med t. For årsak 2 er den årsaksspesifikke tetthetsfunksjonen fra (4.4.3)

$$f_2(t) = \left[\frac{k_0 k_1^2 l_2(a+b-2k_1)}{(a-k_1)^2(b-k_1)^2}\right] e^{-k_1 t} + \left[-\frac{k_0 k_1^3 l_2(a+b-k_1)}{(a-k_1)^2(b-k_1)^2}\right] te^{-k_1 t}$$
$$+ \left[l_1 + \frac{k_0 k_1^2 l_2}{(a-b)(a-k_1)^2}\right] e^{-at} + \left[-\frac{k_0 k_1^2 l_2}{(b-k_1)^2(a-b)}\right] e^{-bt},$$

og for årsak 3 er den tilsvarende med m_1 i stedet for l_1 og m_2 i stedet for l_2.

Kapittel 6

Modelltilpasning: Simulert datasett

Første steg for å teste modellene beskrevet i kapittel 5 er å simulere data fra dem og deretter estimere parameterne når de riktige parameterne er kjente. Dette vil forsikre at metodene brukt er riktige og gi en indikasjon på hvor bra estimeringen er. På denne måten ble det oppdaget at noen av modellene kan ha flere løsninger. Dette er nærmere vist i kapittel 7.

Alle modellene og metodene brukt i denne oppgaven er testet på simulerte datasett, men siden fremgangsmåten er ganske lik og det ikke oppsto noen interessante problemer annet enn for modell 2, er det bare denne som blir beskrevet her, avsnitt 6.3.

Algoritmen for å simulere data fra modellene er beskrevet i avsnitt 6.1. For å maksimere likelihoodfunksjonen brukes R-funksjonen `optim`. Denne er beskrevet i avsnitt 6.2.

6.1 Simuleringsalgoritme

For å simulere data fra modell 2, se figur 5.2, brukes simuleringsalgoritmen 1, som også ble brukt i prosjektoppgaven [Kjølen, 2014]. Denne kan lett modifiseres til å simulere fra de andre modellene.

Algoritme 1 Simulere data fra modell 2

Sett parameterene k, l_1, m_1, l_2, m_2
for $i = 1 \dots n$ **do**
  Trekk x
  Definer A som intensitetsmatrisen
  Trekke t fra eksponentialfordeling med parameter $A[1,1]$
  Trekk r fra multinomisk fordeling
  if $r[2] = 1$ **then**
    $c = 2$
  else if $r[3] = 1$ **then**
    $c = 3$
  else
    Trekk t_2 fra eksponentialfordeling med parameter $A[2,2]$, sett $t = t + t_2$
    Trekk r fra binomisk fordeling
    if $r[1] = 1$ **then**
      $c = 2$
    else$c = 3$
    end if
    Sett $C[i] = c$, $X[i] = x$ og $T[i] = t$
  end if
end for

6.2 R-funksjonen `optim`

For å tilpasse phase-type modellene til datasett brukes sannsynlighetsmaksimering. Den innebygde funksjonen `optim` i R brukes for å maksimere likelihoodfunksjonene. For hver modell blir det skrevet en funksjon i R som gir likelihoodfunksjonen som i avsnitt 4.7. Som standard minimerer `optim`, slik at for å maksimere blir den negative likelihoodfunksjonen brukt.

Funksjonen `optim` kan bruke flere ulike optimeringsalgoritmer. I denne oppgaven blir metoden 'L-BFGS-B' brukt. Dette er en begrenset-minne modifikasjon av en kvasi-Newton metode. Grunnen til at denne er valgt er fordi den gjør det mulig å sette grenser for parameterne. Dette er nyttig siden parameterne i modellen skal være positive. Det blir derfor satt en nedre og øvre grense for parameterne i alle optimeringene.

Det er mulig å finne standardfeil for parameterne fra `optim`. Ved å la argumentet

`hessian==TRUE`, gir `optim` ut hessianmatrisen. Standardfeil finnes fra denne ved å ta kvadratroten av diagonalen til den inverse av hessianmatrisen.

6.3 Modell 2

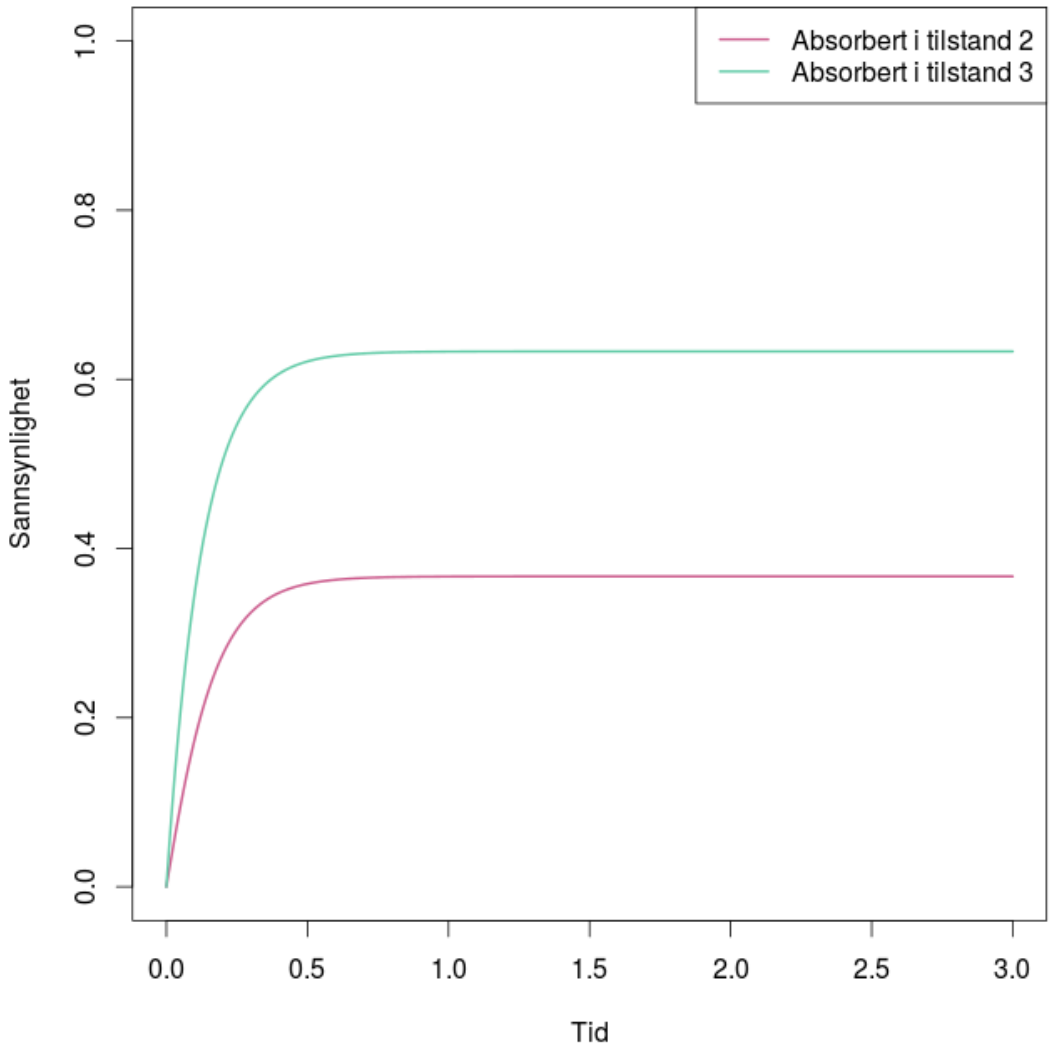
For å simulere data fra modell 2, se figur 5.2, brukes algoritmen 1, og eksempelparameterne $k = 1$, $l_1 = 2$, $l_2 = 16$, $m_1 = 5$ og $m_2 = 1$. Det blir simulert 1 000 000 data, uten kovariater. Ved å maksimere den tilhørende likelihoodfunksjonen, se avsnitt 8.1.2, ved bruk av `optim`, kan parameterestimater oppnås.

Tabell 6.1 viser de estimerte parameterverdiene ved ulike startverdier. Begge estimeringene har nedre grense 10^{-5} og øvre grense 20 for alle parameterne. Dette virker fornuftig da de valgte parameterne har verdier fra 1 til 16. Ved å sette startverdiene lik 5 for alle parameterne ser det ut til at `optim` ikke klarer å estimere riktige parametere. Det eneste den klarer er parameterne m_1 og l_1. Ved å sette startverdiene lik 1 blir parameterne estimert ganske bra, og er svært like de som ble satt i simuleringen. Ekstremparameteren l_2 som var satt til 16 i simuleringen, blir estimert til 15.94, som er ganske bra. Grunnen til at verdiene ikke blir estimert helt eksakt er at de beregnes ut fra en begrenset datamengde. Ved en mindre mengde data blir de enda mer unøyaktige, og dette kommer best til syne for ekstremparameteren. De samme to parameterkombinasjonene kan fås ved å avgrense området for k til å være mellom 10^{-5} og 2.

Tabell 6.1: Estimerte parametere tilhørende modell 2, fra simulerte data. Originalparameterne er de som ble satt i simuleringsalgoritmen. Startverdiene er parameterne satt i `optim`.

	k	l_1	l_2	m_1	m_2
Originalparametere	1	2	16	5	1
Startverdi	5	5	5	5	5
Estimat	9.642560	2.003989	3.402368	5.012515	4.581295
Standardfeil	0.300673	0.011582	0.017120	0.013186	0.116879
Startverdi	1	1	1	1	1
Estimat	0.967338	2.004074	15.942085	5.012401	0.715405
Standardfeil	0.026790	0.011583	0.531679	0.013204	0.325604

For disse to parameterkombinasjonene blir de kumulative insidensfunksjonene de samme, se figur 6.1.



Figur 6.1: Kumulative insidensfunksjoner fra modell 2 for simulerte data. De kumulative insidensfunksjonene for originalparameterne og for de estimerte parameterne sammenfaller og er plottet oppå hverandre.

Ved å studere parameterne i tabell 6.1 er det mulig å merke seg noe interessant. Dersom de øverste parameterne blir uttrykt med hatt, viser det seg følgende sam-

menheng:

$$\begin{aligned} \hat{a} &= 16.65906 \\ a &= 7.98381 \\ \hat{b} &= 7.98366 \\ b &= 16.65749, \end{aligned}$$

der $a = k + l_1 + m_1$ og $b = l_2 + m_2$. Det ser med andre ord ut til at $\hat{a} = b$ og $\hat{b} = a$. I tillegg er $\hat{l_1} = l_1$ og $\hat{m}_1 = m_1$.

Det som er interessant videre er å finne ut om den første estimeringen er feil, eller om det faktisk er slik at det finnes en sammenheng og flere optimale løsninger. Dette blir det sett nærmere på i kapittel 7.

Kapittel 7

Identifiserbarhet av modellene

En phase-type modell sies å være entydig identifiserbar dersom det finnes ett og bare ett sett med parametere som gir den samme fordelingen for (T, C), altså samme kumulative insidensfunksjoner. Det vil igjen si om det finnes ett og bare ett sett av parametere som maksimerer likelihoodfunksjonen. For modell 1 er det greit å finne et analytisk uttrykk for likelihoodfunksjonen, dette blir senere gjort i avsnitt 8.1.1, og det er dermed lett å se at det bare er én løsning for den. De andre modellene derimot er ikke like enkle, og funksjonene blir kompliserte. For å finne ut om det er flere sett parametere som gir samme kumulative insidensfunksjoner blir det sett nærmere på de årsaksspesifikke tetthetsfunksjonene. Grunnen til dette er at de årsaksspesifikke tetthetsfunksjonene har den enkleste formen å jobbe med. Formen til disse bestemmes av egenverdiene til $\mathbf{Q}$-matrisen (4.4.1). De ulike løsningene finnes ved å permutere og det er dermed et endelig antall muligheter for hver modell. Dersom noen egenverdier er multiple vil det bli tilsvarende færre mulige permuteringer. Modell 2 og 3 blir sett nærmere på i dette kapittelet. Resultatene vil bli tilsvarende for modell 4. Kovariater er ikke med i modellene her.

7.1 Modell 2

I denne modellen er det fem ukjente parametere, k, l_1, m_1, l_2 og m_2. Den årsaksspesifikke tetthetsfunksjonen for årsak 2 er som i (5.2.2). Generelt kan disse funksjonene for årsak 2 og 3 skrives som

$$f_2(t) = A_2 e^{\lambda_1 t} + B_2 e^{\lambda_2 t}, \tag{7.1.1}$$

der $A_2 = l_1 - \frac{kl_2}{a-b}$ og $B_2 = \frac{kl_2}{a-b}$, og

$$f_3(t) = A_3 e^{\lambda_1 t} + B_3 e^{\lambda_2 t}, \tag{7.1.2}$$

der $A_3 = m_1 - \frac{km_2}{a-b}$ og $B_3 = \frac{km_2}{a-b}$. Parameterne $\lambda_1 = a$ og $\lambda_2 = b$ er egenverdiene til den tilhørende **Q**-matrisen fra (5.2.1).

La A_2, B_2, A_3, B_3, λ_1 og λ_2 være kjent fra modellen eller uendelig mange datapunkter. Det finnes da en løsning for hver kombinasjon av disse. Det vil si at enten er $\lambda_1 = -a$, $\lambda_2 = -b$, $A_2 = l_1 - \frac{kl_2}{a-b}$, $B_2 = \frac{kl_2}{a-b}$, $A_3 = m_1 - \frac{km_2}{a-b}$ og $B_3 = \frac{km_2}{a-b}$, eller omvendt slik at $\lambda_2 = -a$, $\lambda_1 = -b$, $B_2 = l_1 - \frac{kl_2}{a-b}$, $A_2 = \frac{kl_2}{a-b}$, $B_3 = m_1 - \frac{km_2}{a-b}$ og $A_3 = \frac{km_2}{a-b}$. Dette tilsvarer å velge om $\lambda_1 > \lambda_2$ eller $\lambda_1 < \lambda_2$. Begge disse løsningene vil gi samme funksjoner, og dermed være like gode løsninger. Ved å løse lignignene for A_2, A_3, B_2 og B_3 i tillegg til at $a = k + l_1 + m_1$ og $b = l_2 + m_2$, kan de ulike parameterne finnes ved uttrykkene

$$\begin{aligned} k &= \frac{(B_2 + B_3)(a-b)}{b} \\ l_2 &= \frac{B_2 b}{B_2 + B_3} \\ l_1 &= A_2 + B_2 \\ m_1 &= A_3 + B_3 \\ m_2 &= \frac{B_3 b}{B_2 + B_3}. \end{aligned} \tag{7.1.3}$$

Et eksempel på en slik situasjon er de årsaksspesifikke tetthetsfunksjonene

$$\begin{aligned} f_2(t) &= 5e^{-4t} - 3e^{-5t} \\ f_3(t) &= 3e^{-4t} - 2e^{-5t}. \end{aligned} \tag{7.1.4}$$

De to mulige løsningene har kombinasjonene vist i tabell 7.1. Kolonnen 'kombinasjoner' illustrere rekkefølgen på leddene i (7.1.4). For kombinasjonen AB gjelder $A_2 = 5$ og $B_2 = -3$. For kombinasjonen BA har A_2 og B_2 byttet verdier, og slik fortsetter det.

Tabell 7.1: De to mulige kombinasjonene av hva A_j, B_j, C_j, λ_1 og λ_2 kan være, for eksempelfunksjonene (7.1.4). Kolonnen 'Kombinasjoner' er et uttrykk for rekkefølgen av leddene i (7.1.1) og (7.1.2).

Kombinasjoner	A_2	B_2	A_3	B_3	λ_1	λ_2
AB	5	-3	3	-2	4	5
BA	-3	5	-2	3	5	4

Fra uttrykkene i (7.1.3) kan parameterne for de to tilfellene bestemmes, og disse er vist i tabell 7.2.

Tabell 7.2: De to mulige løsningene for modell 2, for eksempelfunksjonene (7.1.4).

Kombinasjoner	k	l_1	m_1	l_2	m_2
AB	1	2	1	3	2
BA	2	2	1	$\frac{5}{2}$	$\frac{3}{2}$

7.1.1 Identifiserbarhet dersom p er ukjent

Over er det antatt at vektoren av startsannsynligheter er $\mathbf{p} = [1, 0]$. Det betyr at sannsynligheten er 1 for å starte i den første tilstanden. Spørsmålet er om situasjonen blir annerledes dersom starttilstandene er ukjente.

Dersom $\mathbf{p} = [p_1, 1 - p_1]$ er en konstant, vil det alltid finnes to løsninger, der $\mathbf{p}$ er lik for begge. Fra (4.4.3), med samme parameterverdier som over og p_1 ukjent, blir de årsaksspesifikke tetthetsfunksjonene

$$\begin{aligned} f_2(t; p_1) &= 5p_1e^{-4t} + (3 - 6p_1)e^{-5t} \\ f_3(t; p_1) &= 3p_1e^{-4t} + (2 - 4p_1)e^{-5t}. \end{aligned}$$

Dermed følger de samme uttrykkene som (7.1.3), med nye A_2, B_2, A_3 og B_3.

Dersom $\mathbf{p}$ ikke er en konstant, slik som tilfellet er dersom enheter starter i ulike tilstander, vil det for modell 2 kun være én løsning. Dette kommer av at de som starter i tilstand 1 følger modell 1 og har dermed bare én løsning. Da vil l_2 og m_2 være kjent. Da er også $b = l_2 + m_2$ kjent og leddene med a og b kan ikke bytte plass som beskrevet over, og det er dermed kun én mulig løsning. Dette er vist for datasett 1 i kapittel 8 avsnitt 8.2.1.

7.2 Modell 3

Denne modellen har seks ukjente parametere, k_0, k_1, l_1, m_1, l_2 og m_2. Matrisen $\mathbf{Q}$ fra (5.3.1) tilhørende denne modellen har tre egenverdier kalt λ_1, λ_2 og λ_3. Generelt uttrykk for de årsaksspesifikke tetthetsfunksjonene er

$$f_2(t) = A_2 e^{\lambda_1 t} + B_2 e^{\lambda_2 t} + C_2 e^{\lambda_3 t}, \tag{7.2.1}$$

for årsak 2 og

$$f_3(t) = A_3 e^{\lambda_1 t} + B_3 e^{\lambda_2 t} + C_3 e^{\lambda_3 t}, \tag{7.2.2}$$

for årsak 3. Fra (5.3.2) er

$$\begin{aligned}
A_2 &= \frac{k_0 k_1 l_2}{(b-k_1)(a-k_1)} && (7.2.3)\\
B_2 &= l_1 + \frac{k_0 k_1 l_2}{(a-b)(a-k_1)} \\
C_2 &= -\frac{k_0 k_1 l_2}{(b-k_1)(a-b)} && (7.2.4)\\
A_3 &= \frac{k_0 k_1 m_2}{(b-k_1)(a-k_1)} && (7.2.5)\\
B_3 &= m_1 + \frac{k_0 k_1 m_2}{(a-b)(a-k_1)} \\
C_3 &= -\frac{k_0 k_1 m_2}{(b-k_1)(a-b)} && (7.2.6)\\
\lambda_1 &= k_1 \\
\lambda_2 &= a = k_0 + l_1 + m_1 \\
\lambda_3 &= b = l_2 + m_2
\end{aligned}$$

I denne modellen vil det i utgangspunktet være en løsning for alle kombinasjoner av hvilke ledd som svarer til λ_1, λ_2 og λ_3, på samme måte som for modell 2. Det vil si seks mulige løsninger. Disse løsningene, dersom de eksisterer, kan finnes fra

$$
\begin{aligned}
k_0 &= \frac{(A_2 + A_3)(b - k_1)(a - k_1)}{k_1 b} \\
l_1 &= A_2 + B_2 + C_2 \\
l_2 &= \frac{bA_2}{A_2 + A_3} \\
m_1 &= A_3 + B_3 + C_3 \\
m_2 &= \frac{bA_3}{A_2 + A_3}.
\end{aligned}
$$

Parameteren k_1 er gitt siden den er en av de negative av egenverdiene til **Q**-matrisen, på samme måte som a og b.

For at en kombinasjon skal være en mulig løsning må følgende ligninger være oppfylt

$$A_2(a - k_1) = -C_2(a - b) \tag{7.2.7}$$
$$A_3(a - k_1) = -C_3(a - b) \tag{7.2.8}$$
$$a - A_2 - B_2 - C_2 - A_3 - B_3 - C_3 > 0. \tag{7.2.9}$$

Den første restriksjonen (7.2.7) kommer av at (7.2.3) og (7.2.4) gir to løsninger for l_2 og disse må være like. Tilsvarende kommer (7.2.8) av at (7.2.5) og (7.2.6) gir løsninger for m_2. Den siste restriksjonen (7.2.9) kommer av $a = k_0 + l_1 + m_1$.

Et eksempel på en slik situasjon er ved de årsaksspesifikke tetthetsfunksjonene

$$
\begin{aligned}
f_2(t) &= \frac{1}{7}e^{-2t} + \frac{13}{7}e^{-9t} + e^{-10t} \\
f_3(t) &= \frac{3}{14}e^{-2t} + \frac{23}{7}e^{-9t} + \frac{3}{2}e^{-10t}.
\end{aligned}
\tag{7.2.10}
$$

Det er i utgangspunktet seks mulige løsningskombinasjoner illustrert i tabell 7.3.

Tabell 7.3: Mulige løsningskombinasjoner for modell 3, for eksempelfunksjonene (7.2.10). Kolonnen 'Kombinasjoner' er et uttrykk for rekkefølgen av leddene i (7.2.1) og (7.2.2).

Kombinasjoner	k_1	a	b
ABC	2	9	10
ACB	2	10	9
BCA	9	10	2
BAC	9	2	10
CBA	10	9	2
CAB	10	2	9

Ved å sjekke betingelsene i (7.2.7), (7.2.8) og (7.2.9) viser det seg at det bare er kombinasjonene ABC og CBA som faktisk gir mulige løsninger. Disse er gitt i tabell 7.4.

Tabell 7.4: De to mulige løsningene for modell 3 for eksempelfunksjonene (7.2.10).

	k_0	k_1	l_1	l_2	m_1	m_2
Løsning 1	1	2	3	4	5	6
Løsning 2	1	10	3	$\frac{4}{5}$	5	$\frac{6}{5}$

Et interessant spesialtilfelle er dersom $k = k_0 = k_1$. Uttrykk for l_1, l_2, m_1 og m_2 er da gitt som:

$$\begin{aligned} l_1 &= A_2 + B_2 + C_2 \\ l_2 &= \frac{bA_2}{A_2 + A_3} \\ m_1 &= A_3 + B_3 + C_3 \\ m_2 &= \frac{bA_3}{A_2 + A_3} \end{aligned}$$

Nå er $a = k + l_1 + m_1$, der både a og k er negative egenverdier og dermed gitt for hver kombinasjon. Dette gir en ekstra restriksjon for å sikre at denne ligningen holder. Denne kan skrives som

$$a - k = l_1 + m_1 = A_2 + B_2 + C_2 + A_3 + B_3 + C_3.$$

Modifisert modell 3

Dersom overgangene l_1 og m_1 går fra tilstand 1 i stedet for 0, til henholdsvis tilstand 2 og 3, har **Q**-matrisen fortsatt tre egenverdier og årsaksspesifikke tetthetsfunksjonene er på formen

$$\begin{aligned} f_2(t) &= \left[\frac{k_0 l_1}{(a-k_0)} + \frac{k_0 k_1 l_2}{(b-k_0)(a-k_0)}\right] e^{-k_0 t} \\ &\quad + \left[-\frac{k_0 l_1}{(a-k_0)} + \frac{k_0 k_1 l_2}{(a-b)(a-k_0)}\right] e^{-at} \\ &\quad + \left[-\frac{k_0 k_1 l_2}{(a-b)(b-k_0)}\right] e^{-bt} \\ &= A_2 e^{\lambda_1 t} + B_2 e^{\lambda_2 t} + C_2 e^{\lambda_3 t}, \end{aligned}$$

der λ_1, λ_2 og λ_3 er de tre egenverdiene. Funksjonen $f_3(t)$ vil være helt tilsvarende med l_1 byttet ut med m_1 og l_2 byttet ut med m_2. I denne modellen vil det være en løsning for alle kombinasjoner av hvilke ledd som svarer til λ_1, λ_2 og λ_3, som før vil det si seks mulige løsninger.

$$\begin{aligned} k_1 &= -\frac{(C_2+C_3)(b-k_0)(a-b)}{k_0 b} \\ l_1 &= \frac{A_2(a-k_0)+C_2(a-b)}{k_0} \\ l_2 &= \frac{bC_2}{C_2+C_3} \\ m_1 &= \frac{A_3(a-k_0)+C_3(a-b)}{k_0} \\ m_2 &= \frac{bC_3}{C_2+C_3} \end{aligned}$$

For denne modellen er det alltid seks løsninger, da det ikke oppstår noen motsigelser i formlene.

For spesialtilfellet $k = k_0 = k_1$ gjelder restriksjonen $a - k = \frac{(A_2+A_3)(a-k)}{k} + \frac{(C_2+C_3)(a-b)}{k}$, som kommer av at $a = k + l_1 + m_1$.

Kapittel 8

Modelltilpasning: Pneumoni ved innleggelse

I dette kapittelet blir modellene fra kapittel 5 tilpasset datasettet om pneumoni ved innleggelse, også kalt datasett 1, beskrevet i avsnitt 3.1. Her blir modellene tilpasset uten å ta hensyn til kovariater i datasettet. Det vil si at i alle modellene er kovariatkoeffisientene $\beta_j = 0$, $j = 1, 2, 3$. Modeller med kovariater blir studert i kapittel 10.

For dette datasettet er det prøvd to ulike metoder for å tilpasse phase-type modellene. Den første metoden er slik at de med pneumoni ved innleggelse og de uten blir analysert hver for seg, altså tilpasset hver sin modell. Dette blir gjort ved å hente ut alle dataene for henholdsvis de med og uten pneumoni ved innleggelse og tilpasse en modell for hver av dem. Det blir på denne måten estimert et sett med parametere for hver av gruppene.

Den andre metoden er å la de uten pneumoni ved innleggelse og de med starte i ulike tilstander. Dette er en mer intuitiv måte å bruke modellene. De som ikke har pneumoni ved innleggelse starter i tilstand 0 og de med starter i tilstand 1. Dette betyr at tilstand 0 representerer det å ikke ha pneumoni, mens tilstand 1 representerer det å ha pneumoni. På denne måten er det også mulig å estimere alle parametere ved bare én optimering. I tillegg er det fordeler med tanke på identifiserbarhet av denne typen modeller, se kapittel 7.

Uavhengig av modell vil tilstand 2 svare til død og tilstand 3 svare til utskriving fra intensivavdelingen.

Målet med analysen i dette kapittelet er å studere betydningen av å ha pneumoni ved innleggelse på enhetsdødeligheten. Siden pneumoni er en alvorlig sykdom, er det forventet at flere pasienter dør med pneumoni enn uten. Det er likevel ikke selve studien av pneumoni som er lagt vekt på her, men det å tilpasse ulike phase-type

modeller.

For å teste de ulike modellene opp mot hverandre blir det sett på to funksjoner. Det er kumulativ årsaksspesifikk hasardrate og kumulative insidensfunksjoner, begge beskrevet i teoridelen, kapittel 4. Grunnen til dette er at de er enkle å sammenligne med ikke-parametriske estimatorer, se avsnitt 4.8.

De kumulative årsaksspesifikke hasardratene blir sammenlignet med Nelson-Aalen estimatoren som blir funnet ved hjelp av funksjonen `mvna` i R. Denne funksjonen finnes i pakken med samme navn. De kumulative insidensfunksjonene blir sammenlignet med Aalen-Johansen estimatoren funnet ved hjelp av `cuminc`. Denne funksjonen finnes i pakken `cmprsk`. Ved bruk av funksjonen `cuminc` er det også mulig å definere grupper, som for dette datasettet kan være å ha og ikke ha pneumoni. Da kan `cuminc` estimere kumulative insidensfunksjoner for hver av gruppene, slik at de to gruppene kan sammenlignes.

8.1 Pasienter med og uten pneumoni hver for seg

Som nevnt innledningsvis i dette kapittelet, blir modellene tilpasset på to ulike måter. I dette avsnittet blir datasettet delt i to grupper, de med og uten pneumoni ved innleggelse på intensivavdelingen. Alle starter i tilstand 0. Modellene blir så tilpasset hver av de to gruppene separat. Dette gir to sett med parametere, ett for de med og ett for de uten pneumoni. Modell 1, 2, 3 og 4 blir tilpasset på denne måten.

8.1.1 Modell 1

En oversikt over modell 1 er gitt i avsnitt 5.1, og tilhørende tilstandsdiagram er vist i figur 5.1. I denne modellen er det kun to ukjente parametere. Det er raten l_1 for å gå til tilstand 2, altså at pasienten dør, og raten m_1 for å gå til tilstand 3, altså at pasienten blir utskrevet fra intensivavdelingen.

For denne modellen er det mulig å finne sannsynlighetsmaksimeringsestimatoren analytisk. Ved at settet $\mathcal{U}$ deles i to, der $\mathcal{U}_2$ er settet av de usensurerte som absorberes i tilstand 2 og $\mathcal{U}_3$ er settet av de usensurerte som absorberes i tilstand 3, kan likelihoodfunksjonen fra (4.7.1) skrives som

$$\mathcal{L} = \mathcal{L}(\theta; t) = \left\{ \prod_{i \in \mathcal{U}_2} f_2(t_i; \theta) \right\} \left\{ \prod_{j \in \mathcal{U}_3} f_3(t_j; \theta) \right\} \left\{ \prod_{k \in \mathcal{S}} S(t_k; \theta) \right\}, \tag{8.1.1}$$

der $\mathcal{U}_2 \cap \mathcal{U}_3 = \mathcal{U}$, og $\mathcal{U}$ er settet av usensurerte data. Settet $\mathcal{S}$ består av sensurerte data og $\theta = [l_1, m_1]$. I tillegg er

$$\begin{aligned} f_2(t_i; \theta) &= l_1 e^{-(l_1+m_1)t_i} \\ f_3(t_j; \theta) &= m_1 e^{-(l_1+m_1)t_j} \\ S(t_k; \theta) &= e^{-(l_1+m_1)t_k}. \end{aligned}$$

Funksjonene $f_j(t; \theta)$, $j = 2, 3$, er fra (5.1.1) og $S(t; \theta) = 1 - F_2(t, \theta) - F_3(t, \theta)$ fra (5.1.3).

Ved å ta logaritmen av (8.1.1) fås log-likelihood funksjonen

$$\begin{aligned} \ln \mathcal{L} &= \sum_{i \in \mathcal{U}_2} \ln \left(l_1 e^{-t_i(l_1+m_1)} \right) + \sum_{j \in \mathcal{U}_3} \ln \left(m_1 e^{-t_j(l_1+m_1)} \right) \\ &\quad + \sum_{k \in \mathcal{S}} \ln \left(e^{-t_k(l_1+m_1)} \right) \\ &= n_2 \ln(l_1) - (l_1 + m_1) \sum_{i \in \mathcal{U}_2} t_i + n_3 \ln(m_1) - (l_1 + m_1) \sum_{j \in \mathcal{U}_3} t_j \\ &\quad - (l_1 + m_1) \sum_{k \in \mathcal{S}} t_k \\ &= n_2 \ln(l_1) + n_3 \ln(m_1) - (l_1 + m_1) \left(\sum_{i=1}^{n_2} t_i + \sum_{j=1}^{n_3} t_j + \sum_{k=1}^{n_s} t_k \right) \\ &= n_2 \ln(l_1) + n_3 \ln(m_1) - (l_1 + m_1) \sum_{i=1}^{n} t, \end{aligned}$$

der n_2 og n_3 er antall pasienter som går til henholdsvis tilstand 2 og 3, n_s er antall som blir sensurert og $n = n_2 + n_3 + n_s$.

Deriverer med hensyn på l_1 og m_1 og setter lik null:

$$\frac{d \ln \mathcal{L}}{d l_1} = \frac{n_2}{l_1} - \sum_{i=1}^{n} t = 0$$

$$\frac{d \ln \mathcal{L}}{dm_1} = \frac{n_3}{m_1} - \sum_{i=1}^{n} t = 0$$

Dette gir estimatorene

$$\hat{l}_1 = \frac{n_2}{\sum_{i=1}^{n} t} \qquad \text{og} \qquad \hat{m}_1 = \frac{n_3}{\sum_{i=1}^{n} t}.$$

For datasettet, se kapittel 3.1, deles pasientene inn i to grupper, de med og uten pneumoni. Der er antall som dør med pneumoni lik $n_{2p} = 21$, antall som blir utskrevet med pneumoni $n_{3p} = 68$, antall som dør uten pneumoni $n_{2n} = 55$ og antall som blir utskrevet uten pneumoni $n_{3n} = 589$. Videre er for de med pneumoni $\sum_{i=1}^{n} t_{ip} = 2899$, og for de uten$\sum_{i=1}^{n} t_{ip} = 7948$. Dette gir estimatene

$$\begin{aligned}
\hat{l}_{1p} &= \frac{n_{2p}}{\sum_{i=1}^{n} t_{ip}} = \frac{21}{2899} = 0.007244 \\
\hat{m}_{1p} &= \frac{n_{3p}}{\sum_{i=1}^{n} t_{ip}} = \frac{68}{2899} = 0.023456 \\
\hat{l}_{1n} &= \frac{n_{2n}}{\sum_{i=1}^{n} t_{in}} = \frac{55}{7948} = 0.006920 \\
\hat{m}_{1n} &= \frac{n_{3n}}{\sum_{i=1}^{n} t_{in}} = \frac{589}{7948} = 0.074107,
\end{aligned}$$

der l_{1p} og m_{1p} er for tilfellet med pneumoni og l_{1n} og m_{1n} er for tilfellet uten.

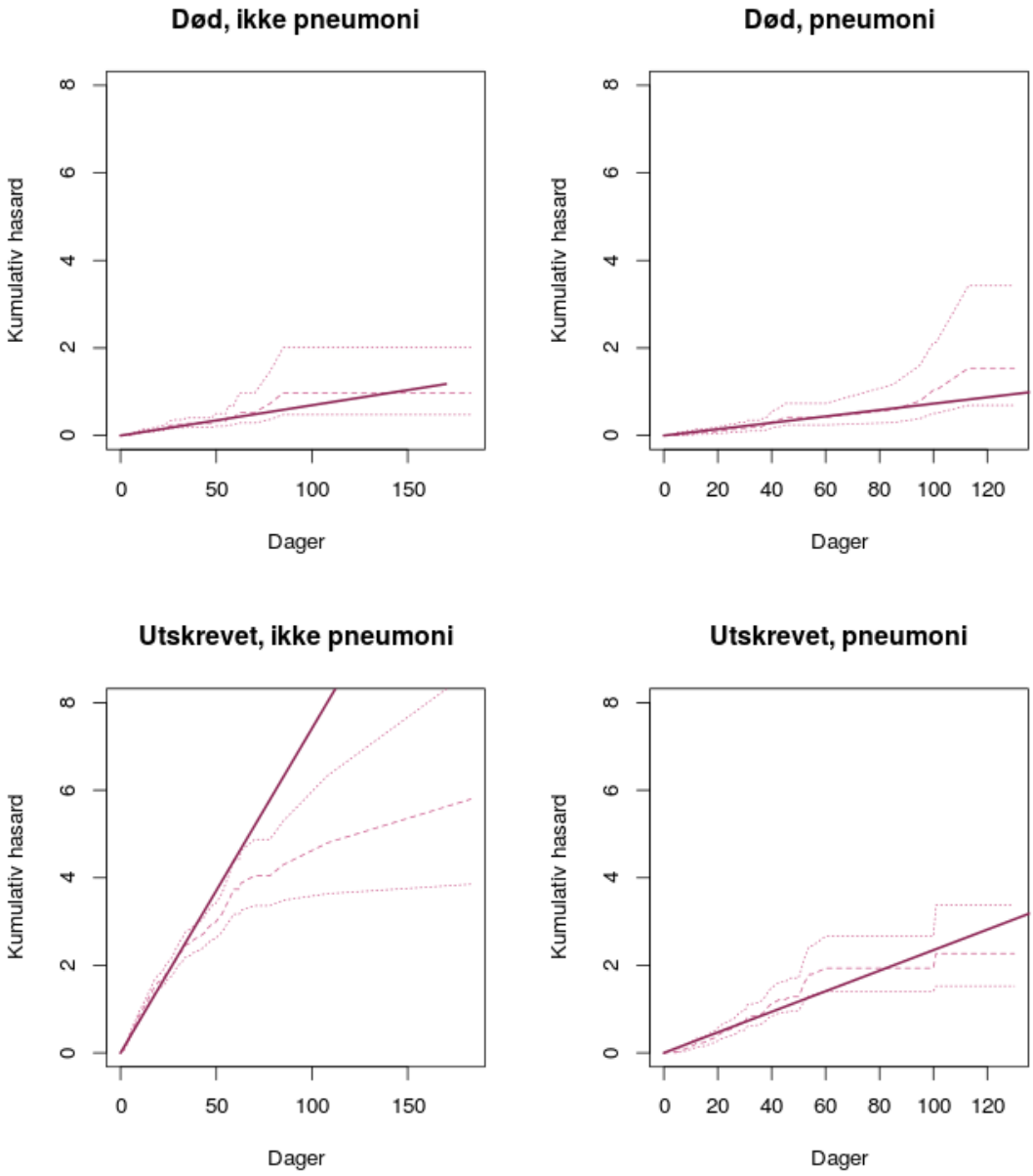
Parameterestimatene kan også finnes fra `optim`, se avsnitt 6.2. Disse er gitt i tabell 8.1, for to ulike startverdier. Øvre og nedre grense for parameterne er her satt til henholdsvis 10^{-5} og 2 for både l_1 og m_1. Det viser seg at estimatene endres veldig lite ved ulike nedre grenser, men er uendret ved ulike øvre grenser. Estimatene er ganske like for startverdi 1 og startverdi 0.1, men likevel ikke helt like, noe som kan tyde på at det er viktig å velge fornuftige startverdier. Dette er noe som kommer tydeligere frem for de øvrige modellene. I forhold til den analytiske løsningen finner `optim` rimelige verdier.

Tabell 8.1: Estimerte parametere fra modell 1, for datasett 1. Optimeringen er gjort med to ulike startverdier.

	l_1	m_1
Startverdi	1	1
Pneumoni	0.007290	0.023469
Ikke pneumoni	0.006927	0.074055
Startverdi	0.1	0.1
Pneumoni	0.007290	0.023471
Ikke Pneumoni	0.006927	0.074122

De kumulative årsaksspesifikke hasardratene fra (4.4.6), med årsaksspesifikk hasardrate som i (5.1.2) og parametere fra tabell 8.1, er vist i figur 8.1. Her er de striplede trappefunksjonene Nelson-Aalen estimatoren, og tilhørende konfidensintervall, funnet ved R-funksjonen `mvna`. Fremgangsmåten for dette er beskrevet i [Beyersmann, 2012]. De heltrukne tykke linjene er de kumulative hasardfunksjonene fra phase-type modellen. For denne enkleste modellen er disse lineære, og dermed ikke veldig bra tilpasset, men følger likevel bra i starten, for små tider.

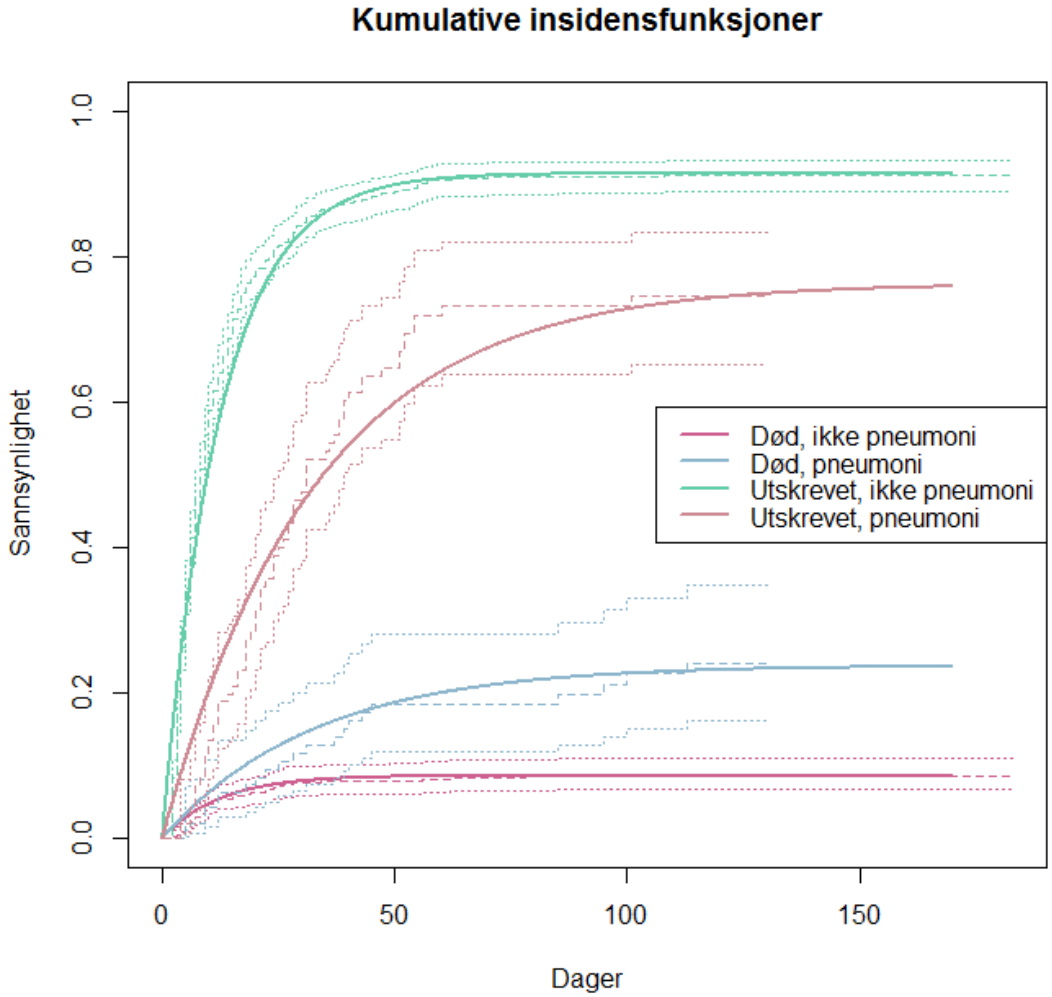
Som beskrevet i [Beyersmann, 2012] ser det ut til, fra figur 8.1, at pneumoni ikke har noen effekt på dødshasarden. Det ser ut som hasardraten for død uten pneumoni oppe til venstre er av samme type som død med pneumoni oppe til høyre. Dette betyr ikke at pneumoni ikke har noen effekt på dødeligheten. Tvert imot reduserer pneumoni utskrivelseshasarden, på de to nederste plottene. Dette er et typisk konkurrerende risikofenomen.



Figur 8.1: Kumulativ årsaksspesifikk hasardrate fra modell 1, for datasett 1. De stiplede trappefunksjonene er Nelson-Aalen estimatoren og tilhørende konfidensintervall.

Kumulative insidensfunksjoner fra (5.1.3), med parameterne fra tabell 8.1 er vist i figur 8.2. De stiplede trappefunksjonene er Aalen-Johansen estimatorene fra

`cuminc`, med tilhørende konfidensintervall. Her er det tydelig at pneumoni har en påvirkning på sannsynligheten for å dø.



Figur 8.2: Kumulative insidensfunksjoner fra modell 1, for datasettet 1. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

Fra figur 8.2 ser det ut som tilpasningen fungerer i grove trekk. Likevel er det rom for forbedring. Spesielt de kumulative insidensfunksjonene for utskrevet med pneumoni følger kurven litt dårlig. Dette kan forbedres ved å utvide modellen.

8.1.2 Modell 2

Modell 2 er en utvidelse av modell 1 med en ekstra transient tilstand, se avsnitt 5.2 figur 5.2. Dette gjør at det også blir flere parametere å estimere. I tillegg til l_1 og m_1 som er tilsvarende som i modell 1, er k intensiteten mellom de transiente tilstandene, og l_2 og m_2 er intensitetene fra den andre transiente tilstanden til henholdsvis tilstand 2 og 3.

Likelihoodfunksjonen for denne modellen er som i (8.1.1), med

$$\begin{aligned} f_2(t_i;\theta) &= \left[l_1 - \frac{kl_2}{a-b}\right] e^{-at_i} + \frac{kl_2}{a-b} e^{-bt_i} \\ f_3(t_j;\theta) &= \left[m_1 - \frac{km_2}{a-b}\right] e^{-at_j} + \frac{km_2}{a-b} e^{-bt_j} \\ S(t_k;\theta) &= \frac{l_1 + m_1 - b}{a-b} e^{-at_k} + \frac{k}{a-b} e^{-bt_k}, \end{aligned}$$

funnet fra (5.2.2) og (5.2.4).

Det at hver av funksjonene består av to ledd, gjør at det ikke er mulig å finne en eksplisitt løsning for maksimum likelihoodestimatene. Ved å maksimere likelihoodfunksjonen ved hjelp av `optim` kan parameterne estimeres. Tabell 8.2 viser parameterestimater ved ulike startverdier. Her er nedre og øvre grense for alle parameterne satt til henholdsvis 10^{-10} og 10^{10}.

Tabell 8.2: Estimerte parametere fra modell 2, for datasett 1. Optimeringen er gjort med tre ulike startverdier.

	k	l_1	l_2	m_1	m_2
Startverdi	1	1	1	1	1
Pneumoni	0.124151	0	0.009901	0	0.031970
Ikke pneumoni	0.653230	0	0.007949	0	0.084730
Startverdi	10	1	1	1	1
Pneumoni	9.999997	0	0.007314	0	0.023549
Ikke pneumoni	9.999973	0	0.007017	0	0.074712
Startverdi	0.000001	1	1	1	1
Pneumoni	1e-10	0.007289	0.930693	0.023471	0.931659
Ikke pneumoni	0.092607	0.000000	0.055883	0.000000	0.598335

Tabell 8.2 viser at `optim` finner fornuftige parametere dersom startverdien er 1 for alle parameterne. Dersom startverdien for k settes stor, for eksempel lik 10

som vist i tabellen, blir k estimert til å være tilnærmet lik startverdi. Det at den blir estimert til omtrent samme verdi som den starter på, kan tyde på at det er en svakhet i optimeringen. Ved å prøve ulike store startverdier fås samme resultat, altså at k blir tilnærmet lik startverdien. De resterende parameterne blir like i alle disse tilfellene. Faktisk er det slik at disse tilsvarer parameterne i modell 1. Det som skjer er at modellen blir forenklet til modell 1 fordi k er så stor at alle hopper til tilstand 1 så raskt at det blir forholdsvis tilnærmet uendelig raskt. Det tilsvarende ser ut til å skje dersom startverdien for k settes svært liten. Da blir k estimert til å være svært liten, slik at så godt som ingen hopper til tilstand 1. Dette gir igjen modell 1, nå med parameterne l_1 og m_1. Dette gjelder riktignok kun i tilfellet med pneumoni. I tilfellet uten pneumoni, skjer ikke dette uansett hvor liten k settes i starten.

Det er parameterne i øverste rad i tabell 8.2 som er de riktige, da disse gir størst likelihood. Det skal, som vist i kapittel 6, være to optimale løsninger, der den andre kan finnes entydig dersom man kjenner den første. Ved bruk av ligningene 7.1.3 kan den andre optimale løsningen bli funnet. De to optimale løsningene er vist i tabell 8.3.

Tabell 8.3: De to optimale parameterestimatene for modell 2. Begge settene med parametere gir samme kumulative insidensfunksjoner.

Løsning 1

	k	l_1	l_2	m_1	m_2
Pneumoni	0.124372	0	0.009904	0	0.031968
Ikke pneumoni	0.653230	0	0.007949	0	0.084730

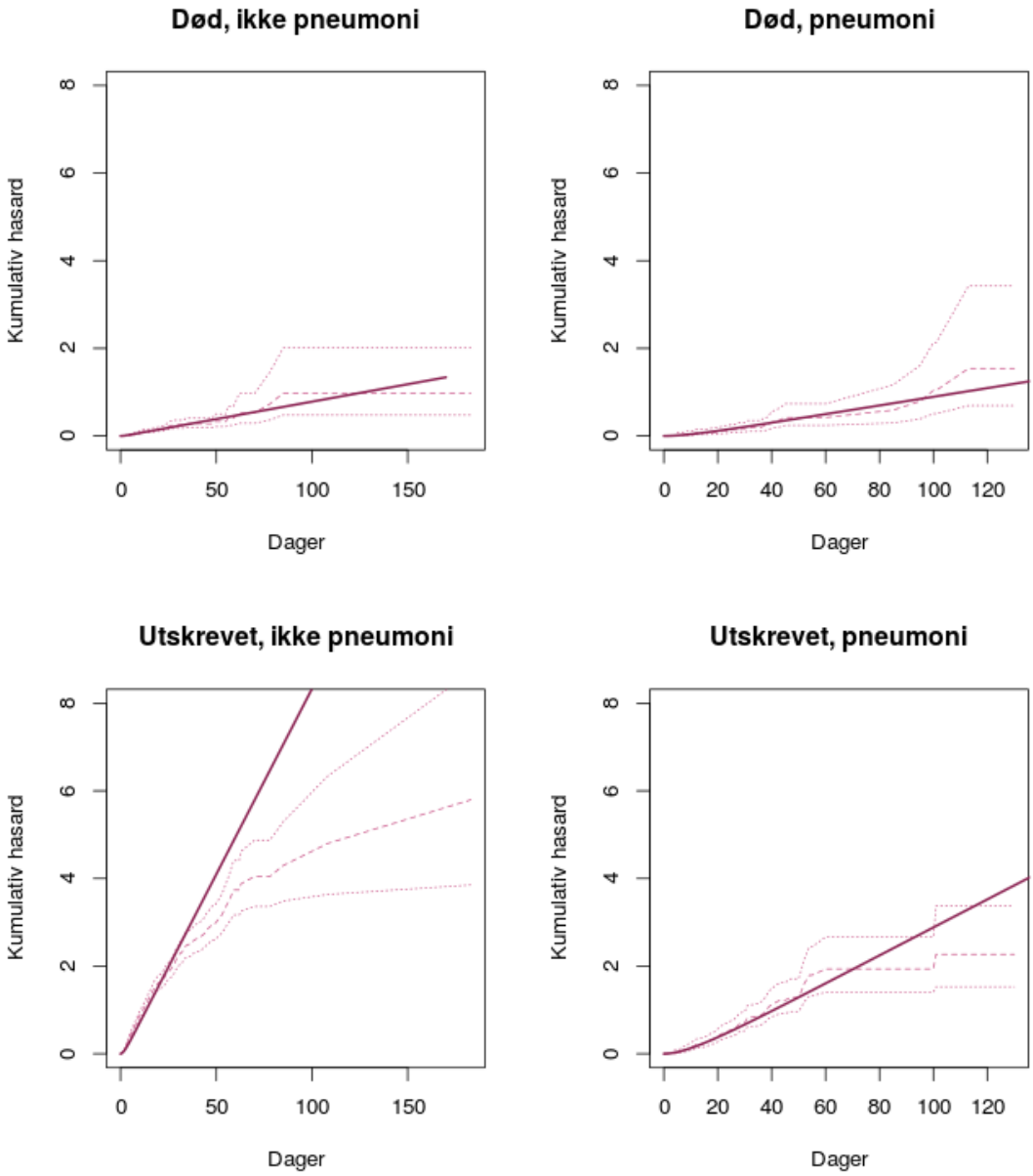
Løsning 2

	k	l_1	l_2	m_1	m_2
Pneumoni	0.041811	0	0.029389	0	0.095165
Ikke pneumoni	0.092679	0	0.056026	0	0.597205

Løsning 2 kan også finnes numerisk ved hjelp av `optim`. Dette kan gjøres på to måter, enten ved at nedre og øvre grense for parameteren k begrenses, slik at den tvinges til den andre løsningen, eller ved å endre startverdier slik at de er nærmest den andre løsningen.

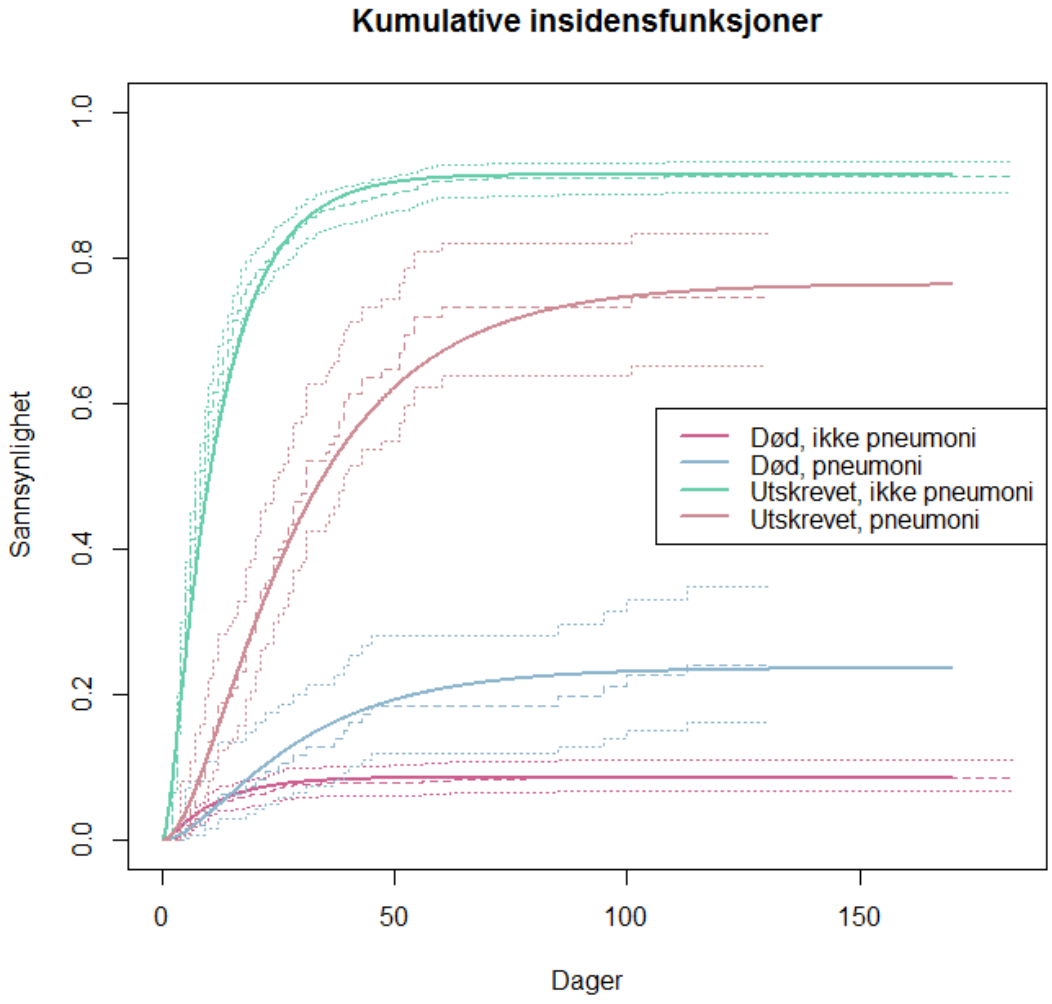
De kumulative årsaksspesifikke hasardratene (4.4.6), med årsaksspesifikke hasardrater som i (5.2.3) og parametere fra tabell 8.3, er vist i figur 8.3. Det er for denne modellen en liten knekk i starten som gjør at de følger litt bedre enn modell 1 for små tider, men spesielt utskrevet uten pneumoni er fortsatt svært dårlig etter

lengre tid. Disse funksjonene så ikke ut til å forbedres med de andre modellene og er derfor ikke plottet for dem.



Figur 8.3: Kumulativ årsaksspesifikk hasardrate for datasettet om pneumoni, for modell 2. De stiplede trappefunksjonene er Nelson-Aalen estimatoren og tilhørende konfidensintervall.

De kumulative insidensfunksjonene blir like for de to settene med parametere og er som vist i figur 8.4. For modell 1 var det utskrevet med pneumoni som fulgte kurven dårligst. Her følger den noe bedre, men fortsatt har den en del forbedringspotensial. For de andre tilfellene ser det ut til å bli tilpasset bra også for denne modellen.



Figur 8.4: Kumulative insidensfunksjoner fra modell 2, for datasettet 1. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

8.1.3 Modell 3

Denne modellen har en ekstra transient tilstand, som beskrevet i avsnitt 5.3, kalt 1'. Overgangsraten mellom tilstand 0 og 1 er nå k_0 og mellom 1 og 1' er den k_1, se figur 5.3. Denne modellen ble studert spesielt i kapittel 7 om identifiserbarhet. Det ble vist at det i utgangspunktet er seks løsninger, men at visse restriksjoner må være oppfylt for at de skal være gyldige.

Likelihoodfunksjonen er her funnet ved samme fremgangsmåte som for de andre modellene. Ved bruk av `optim` er parameterne funnet som vist i tabell 8.4, for ulike startverdier. Her er nedre grense $(10^{-5}, 10^{-5}, 0, 10^{-5}, 0, 10^{-5})$ og øvre grense 2 for alle parameterne.

Tabell 8.4: Estimerte parametere fra modell 3, for datasett 1. Optimeringen er gjort med to ulike startverdier.

	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1	0.1
Pneumoni	0.232450	0.040028	0	0.087421	0	0.283142
Ikke pneumoni	1.185734	0.093999	0	0.101280	0	1.084301
Startverdi	0.01	1	0.1	0.1	0.1	0.1
Pneumoni	1.722625	0.140143	0	0.009768	0	0.031509
Ikke pneumoni	1.177975	1.185317	0	0.008048	0	0.086046

Tabell 8.4 viser at ulike startverdier gir ulike parametere. Dette er fordi det finnes flere mulige løsninger, altså har funksjonen flere toppunkt, se kapittel 7. Hvilket som blir funnet først er avhengig av startverdiene. Det er en svært liten forskjell i likelihooden for de ulike løsningene, men dette henger sammen med unøyaktighet i `optim`, og er derfor et numerisk problem. For de uten pneumoni blir det funnet en løsning som ikke er optimal. Dette er enkelt å se dersom de kumulative insidensfunksjoner med disse parameterne blir tegnet. Da stemmer ikke funksjonene i det hele tatt. Grunnen til dette kan være at funksjonen som optimeres har lokale maksima, som blir funnet av `optim`. Det er derfor viktig å velge fornuftige startverdier, og også sjekke at parameterne faktisk gir fornuftige funksjoner ved å tegne plott. Alle de mulige løsningene funnet teoretisk fra kapittel 7 er vist i tabell 8.5, med utgangspunkt i den første løsningen.

Tabell 8.5: De seks optimale parameterestimatene for modell 3. Alle settene med parametere gir samme kumulative insidensfunksjoner.

Løsning 1

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.232450	0.040028	0	0.087421	0	0.283142
Ikke pneumoni	1.185734	0.093999	0	0.101280	0	1.084301

Løsning 2

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.370563	0.040028	0	0.054838	0	0.177611
Ikke pneumoni	1.185581	0.093999	0	0.101280	0	1.084441

Løsning 3

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.040028	0.370563	0	0.054838	0	0.177611
Ikke pneumoni	0.093855	1.185581	0	0.101280	0	1.084441

Løsning 4

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.040028	0.232450	0	0.087421	0	0.283142
Ikke pneumoni	0.093855	1.185734	0	0.101280	0	1.084301

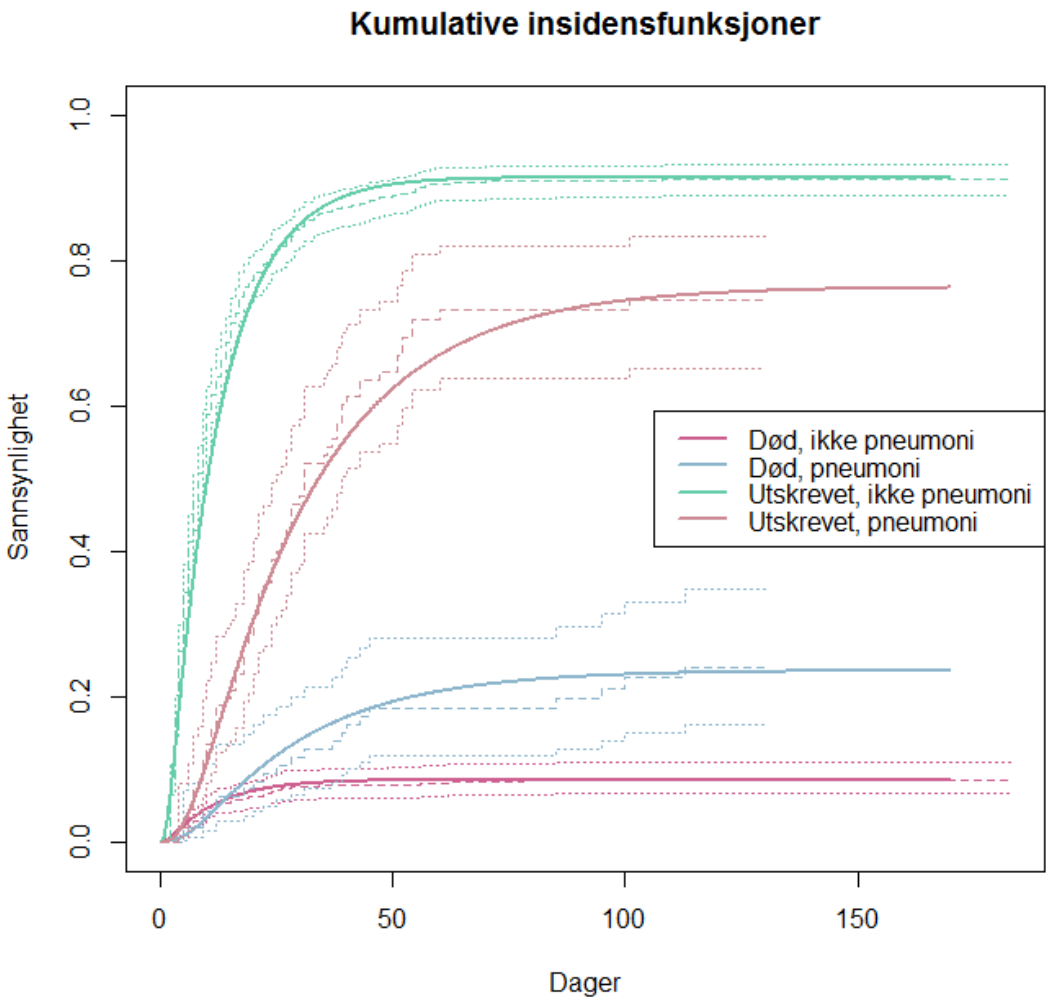
Løsning 5

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.370563	0.232450	0	0.009443	0	0.030585
Ikke pneumoni	1.183775	1.185734	0	0.008030	0	0.085969

Løsning 6

	k_0	k_1	l_1	l_2	m_1	m_2
Pneumoni	0.232449	0.370563	0	0.009443	0	0.030585
Ikke pneumoni	1.183928	1.185581	0	0.008030	0	0.085969

De kumulative insidensfunksjonene fra (4.4.2) med matriser som i avsnitt 5.3, og med parametere fra tabell 8.5 er vist i figur 8.5. Alle de seks ulike løsningene gir samme funksjoner.



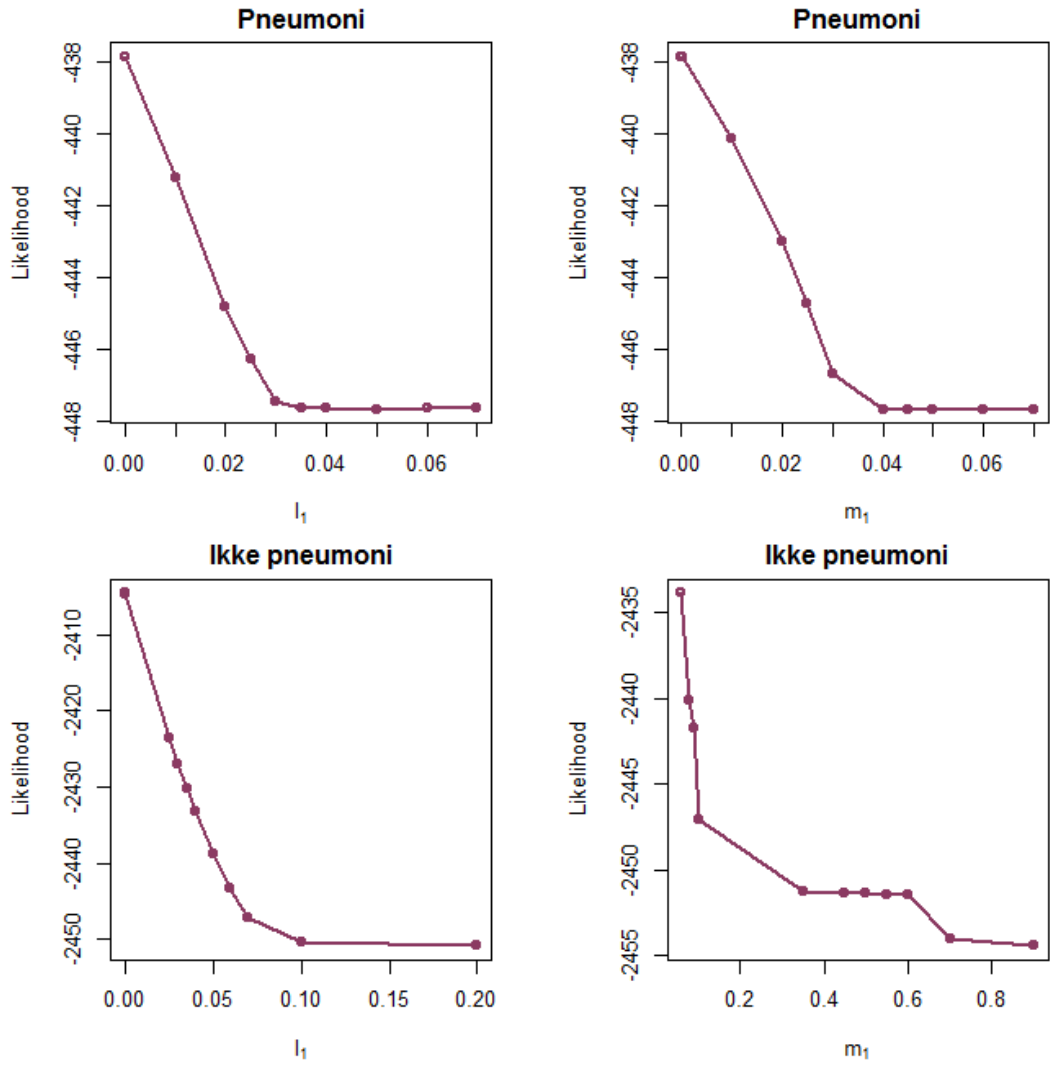
Figur 8.5: Kumulative insidensfunksjoner fra modell 3, for datasettet 1. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

Ved å sammenligne de kumulative insidensfunksjonene med de oppnådd ved bruk av modell 2, ser det faktisk ikke ut som denne modellen gir så mye forbedring. Grunnen til dette er at siden l_1 og m_1 blir null og k_0 og k_1 er ulike, blir denne modellen i praksis nærmest lik modell 2. Det er likevel mulig å se at den tilpasses litt bedre for små tider, at den følger litt bedre helt i starten.

Dersom den ekstra transiente tilstanden ble lagt før tilstand 0 i stedet, ville det

gitt samme modell. Igjen er dette på grunn av at l_1 og m_1 blir 0. I kapittel 7 ble det vist at det alltid er seks løsninger for denne modellen. Dette stemmer her.

I både modell 2 og 3 ble l_1 og m_1 estimert til å være null. Det er grunn til å tro at det er en svakhet i modellen, men ved å simulere data med disse ulike null, klarer den å finne disse. For å være sikker sjekkes dette ved profil-likelihood. Dette gjøres ved å sette en av parameterne og deretter maksimere likelihooden med hensyn på de andre parameterne, se avsnitt 4.7.1. Figur 8.6 viser profil-likelihood for l_1 og m_1 for modell 2, tilsvarende resultat kan oppnås for modell 3. For alle tilfellene er likelihooden størst i null.



Figur 8.6: Profil-likelihood med hensyn på l_1 og m_1, for tilfellet med og uten pneumoni. Punktene viser for hvilke verdier maksimum likelihooden ble funnet.

Spesialtilfeller

Over ble det vist at optimale parametere oppnås ved $l_1 = 0$ og $m_1 = 0$. For spesialtilfellet der $k_0 = k_1$ er det ikke mulig at l_1 og m_1 er null. Dette er fordi det fører til at ledd blir null i nevneren. Årsaken er at ved spesialtilfellet $k_0 = k_1$, $l_1 = 0$ og $m_1 = 0$, har $\mathbf{Q}$-matrisen en egenverdi av multiplisitet 2, og årsaksspesifikk tetthetsfunksjon for årsak 2 ville blitt

$$f_2(t) = \left[\frac{k^2 l_2}{b-k}\right] te^{-kt} + \left[-\frac{k^2 l_2}{(b-k)^2}\right] e^{-kt} + \left[\frac{k^2 l_2}{(b-k)^2}\right] e^{-bt}.$$

Her er ett av leddene multiplisert med t. Dette gir gammafordelte overgangstider mellom tilstand 0 og 1'. Tilsvarende gammafordeling oppstår for den siste modellen, modell 4.

8.1.4 Modell 4

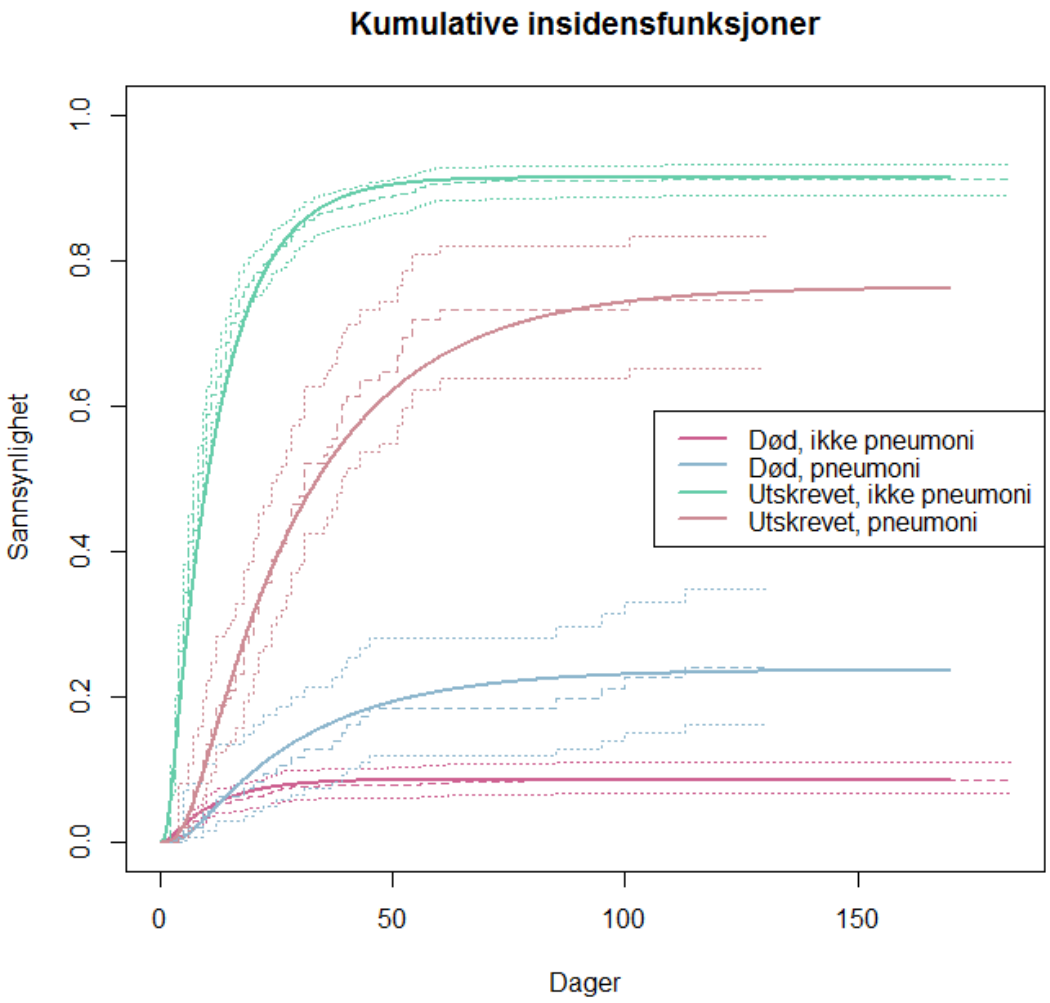
Denne modellen har to ekstra transiente tilstander, som vist i avsnitt 5.4, med tilstandsdiagram som vist i figur 5.4. Overgangsraten mellom tilstand 1 og 1' og mellom 1' og 1" er begge k_1. Dette gjør at en av egenverdiene til $\mathbf{Q}$-matrisen har multiplisitet 2 slik at overgangstidene mellom tilstand 1 og 1" blir gammafordelte.

Likelihoodfunksjonen blir funnet på samme måte som for de andre modellene. Funksjonen `optim` virker å gi noe ustabile resultater for denne modellen. Den er veldig sensitiv for ulike startparametere og ulike øvre og nedre grenser. Det kan tyde på at det finnes flere lokale maksima. Det har blitt prøvd med ulike startparametere og grenser, og de parameterne som maksimerer likelihooden er gitt i tabell 8.6. Her er nedre grense satt til $(10^{-5}, 10^{-5}, 0, 10^{-5}, 0, 10^{-5})$ for både tilfellet med og uten pneumoni. Øvre grense er satt til 2 for alle parameterne bortsett fra k for tilfellet med pneumoni, der den ble satt til 0.6 for at `optim` skulle finne optimale parametere.

Tabell 8.6: Estimerte parametere fra modell 4, for datasett 1.

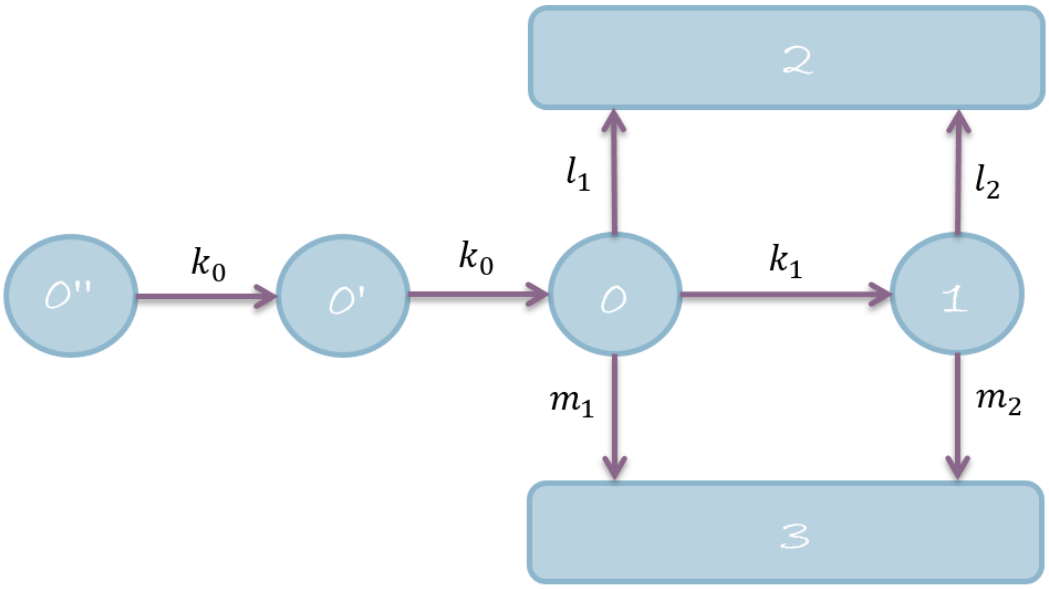
	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	1	1	0.1	1	0.1
Pneumoni	0.582787	0.450303	0	0.009153	0	0.029531
Ikke pneumoni	1.732646	1.731734	0	0.008102	0	0.086332

De kumulative insidensfunksjonene fra (4.4.2) med matriser som i avsnitt 5.4, og med parametere fra tabell 8.6 er vist i figur 8.7. Det ser ikke ut til at denne modellen gav noe forbedring i forhold til modell 3. Det kan se ut til at det ikke hjalp med gammafordelte overgangstider slik tilstandene er plassert her.



Figur 8.7: Kumulative insidensfunksjoner fra modell 4, for datasettet 1. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

I og med at det er grunn til å tro at tilpasningen skulle blitt bedre med gammafordelte overgangstider blir det testet å legge de ekstra transiente tilstandene før tilstand 0 i stedet. Dette fører til gammafordelte overgangstider før den opprinnelige modell 2. Tilstandsdiagram er vist i figur 8.8.



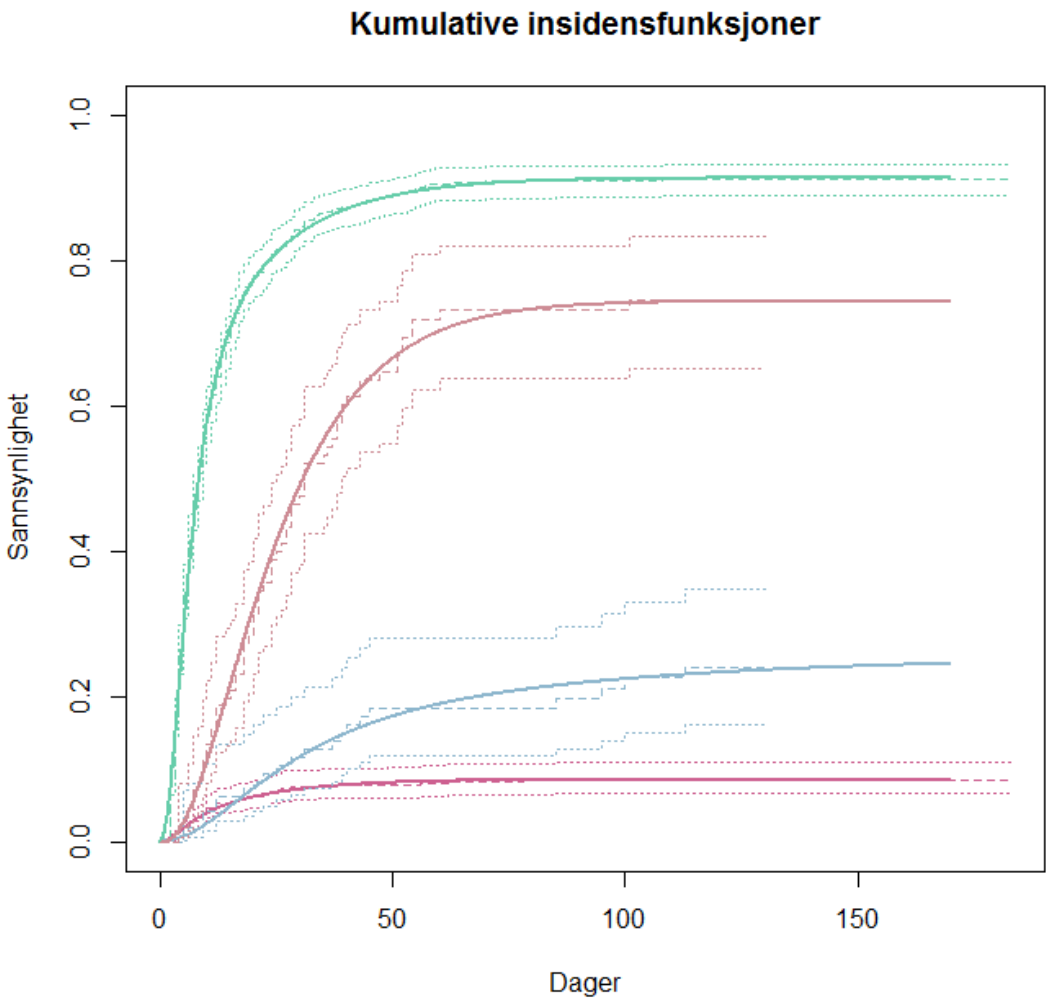
Figur 8.8: Tilstandsdiagram for en modifisert versjon av modell 4. Ratene ut fra tilstand 0" og 0' er like.

Ved å finne likelihoodfunksjonen og bruke `optim` som tidligere blir parameterne for denne modellen estimert som vist i tabell 8.7. For denne modellen blir l_1 og m_1 ulik null. Nedre grense er satt til 10^{-5} for alle parameterne. Øvre grense er satt til 2 for tilfellet med pneumoni og 5 for tilfellet uten pneumoni.

Tabell 8.7: Estimerte parametere fra modifisert versjon av modell 4, for datasett 1.

	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	1	0.1	0.1	0.1	0.1	0.1
Pneumoni	0.079642	0.064075	0.102990	0.015845	0.479752	0.000569
Ikke pneumoni	0.749985	0.072420	0.009248	0.008576	0.155579	0.047151

De kumulative insidensfunksjonene er vist i figur 8.9. For denne modellen ble de tilpasset svært bra.



Figur 8.9: Kumulative insidensfunksjoner fra modifisert modell 4 vist i figur 8.8, for datasettet 1. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

8.2 Med ulike starttilstander

En måte å bruke phase-type modellene på, som kan virke mer intuitiv enn den beskrevet over, er å la tilstandene få en betydning. Ved å definere tilstandene i modellen slik at en bestemt tilstand betyr at en pasient har pneumoni og en annen betyr at pasienten ikke har pneumoni, kan også de transiente tilstandene få en

intuitiv betydning. De absorberende tilstandene har allerede betydningen død og utskrevet.

En av forskjellene her er at parameteren $\mathbf{p}$, som gir sannsynligheten for å starte i de ulike tilstandene, blir ulik for de med og uten pneumoni. Før startet alle i samme tilstand. Modell 2, 3 og 4 blir tilpasset på denne måten. I alle disse modellene representerer tilstand 0 det å ikke ha pneumoni og tilstand 1 det å ha pneumoni. På denne måten er initialtilstanden 0 for de som ikke har pneumoni ved innleggelse og 1 for de som har det.

8.2.1 Modell 2

Det som skjer i praksis ved denne typen tilpasning er, for modell 2, at de som ikke har pneumoni ved innleggelse tilpasses modell 2, mens de som har pneumoni ved innleggelse, og dermed starter i tilstand 1, tilpasses modell 1.

For å finne likelihoodfunksjonen må alle de ulike gruppene av pasienter defineres. La settet $\mathcal{U}_{2n}$ bestå av alle pasienter som dør uten pneumoni ved innleggelse. La videre settet $\mathcal{U}_{2p}$ bestå av alle pasienter som dør med pneumoni ved innleggelse. Tilsvarende defineres de resterende gruppene som $\mathcal{U}_{3n}$, $\mathcal{U}_{3p}$, $\mathcal{S}_n$ og $\mathcal{S}_p$ som settene av henholdsvis pasienter som blir utskrevet uten og med pneumoni og blir sensurert uten og med pneumoni. Med bruk av disse definisjonene kan likelihoodfunksjonen skrives som

$$\begin{aligned}\mathcal{L}(\theta) = &\left\{\prod_{i\in\mathcal{U}_{2n}} f_{2n}(t_i;\theta)\right\}\left\{\prod_{j\in\mathcal{U}_{3n}} f_{3n}(t_j;\theta)\right\}\left\{\prod_{k\in\mathcal{S}_n} S_n(t_k;\theta)\right\}\\ &\left\{\prod_{l\in\mathcal{U}_{2p}} f_{2p}(t_l;\theta)\right\}\left\{\prod_{m\in\mathcal{U}_{3p}} f_{3p}(t_m;\theta)\right\}\left\{\prod_{n\in\mathcal{S}_p} S_p(t_n;\theta)\right\},\end{aligned} \tag{8.2.1}$$

der

$$\begin{aligned}
f_{2n}(t_i;\theta) &= \left[l_1 - \frac{kl_2}{a-b}\right]e^{-at_i} + \frac{kl_2}{a-b}e^{-bt_i} \\
f_{3n}(t_j;\theta) &= \left[m_1 - \frac{km_2}{a-b}\right]e^{-at_j} + \frac{km_2}{a-b}e^{-bt_j} \\
S_n(t_k;\theta) &= \frac{l_1+m_1-b}{a-b}e^{-at_k} + \frac{k}{a-b}e^{-bt_k} \\
f_{2p}(t_l;\theta) &= l_2e^{-bt_l} \\
f_{3p}(t_m;\theta) &= m_2e^{-bt_m} \\
S_p(t_n;\theta) &= e^{-bt_n}.
\end{aligned}$$

Her er $f_{jn}(t;\theta)$ og $f_{jp}(t;\theta)$, $j = 2, 3$, funnet fra (4.4.3) med henholdsvis $\mathbf{p} = [1, 0]$ og $\mathbf{p} = [0, 1]$. Funksjonene $S_n(t;\theta)$ og $S_p(t;\theta)$ er funnet fra (4.4.4), der $S(t) = 1-F(t)$, også her med henholdsvis $\mathbf{p} = [1, 0]$ og $\mathbf{p} = [0, 1]$. Funksjonene $f_{jn}(t;\theta)$ og $S_n(t;\theta)$ er på formen til modell 1.

Ved å maksimere likelihoodfunksjonen (8.2.1) ved hjelp av `optim`, estimeres parameterne som vist i tabell 8.8. Her er nedre grense for parameterne satt til $(10^{-5}, 0, 10^{-5}, 0, 10^{-5})$ og øvre grense er satt til 5 for alle. Tabellen viser også standardfeil for parameterne funnet fra `optim` som beskrevet i avsnitt 6.2.

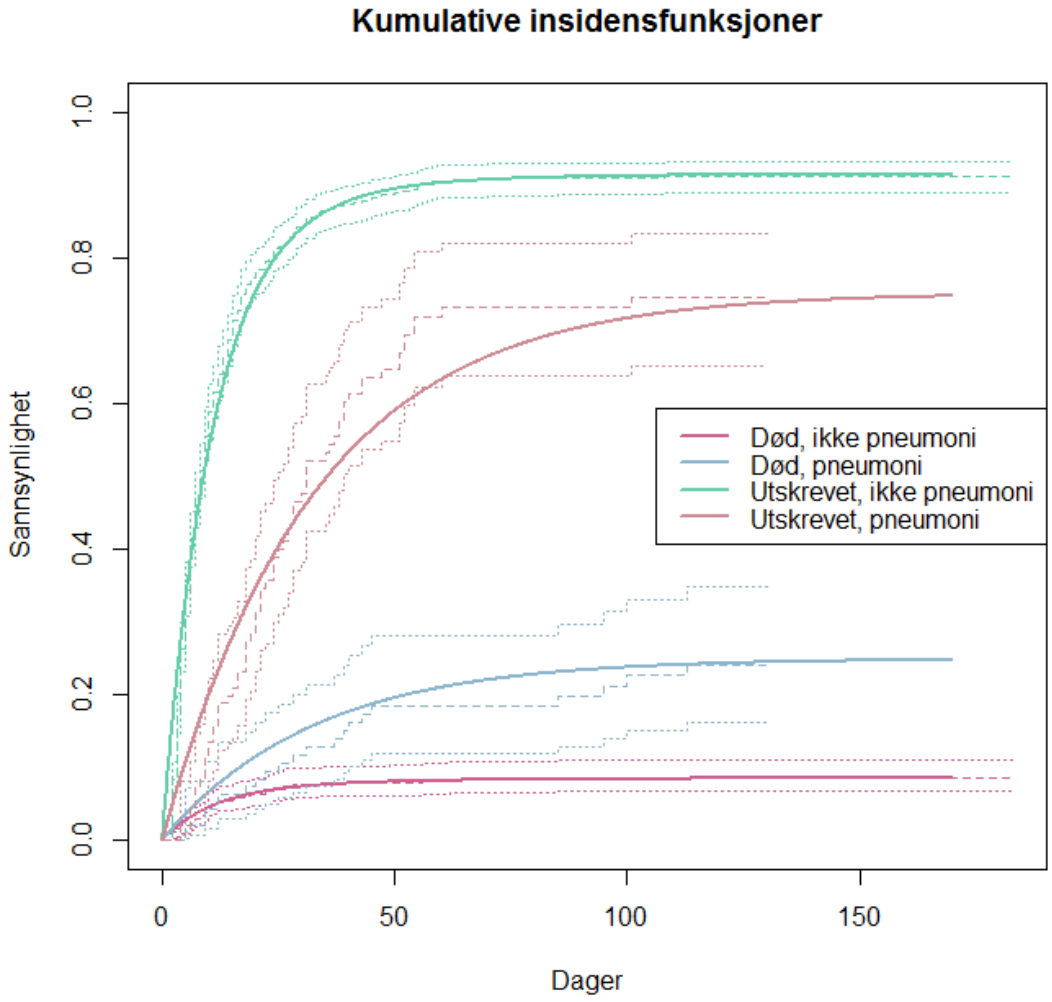
Tabell 8.8: Estimerte parametere fra modell 3 og 4 med tilhørende standardfeil, for datasett 1, i tilfellet med ulike starttilstander.

	k	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1
Estimat	0.004925	0.006682	0.007647	0.082213	0.023178
Standardfeil	0.002039	0.001043	0.001516	0.004034	0.002900

Denne modellen har kun én mulig løsning, se kapittel 7 avsnitt 7.1.1. Fra tabell 8.8 er l_2 og m_2 omtrent like som l_1 og m_1 for modell 1, med pneumoni, se tabell 8.1. Dette er på grunn av at de som starter i tilstand 1 egentlig følger en modell tilsvarende modell 1. Parameterne er likevel ikke helt like siden det i denne modellen også er noen av de som starter i tilstand 0 som hopper over til tilstand 1 og påvirker parameterne l_2 og m_2.

De kumulative insidensfunksjonene er vist i figur 8.10. For tilfellet uten pneumoni tilpasses funksjonene bra siden de følger modell 2. For tilfellet med pneumoni blir

de ikke tilpasset så bra, siden de følger modell 1. Faktisk er det mulig å se at de kumulative insidensfunksjonene for de med pneumoni er veldig like som de fra modell 1 i figur 8.2.



Figur 8.10: Kumulative insidensfunksjoner fra modell 2, for datasettet 1, i tilfellet med ulike starttilstander. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

8.2.2 Modell 3 og 4

For modell 2 ble det vist at de med pneumoni egentlig ble tilpasset modell 1, og dermed ble dårlig tilpasset. Dette kan forbedres ved å utvide med flere tilstander, som i modell 3 og 4 med henholdsvis en og to ekstra tilstander etter tilstand 0.

Likelihoodfunksjonene blir funnet som beskrevet over for modell 2, bare at det nå er **Q**-matrisen til modell 3 og 4 som blir benyttet. Startsannsynlighetene blir nå for modell 3 $\mathbf{p} = [1, 0, 0]$ for de uten pneumoni og $\mathbf{p} = [0, 1, 0]$ for de med. For modell 4 blir de $\mathbf{p} = [1, 0, 0, 0]$ for de uten pneumoni og $\mathbf{p} = [0, 1, 0, 0]$ for de med pneumoni.

Parameterne blir fra `optim` estimert til å være som vist i tabell 8.9. Her er nedre grense satt til 10^{-5} for alle parameterne for begge modellene. Øvre grense er satt til 5 for alle parameterne i modell 3 og til $(1, 0.2, 0.4, 1, 0.3, 1)$ for modell 4.

Tabell 8.9: Estimerte parametere fra modell 3 og 4 med tilhørende standardfeil, for datasett 1, i tilfellet med ulike starttilstander.

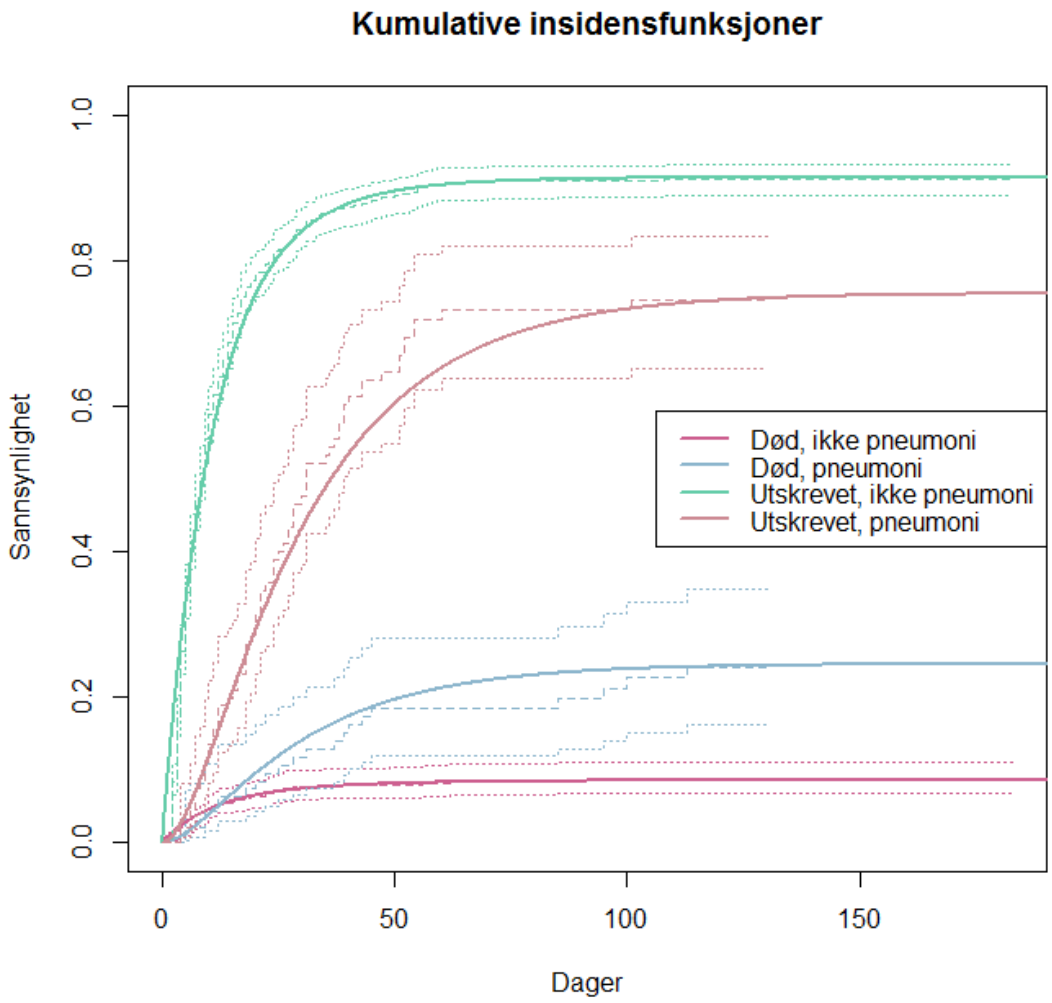
Modell 3

	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1	0.1
Estimat	0.004541	0.038829	0.006855	0.033774	0.082765	0.103709
Standardfeil	0.001917	0.005309	0.001036	0.013468	0.004140	0.037240

Modell 4

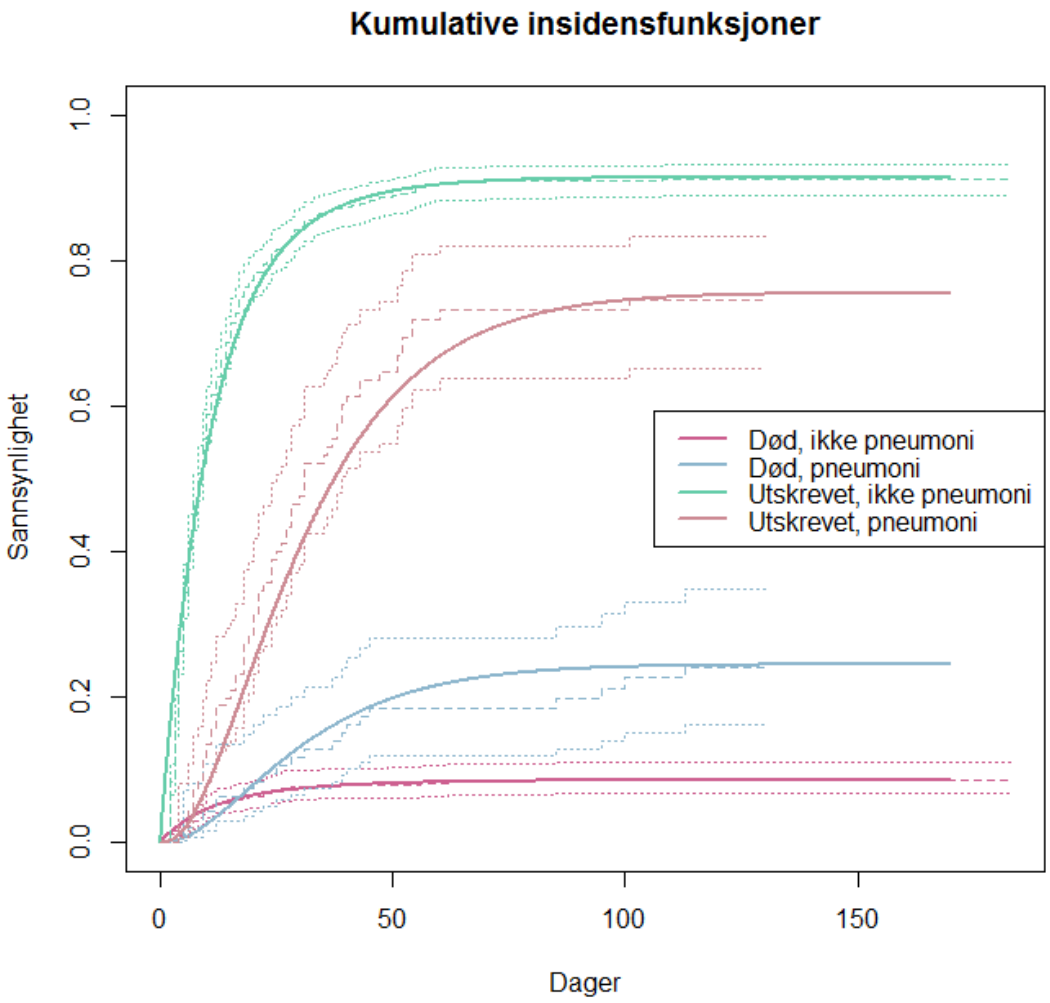
	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1	0.1
Estimat	0.004328	0.063724	0.006933	0.135531	0.082949	0.419559
Standardfeil	0.001750	0.004782	0.001032	0.066624	0.004180	0.193430

De kumulative insidensfunksjonene for modell 3 er vist i figur 8.11. De med pneumoni har blitt bedre tilpasset og de uten er fortsatt bra.



Figur 8.11: Kumulative insidensfunksjoner fra modell 3, for datasettet 1, i tilfellet med ulike starttilstander. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

De kumulative insidensfunksjonene for modell 4 er vist i figur 8.12. Det er litt vanskelig å avgjøre om dette er en bedre tilpasning enn modell 3. De uten pneumoni er tilpasset bra for alle modellene. De med pneumoni ser ut til å følge kurven litt bedre her, men på den andre siden ser det ut til at de blir estimert til å være litt for liten helt fra starten av for utskrevet med pneumoni.



Figur 8.12: Kumulative insidensfunksjoner fra modell 4, for datasettet 1, i tilfellet med ulike starttilstander. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

Kapittel 9

Modelltilpasning: Sykehuservervet pneumoni

I dette kapittelet blir modellene fra kapittel 5 tilpasset datasett 2, om sykehuservervet pneumoni, beskrevet i kapittel 3, avsnitt 3.2. Kovariater blir ikke analysert for dette datasettet, og alle kovariatkoeffisientene $\beta_j = 0$, $j = 1, 2, 3$, i modellene.

Det som er spesielt med dette datasettet sammenlignet med det analysert i kapittel 8, er at det nå ikke bare er observert tid til pasientene dør eller blir utskrevet, men også tid til de eventuelt får pneumoni. Det vil si at etter pasienter har blitt observert med pneumoni, er de fortsatt med i studien og absorberes først i død eller utskrivelse. I motsatt tilfelle er det ikke mulig å få pneumoni etter død eller utskrivelse. Denne typen data kalles semi-konkurrerende risikoer, og er beskrevet i teoridelen avsnitt 4.6.

Analysen av dette datasettet blir gjort på to ulike måter. Først blir datasettet modifisert slik at det er av samme type som datasett 1 om pneumoni ved innleggelse. Da kan sammenhengen mellom pneumoni og død analyseres. Da er det bare interessant om pasienten fikk pneumoni på et eller annet tidspunkt. Dette for å sjekke at metodene fra forrige kapittel kan brukes generelt, og ikke bare akkurat i det ene tilfellet. Det er også interessant å se om det er de samme modellene som er gode for begge datasettene, eller om det er ulikheter her.

Det andre som blir lagt vekt på er sannsynligheten for å få pneumoni. For å studere dette blir modellene modifisert til flertilstandsmodeller. Overgangssannsynligheten som funksjon av tid som oppnås da, kan sammenlignes med empirisk overgangssannsynlighet funnet ved funksjonen `etm` i R fra pakken med samme navn. Dette er nærmere vist i boken [Beyersmann, 2012].

Da det viste seg at det ikke ble noe særlig forbedring med modell 4 for datasett 1, er bare modell 2 og 3 vist her. Det er selvsagt også mulig å bruke modellene slik

at pasientene med og uten pneumoni ved innleggelse starter i ulike tilstander, se avsnitt 8.2. Dette er testet, men siden resultatene ble ganske like, er det ikke vist her.

9.1 Modifisert datasett

For å analysere sannsynligheten for å dø eller bli utskrevet med og uten pneumoni, blir datasettet om sykehuservervet pneumoni modifisert til å bli av samme type som datasettet om pneumoni ved innleggelse. Det blir da sett bort fra de faktiske tidene pasientene får pneumoni, og bare registrert om de i det hele tatt får det eller ikke. På denne måten kan modelltilpasningen foregå på akkurat samme måte som i kapittel 8. To eksempler er vist under, for modell 2 og modell 3, der de med og uten pneumoni tilpasses separat.

9.1.1 Modell 2 og 3

Modell 2 og modell 3 er som før, og vist i avsnitt 5.2 og 5.3. Likelihoodfunksjonene finnes på samme måte som vist i avsnitt 8.1.2. Estimatene blir fra `optim` som vist i tabell 9.1. Her er nedre grense satt til 0 for l_1 og m_1 i begge modellene, og 10^{-5} for de resterende. Øvre grense er satt til 5 for alle parameterne i begge modellene.

Tabell 9.1: Estimerte parametere fra modell 2 og 3, for datasett 2. Modell 2 har ikke parameteren k_1, men $k = k_0$ her.

Modell 2

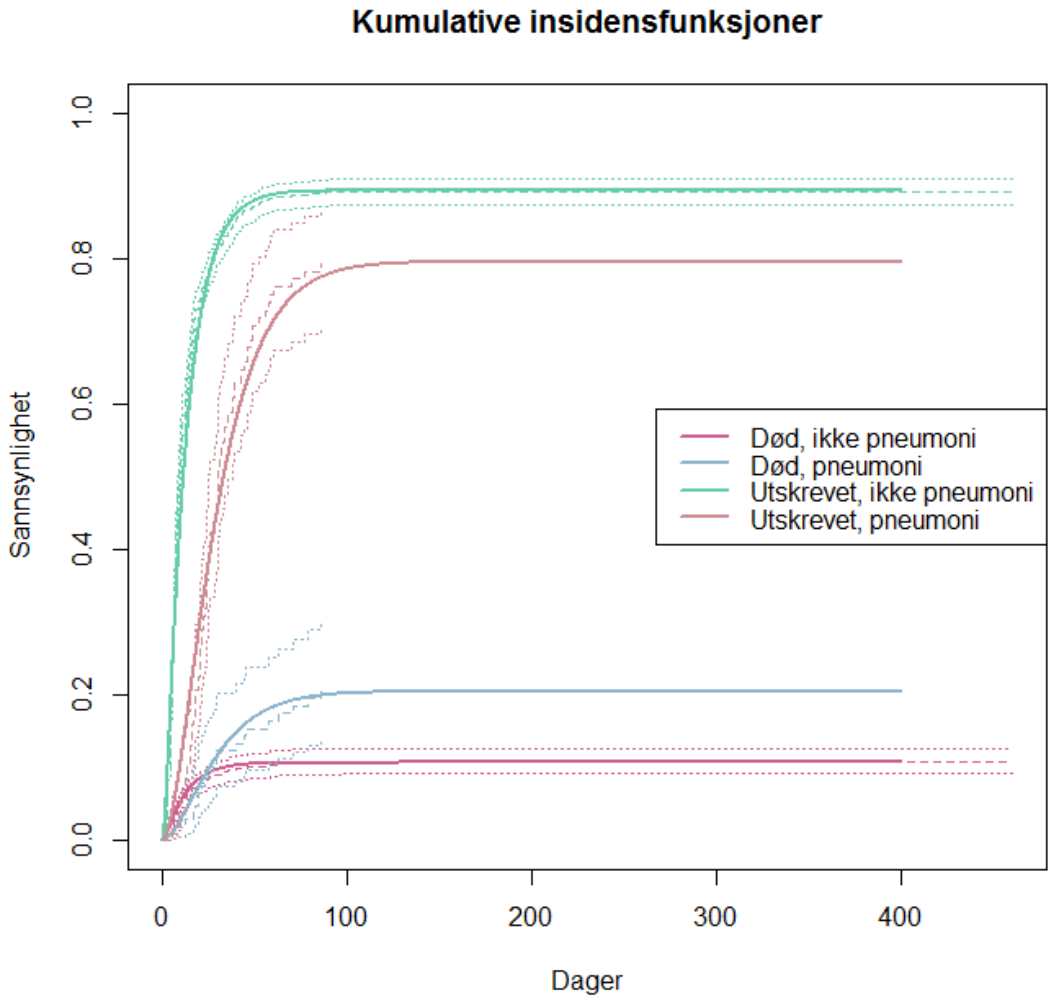
	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	-	0.1	0.1	0.1	0.1
Pneumoni	0.064052	-	0	0.013136	0	0.051220
Ikke pneumoni	0.483401	-	0	0.009458	0	0.079502

Modell 3

	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1	0.1
Pneumoni	0.127559	0.077733	0	0.019833	0	0.077357
Ikke pneumoni	0.866981	0.866972	0	0.009665	0	0.081247

De kumulative insidensfunksjonene for modell 2 er vist i figur 9.1. De ser ut til

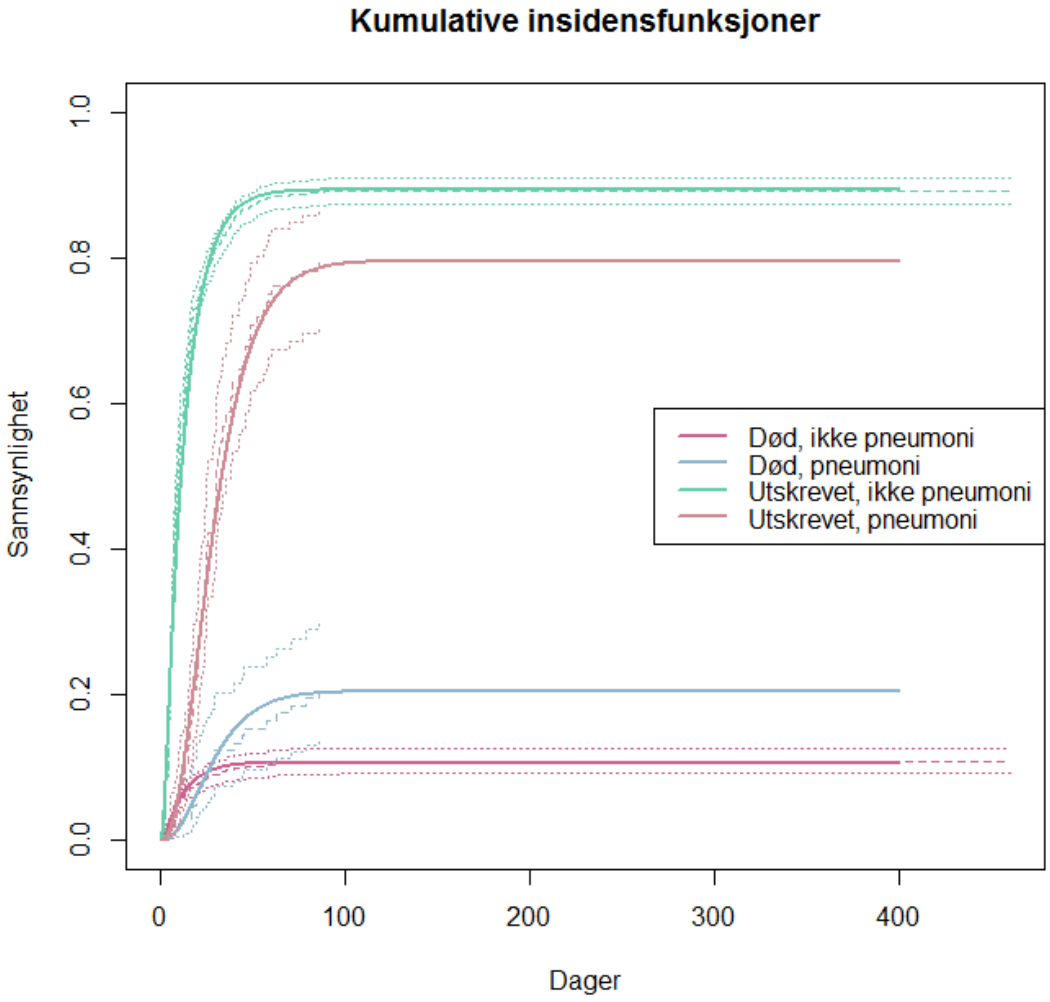
å bli ganske bra, men også for dette datasettet ser det ut til at utskrevet med pneumoni er den som tilpasses dårligst. Dette kan tyde på at det er en generell svakhet ved modellen, siden det samme har skjedd for to ulike datasett. Det kan også henge sammen med at det er færre observasjoner av pasienter med pneumoni og dermed blir også de ikke-parametriske estimatene dårligere.



Figur 9.1: Kumulative insidensfunksjoner fra modell 2, for datasettet 2. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

De kumulative insidensfunksjonene fra modell 3 er vist i figur 9.2. Her ser det ut

til at utskrevet med pneumoni er blitt svært bra tilpasset. Dette er ulik tendens som for datasett 1, der det nesten ikke var noen forskjell på de to modellene.



Figur 9.2: Kumulative insidensfunksjoner fra modell 3, for datasettet 2. De stiplede trappefunksjonene er Aalen-Johansen estimatorene med tilhørende konfidensintervall.

9.2 Semi-konkurrerende risikoer

Datasettet om sykehuservervet pneumoni inneholder informasjon om de eksakte tidene pasienter får pneumoni. I dette avsnittet blir det sett på hvordan phase-type modellene kan brukes på et semi-konkurrerende risikoproblem. Dette innebærer å bruke phase-type modellene som flertilstandsmodeller. Semi-konkurrerende risikoer er beskrevet i teoridelen 4.6. Modell 2 og modell 3 blir sett på på denne måten.

9.2.1 Modell 2

For å tilpasse modell 2 til det semi-konkurrerende risikoproblemet starter alle i tilstand 0, som betyr at man ikke har pneumoni. Det er naturlig nok ingen som har sykehuservervet pneumoni ved innleggelse. Tilstand 1 representerer de som får sykehuservervet pneumoni og ankomsttiden til denne blir observert. Det er som før mulig å bli absorbert i død eller utskrivelse både med og uten pneumoni, altså fra tilstand 0 eller 1. I tillegg er det som alltid mulig å bli sensurert. Alt i alt vil det si at pasientene har seks mulige utfall, og det er dermed seks typer data i datasettet. Disse er definert som:

Type I Absorbert direkte i tilstand 2 ved tid t.

Type II Absorbert direkte i tilstand 3 ved tid t.

Type III Går til tilstand 1 ved tid t_1 og derfra til tilstand 2 ved tid t_2.

Type IV Går til tilstand 1 ved tid t_1 og derfra til tilstand 3 ved tid t_2.

Type V Sensurert i tilstand 0

Type VI Sensurert i tilstand 1

Likelihoodfunksjonen er satt sammen av sannsynligheten for de ulike tilfellene over. Disse er gitt i (9.2.1). Funksjonene $f_{2n}(t,\theta)$, $f_{3n}(t;\theta)$ og $S_n(t;\theta)$ er som for modell 1. Disse svarer til å henholdsvis dø, bli utskrevet og sensurert uten å ha fått pneumoni. Funksjonene $f_{2p}(t,\theta)$, $f_{3p}(t;\theta)$ og $S_p(t;\theta)$ er egentlig sammensatt av to deler. Den første er sannsynligheten for å få pneumoni og den andre er sannsynligheten for å dø, bli utskrevet og sensurert etter å ha fått pneumoni.

$$\begin{aligned} P(\text{Type I}) = f_{2n}(t;\theta) &= l_1 e^{-at} \\ P(\text{Type II}) = f_{3n}(t;\theta) &= m_1 e^{-at} \\ P(\text{Type III}) = f_{2p}(t;\theta) &= k e^{-at_1} l_2 e^{-b(t_2-t_1)} \\ P(\text{Type IV}) = f_{3p}(t;\theta) &= k e^{-at_1} m_2 e^{-b(t_2-t_1)} \\ P(\text{Type V}) = S_n(t;\theta) &= e^{-at} \\ P(\text{Type IV}) = S_p(t_\theta) &= k e^{-at_1} e^{-b(t_2-t_1)} \end{aligned} \tag{9.2.1}$$

Likelihoodfunksjonen kan nå skrives som i (8.2.1). Ved å maksimere denne ved hjelp av `optim`, fås parameterne som vist i tabell 9.2. Her er nedre grense for parameterne satt til 10^{-5} og øvre til 5 for alle parameterne. Standardfeil blir funnet som beskrevet i avsnitt 6.2.

Tabell 9.2: Estimerte parametere fra modell 2 med tilhørende standardfeil, for datasett 2, i tilfellet med semi-konkurrerende risikoer.

	k	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1
Estimat	0.006411	0.007468	0.009710	0.062852	0.038066
Standardfeil	0.000601	0.000653	0.002096	0.001927	0.004201

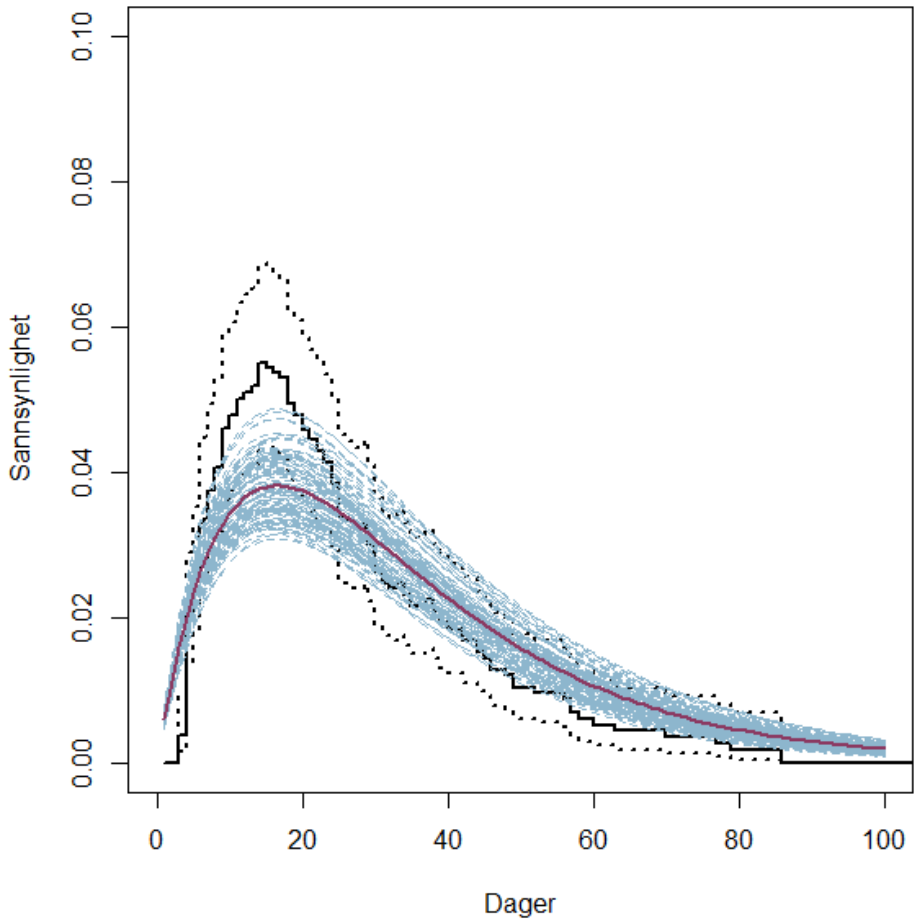
Overgangsmatrisen som funksjon av tid t finnes fra (4.3.1). Sannsynligheten for å gå fra tilstand 0 til tilstand 1 ved tid t, kalt $P_{01}(t)$ finnes ved å hente ut tilsvarende verdi fra matrisen. Denne vil være lik sannsynligheten for å være i tilstand 0 ved tid 0, gå til tilstand 1 ved tid t og ikke være absorbert enda. Denne er plottet i figur 9.3, sammen med ikke-parametrisk estimat fra R-funksjonen `etm`.

Standardfeil for hver av parameterne er ikke et godt estimat på usikkerheten i overgangssannsynligheten $P_{01}(t)$. Det brukes derfor en annen metode for å sjekke usikkerhet, kalt bootstrapping. Dette er en måte å gjøre statistisk inferens ved å resample fra originalutvalget. Målet er å oppnå et estimat på standardfeilen til en estimator, si $\hat{\theta}$. Ved å resample utvalget B ganger, kan det finnes B estimater av $\hat{\theta}$, si $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$. Da er bootstrap estimatet av variansen gitt som

$$\widehat{\text{var}}_{\text{boot}}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^*_b - \overline{\hat{\theta}^*})^2, \tag{9.2.2}$$

der $\overline{\hat{\theta}^*} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*_b$ er gjennomsnittet av de B bootstrap estimatene.

Datasettet om sykehuservervet pneumoni består av et utvalg på 1313 pasienter. Ved bootstrapping resamples disse med tilbakelegging. Uten tilbakelegging ville utvalget blitt det samme hver gang. Her er det valgt å resample $B = 100$ ganger.



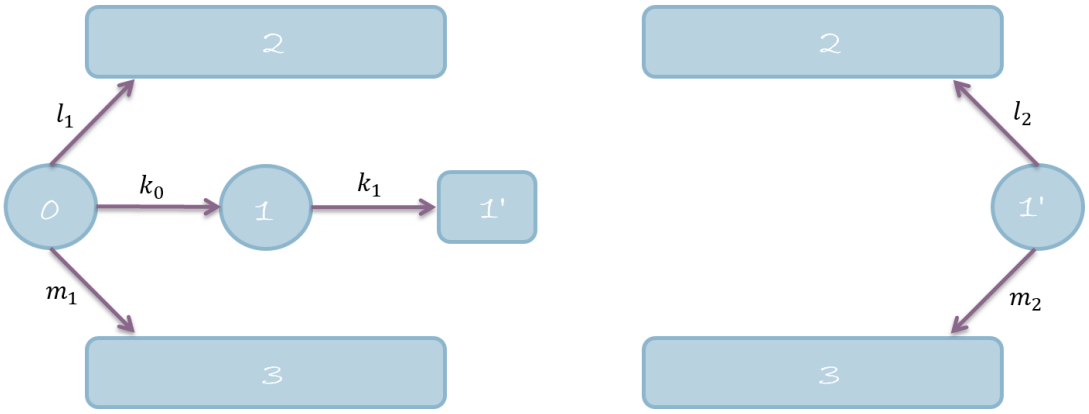
Figur 9.3: Overgangsannsynlighet $P_{01}(t)$ sammen med 100 bootstrapestimater. Den lilla kurven er snittet av disse som sammenfaller med den estimerte overgangsannsynligheten fra modell 2. Trappefunksjonene i bakgrunnen er Aalen-Johansen estimatoren fra R-funksjonen `etm`, og dens konfidensintervall.

Figur 9.3 viser 100 bootstrapestimater av $P_{01}(t)$, samt snittet av disse. Standardavvik kan finnes fra (9.2.2). Den lilla linjen er fra originalutvalget, denne sammenfal-

ler med snittet av bootstrapestimatene. Trappefunksjonene i bakgrunnen er Aalen-Johansen estimatoren for overgangssannsynligheten funnet ved R-pakken `etm` og tilhørende konfidensintervall. Phase-type modellen klarer ikke estimere toppen av kurven, men usikkerheten er av samme størrelsesorden.

9.2.2 Modell 3

Modell 3 har de samme datatypene som modell 2, men nå med noe forskjellige sannsynligheter. Disse er ikke like intuitive som for modell 2. For å finne dem deles modellen opp i to konkurrerende risikosystemer. Dette er illustrert i figur 9.4. Den første delen, figuren til venstre, er et konkurrerende risikoproblem med 3 risikoer, død, utskrivelse og å få pneumoni. De som dør eller blir utskrevet blir absorbert. De som får pneumoni blir utsatt for en ny konkurrerende risikosituasjon, til høyre, der de kan dø eller bli utskrevet.



Figur 9.4: Modell 3 delt opp i to separate konkurrerende risikoproblemer. Figuren til venstre har tre konkurrerende risikoer. Figuren til høyre er tilsvarende modell 1 med to konkurrerende risikoer.

Intensitetsmatrisene kan settes opp på samme måte som for modellene i kapittel 5, slik at sannsynlighetene for de ulike datatypene kan finnes. Disse er vist i (9.2.3).

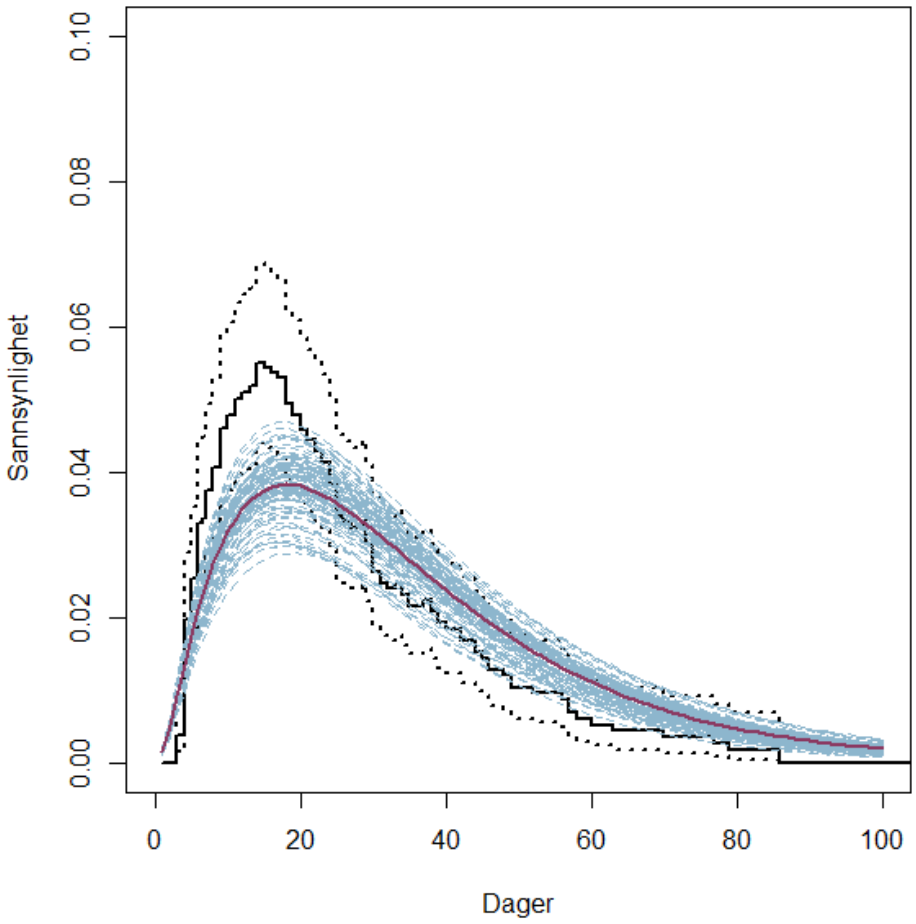
$$
\begin{aligned}
P(\text{Type I}) &= l_1 e^{-at} \\
P(\text{Type II}) &= m_1 e^{-at} \\
P(\text{Type III}) &= \frac{k_0 k_1 \left(e^{-k_1 t_1} - e^{-a t_1}\right)}{a - k1} l_2 e^{-b(t_2 - t_1)} \\
P(\text{Type IV}) &= \frac{k_0 k_1 \left(e^{-k_1 t_1} - e^{-a t_1}\right)}{a - k1} m_2 e^{-b(t_2 - t_1)} \\
P(\text{Type V}) &= e^{-at} \\
P(\text{Type IV}) &= \frac{k_0 k_1 \left(e^{-k_1 t_1} - e^{-a t_1}\right)}{a - k1} e^{-b(t_2 - t_1)}
\end{aligned}
\tag{9.2.3}
$$

Ved å sette opp likelihoodfunksjon ut fra disse og maksimere ved `optim` i R, blir parameterne estimert som vist i tabell 9.3 Her er nedre grense 10^{-5} og øvre grense 1 for alle parameterne.

Tabell 9.3: Estimerte parametere fra modell 3 med tilhørende standardfeil, for datasett 2, i tilfellet med semi-konkurrerende risikoer.

	k_0	k_1	l_1	l_2	m_1	m_2
Startverdi	0.1	0.1	0.1	0.1	0.1	0.1
Estimat	0.006478	0.721265	0.007548	0.009717	0.063266	0.037974
Standardfeil	0.000608	0.232624	0.000661	0.002098	0.001948	0.004191

Overgangssannsynlighet med bootstrapusikkerhet som for modell 2 er vist i figur 9.5. Det ser ikke ut til at modell 3 gir noe bedre resultat enn modell 2.



Figur 9.5: Overgangsannsynlighet $P_{01}(t)$ sammen med 100 bootstrapestimater. Den lilla kurven er snittet av disse som sammenfaller med den estimerte overgangsannsynligheten fra modell 3. Trappefunksjonene i bakgrunnen er fra R-funksjonen `etm`, og dens konfidensintervall.

Kapittel 10

Kovariater i phase-type modellene

Kovariater kan tas med i phase-type modellene som vist i kapittel 5. Som beskrevet der, er det gjort for modell 1 og modell 2. Det er derfor disse modellene som blir studert i dette kapittelet. For å teste disse modellene brukes datasett 1 om pneumoni ved innleggelse på intensivavdelingen. Datasettet inneholder informasjon om alder og kjønn på pasienter. Det som er interessant med kovariater generelt er om disse påvirker resultatene på noen måte. Med andre ord er det mulig å finne ut om det er forskjeller mellom kjønn eller mellom ulike aldersgrupper.

I datasettet kan man også tenke seg tilstedeværelse av pneumoni ved innleggelse som en kovariat. Dette er ikke studert her, da dette er lagt vekt på i [Beyersmann, 2012].

Igjen er det greit å merke seg at fokuset ligger i hvorvidt phase-type modellene klarer å fange opp forskjellene via kovariater, og sammenligne med andre metoder. Sammenligningen blir gjort på tre ulike måter. Først blir de estimerte kovariatkoeffiseintene sammenlignet med de man får ved å bruke Cox-regresjon. Etterpå sammenlignes resultatene ved bruk av kovariater i modellene med resultatene ved å kjøre optimering for de ulike gruppene separat. Til slutt blir de kumulative insidensfunksjonene fra modellene med kovariater sammenlignet med de som fås fra den innebygde funksjonen `cuminc` i R.

10.1 Sammenligne med Cox-regresjon

En standard metode for å modellere konkurrerende risikoer er ved proporsjonale hasardrater. Dette ble først gjort av [Cox, 1972]. Tanken bak dette er å først finne en underliggende hasardrate uten kovariater kalt $\lambda_0(t)$. Hasardraten kan videre skrives som et multiplum av den gjennomsnittlige hasardraten

$$\lambda(t) = \psi(t)\lambda_0(t).$$

Dette kan skrives om til et forhold mellom gjennomsnittlig hasard og hasard for en bestemt enhet

$$\psi(t) = \frac{\lambda(t)}{\lambda_0(t)},$$

slik at $\psi(t)$ er konstant med hensyn på t dersom $\lambda(t)$ og $\lambda_0(t)$ er proporsjonale. Det betyr at hasardraten for en bestemt enhet, på et hvilket som helst tidspunkt, vil være $\psi(t)$ ganger hasarden for en gjennomsnittlig enhet. Videre kan $\psi(t)$ avhenge av kovariatene x. Siden det ofte er brukt logaritmer i forbindelse med hasardrater blir det satt at $\psi(x) = e^{\beta x}$. Da er de årsaksspesifikke hasardratene gitt ved

$$\lambda_j(t; x) = \lambda_{0j}(t)e^{\beta_j x},$$

der j svarer til de ulike konkurrerende risikoene.

For å finne kovariatkoeffisientene fra Cox-regresjon brukes den innebygde funksjonen `coxph` i R. Dette resultatet blir sammenlignet med resultatene fra modell 1 og 2. Estimatene fra modell 1 og 2 er funnet ved å definere likelihoodfunksjonene som i (8.1.1), bare med β_2 og β_3 som ukjente parametere. I modell 2 er $\beta_1 = 0$ for å kunne sammenligne med Cox-regresjon.

Resultat for kovariaten kjønn er vist i tabell 10.1. Her er β-verdiene for de uten pneumoni ganske bra estimerte, mens de for pneumoni ikke er så bra. Dette kan henge sammen med at det er mange flere pasienter uten pneumoni, og dermed mye mer data. Likevel er alle innenfor konfidensintervallet fra Cox-regresjonen. Modell 2 har to løsninger, den gitt i tabell 10.1 er den ene som har tilsvarende β'er som for modell 1 og Cox-regresjonen.

Tabell 10.1: Sammenligning av kovariatkoeffisientene estimert fra modell 1 og 2 med de fra Cox-regresjon, for kovariaten kjønn.

Cox-modell

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	0.03557	0.47383	0.07026	0.25304
Ikke pneumoni	-0.5077	0.2728	-0.12249	0.08422

Phase-type modell 1

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	-0.00054	0.44281	0.01203	0.253608
Ikke pneumoni	-0.49018	0.26643	-0.16562	0.083679

Phase-type modell 2

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	0.042	0.505	-0.0369	0.302
Ikke pneumoni	-0.518	0.1738	-0.196	0.0995

Kovariaten alder er en kontinuerlig variabel. Resultatet for denne er gitt i tabell 10.2. Også her er tendensen at phase-type modellene gir likest resultat som Cox-regresjon for de uten pneumoni. Likevel er de med pneumoni også ganske likt estimert her.

Tabell 10.2: Sammenligning av kovariatkoeffisientene estimert fra modell 1 og 2 med de fra Cox-regresjon, for kovariaten alder.

Cox-regresjon

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	0.00834	0.01502	-0.006017	0.007363
Ikke pneumoni	0.016572	0.008765	-0.0025	0.002281

Phase-type modell 1

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	0.005255	0.0100	-0.003601	0.00711
Ikke pneumoni	0.010122	0.00474	-0.003106322	0.002245

Phase-type modell 2

	β_2	σ_{β_2}	β_3	σ_{β_3}
Pneumoni	0.005234	0.009061	-0.005471	0.009552
Ikke pneumoni	0.011018	0.005087	-0.002192	0.002368

10.2 Sammenligne med å ta grupper hver for seg

Dersom kovariaten er diskret er det mulig å kjøre tilpasningen for hver av kovariatens mulige verdier separat. Kovariaten kjønn er diskret, og kan være 'M' for mann eller 'F' for kvinne. På samme måte som de med og uten pneumoni ble tatt hver for seg tidligere, kan nå mann og kvinne bli tilpasset hver sin modell også. Det er da mulig å se forskjellen mellom dem. Dette er det samme som modeller med kovariater skal kunne fange opp. Dermed kan disse sammenlignes, og gi en indikasjon på hvor godt modellene med kovariater klarer å fange opp forskjellene.

10.2.1 Modell 1

Modell 1 har som tidligere to parametere, l_1 og m_1. Dersom det tilpasses en modell for menn og en for kvinner kan parameterne kalles l_{1M}, m_{1M}, l_{1F} og m_{1F}. Dersom kovariaten x_1 gjelder fom menn og x_2 for kvinner, vil de tilsvarende parameterne ved bruk av kovariater i modellen være $l_1 e^{\beta_2 x_1}$, $m_1 e^{\beta_3 x_1}$, $l_1 e^{\beta_2 x_2}$ og $m_1 e^{\beta_3 x_2}$.

Tabell 10.3: Sammenligning av parameterne for modell 1 med kovariaten kjønn og det å ta de to gruppene hver for seg. Parameterene på hver rad tilsvarer hverandre. For hver av modellene er maksimum log-likelihood gitt i nederste rad, kalt 'loglik'.

Pneumoni

Hver for seg		Kovariat	
l_{1M}	0.007488	$l_1 e^{\beta_2 x_1}$	0.007309
m_{1M}	0.023406	$m_1 e^{\beta_3 x_1}$	0.023549
l_{1F}	0.006925	$l_1 e^{\beta_2 x_2}$	0.007303
m_{1F}	0.023590	$m_1 e^{\beta_3 x_2}$	0.023324
loglik	-447.6425	loglik	-447.6607

Ikke pneumoni

Hver for seg		Kovariat	
l_{1M}	0.005673	$l_1 e^{\beta_2 x_1}$	0.005620
m_{1M}	0.069505	$m_1 e^{\beta_3 x_1}$	0.069452
l_{1F}	0.009140	$l_1 e^{\beta_2 x_2}$	0.009177
m_{1F}	0.081933	$m_1 e^{\beta_3 x_2}$	0.081939
loglik	-2446.753	loglik	-2446.753

Tabell 10.3 viser de ulike parameterestimatene både for de med pneumoni og de uten. For de uten pneumoni er de korresponderende parameterne svært likt estimert. For de med pneumoni er det en litt større forskjell, noe som mest sannsynlig skyldes mindre datagrunnlag.

10.2.2 Modell 2

Modell 2 har fem ukjente parametere som tidligere. I tillegg til tilsvarende inndeling som for modell 1, er det for modell 2 mulig å lage ulike kombinasjoner av hvor kovariatene plasseres. Det kan være kovariater på bare overgangene til de absorberende tilstandene, eller bare mellom de transiente tilstandene, eller på alle overgangene. Alle tre tilfellene er vist i tabell 10.4.

Tabell 10.4: Sammenligning av parameterne for modell 2 med kovariaten kjønn og det å ta de to gruppene hver for seg. Parameterene på hver rad tilsvarer hverandre. 'Kovariat abs' betyr at det er kovariater bare på overgangene til de absorberende tilstandene. 'Kovariat alle' betyr at det er kovariater på alle overgangene og 'Kovariat trans' betyr at det kun er kovariater på overgangen mellom de transiente tilstandene. 'Hver for seg' betyr at parameterne blir tilpasset hver sin modell for med og uten pneumoni. For hver av modellene er maksimum log-likelihood gitt i nederste rad, kalt 'loglik'.

Pneumoni

Hver for seg		**Kovariat abs**		**Kovariat alle**		**Kovariat trans**	
k_M	0.1870	k	0.123715	$ke^{\beta_1 x_1}$	0.186002	$ke^{\beta_1 x_1}$	0.159406
l_{1M}	0	$l_1 e^{\beta_2 x_1}$	0	$l_1 e^{\beta_2 x_1}$	0	l_1	0
l_{2M}	0.0091	$l_2 e^{\beta_2 x_1}$	0.010056	$l_2 e^{\beta_2 x_1}$	0.009104	l_2	0.009733
m_{1M}	0	$m_1 e^{\beta_3 x_1}$	0	$m_1 e^{\beta_3 x_1}$	0	m_1	0
m_{2M}	0.0285	$m_2 e^{\beta_3 x_1}$	0.031627	$m_2 e^{\beta_3 x_1}$	0.028407	m_2	0.031505
k_F	0.0622	k	0.123715	$ke^{\beta_1 x_2}$	0.060934	$ke^{\beta_1 x_2}$	0.095387
l_{1F}	0	$l_1 e^{\beta_2 x_2}$	0	$l_1 e^{\beta_2 x_2}$	0	l_1	0
l_{2F}	0.0141	$l_2 e^{\beta_2 x_2}$	0.009668	$l_2 e^{\beta_2 x_2}$	0.014370	l_2	0.009733
m_{1F}	0	$m_1 e^{\beta_3 x_2}$	0	$m_1 e^{\beta_3 x_2}$	0	m_1	0
m_{2F}	0.0483	$m_2 e^{\beta_3 x_2}$	0.032785	$m_2 e^{\beta_3 x_2}$	0.049734	m_2	0.031505
loglik	-436.5852	loglik	-437.8434	loglik	-436.5876	loglik	-437.3097

Ikke pneumoni

Hver for seg		**Kovariat abs**		**Kovariat alle**		**Kovariat trans**	
k_M	0.6752	k	0.6439	$ke^{\beta_1 x_1}$	0.679641	$ke^{\beta_1 x_1}$	0.68503
l_{1M}	0	$l_1 e^{\beta_2 x_1}$	0	$l_1 e^{\beta_2 x_1}$	0	l_1	0
l_{2M}	0.0064	$l_2 e^{\beta_2 x_1}$	0.00655	$l_2 e^{\beta_2 x_1}$	0.006324	l_2	0.0079
m_{1M}	0	$m_1 e^{\beta_3 x_1}$	0	$m_1 e^{\beta_3 x_1}$	0	m_1	0
m_{2M}	0.07830	$m_2 e^{\beta_3 x_1}$	0.07866	$m_2 e^{\beta_3 x_1}$	0.078136	m_2	0.085
k_K	0.6005	k	0.6439	$ke^{\beta_1 x_2}$	0.596003	$ke^{\beta_1 x_2}$	0.62399
l_{1K}	0	$l_1 e^{\beta_2 x_2}$	0	$l_1 e^{\beta_2 x_2}$	0	l_1	0
l_{2K}	0.0108	$l_2 e^{\beta_2 x_2}$	0.011	$l_2 e^{\beta_2 x_2}$	0.010816	l_2	0.0079
m_{1K}	0	$m_1 e^{\beta_3 x_2}$	0	$m_1 e^{\beta_3 x_2}$	0	m_1	0
m_{2K}	0.0967	$m_2 e^{\beta_3 x_2}$	0.0956	$m_2 e^{\beta_3 x_2}$	0.096775	m_2	0.085
loglik	-2400.0866	loglik	-2400.188	loglik	-2400.087	loglik	-2404.907

Likelihood ratio test

Ved å studere tabell 10.4 ser det ut til at alle de ulike versjonene av kovariater klarer å tilpasse seg bra, men at det er den med flest kovariater som er best. Spørsmålet er om den er signifikant best. For å finne ut dette brukes likelihood ratio testen. Definer de to hypotesene

$$H_0: \ \beta_j = 0 \quad \text{og} \quad H_1: \ \beta_j \neq 0.$$

Hypotesene trenger ikke dreie seg om bare én parameter β_j som er satt lik null, det kan være flere. Teststatistikken defineres som

$$D = 2(L(\theta_0|t,x) - L(\theta_1|t,x)),$$

der θ_0 er en vektor av de ukjente parameterne i null-hypotesen og θ_1 er en vektor av de ukjente parameterne i den alternative hypotesen.

Modellen med flest parametere vil alltid passe minst like bra som den med mindre. Spørsmålet er om den er signifikant bedre. For å finne ut det trengs p-verdien til differansen D. Der null-hypotesen representerer et spesialtilfelle av den alternative hypotesen, er test-statistikken kji-kvadrat fordelt med $df = df_1 - df_0$ frihetsgrader, der df_1 er antall frihetsgrader for funksjonen $L(\theta_1|t,x)$ og df_0 er antall frihetsgrader for funksjonen $L(\theta_0|t,x)$. Antall frihetsgrader er her det samme som antall ukjente parametere.

Tabell 10.5 viser resultat av likelihood ratio testen for modellene med ulike kombinasjoner av hvor kovariatene sitter, for tilfellene med og uten pneumoni ved innleggelse. For tilfellet med pneumoni er modellen med kovariater på alle overgangene ikke signifikant bedre enn de to andre. For tilfellet uten pneumoni ved innleggelse er modellen med alle parameterne ikke signifikant bedre enn modellen der $\beta_1 = 0$. Derimot er modellen med alle parameterne signifikant bedre enn modellen der både $\beta_2 = 0$ og $\beta_3 = 0$.

Tabell 10.5: Likelihood ratio test for modellene med kovariater på overganger til absorberende, transiente og alle tilstander, for pasientene med og uten pneumoni ved innleggelse. H_0 er null-hypotesen, H_1 er den alternative hypotesen, 'loglik H_0' og 'loglik H_1' er sannsynlighetsmaksimeringsestimatene for null- og alternativhypotesen. 'χ^2_{df} 95%' er kji-kvadratfordeling med *df* frihetsgrader og signifikansnivå 95%.

Pneumoni

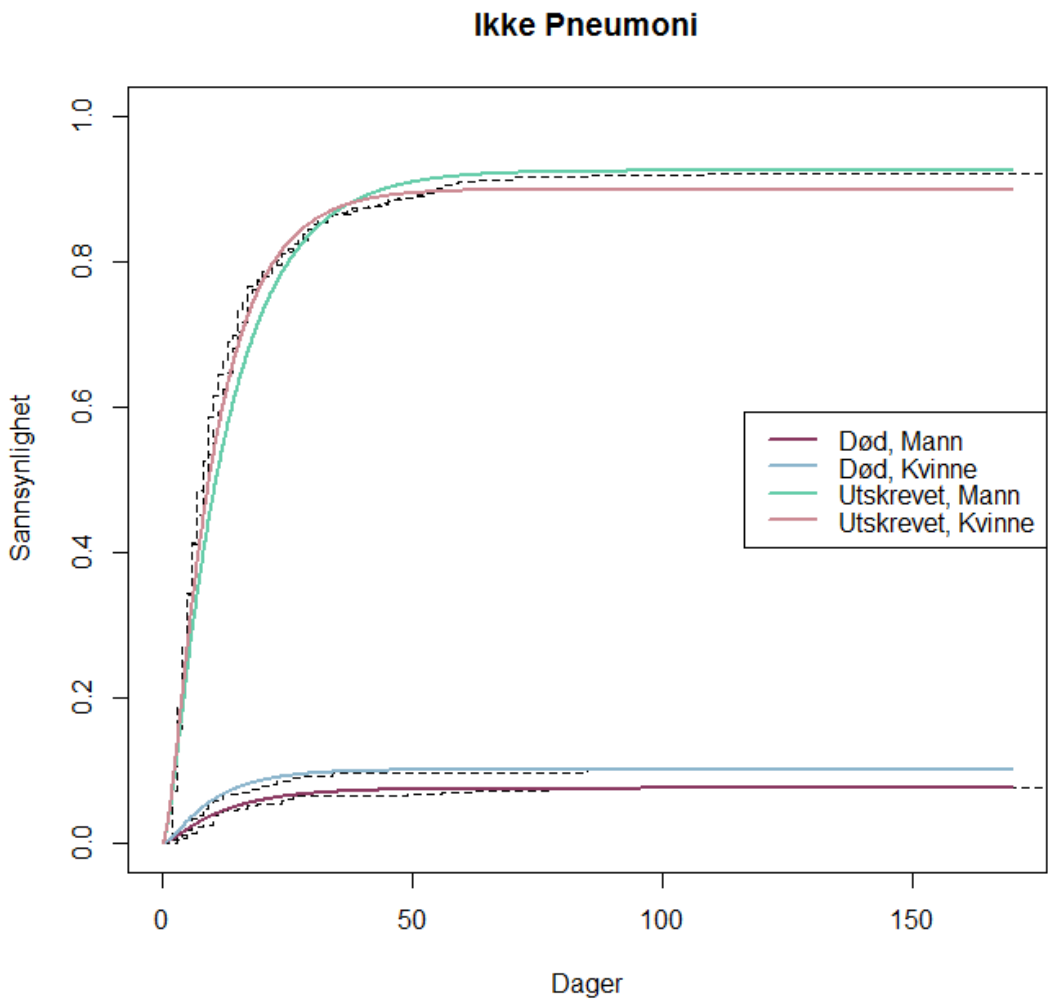
H_0	H_1	loglik H_0	loglik H_1	D	*df*	χ^2_{df} 95%
$\beta_1 = 0$	$\beta_1 \neq 0$	-437.8434	-436.5876	2.517	1	3.841
$\beta_2, \beta_3 = 0$	$\beta_2, \beta_3 \neq 0$	-437.3097	-436.5876	1.444	2	5.991

Ikke pneumoni

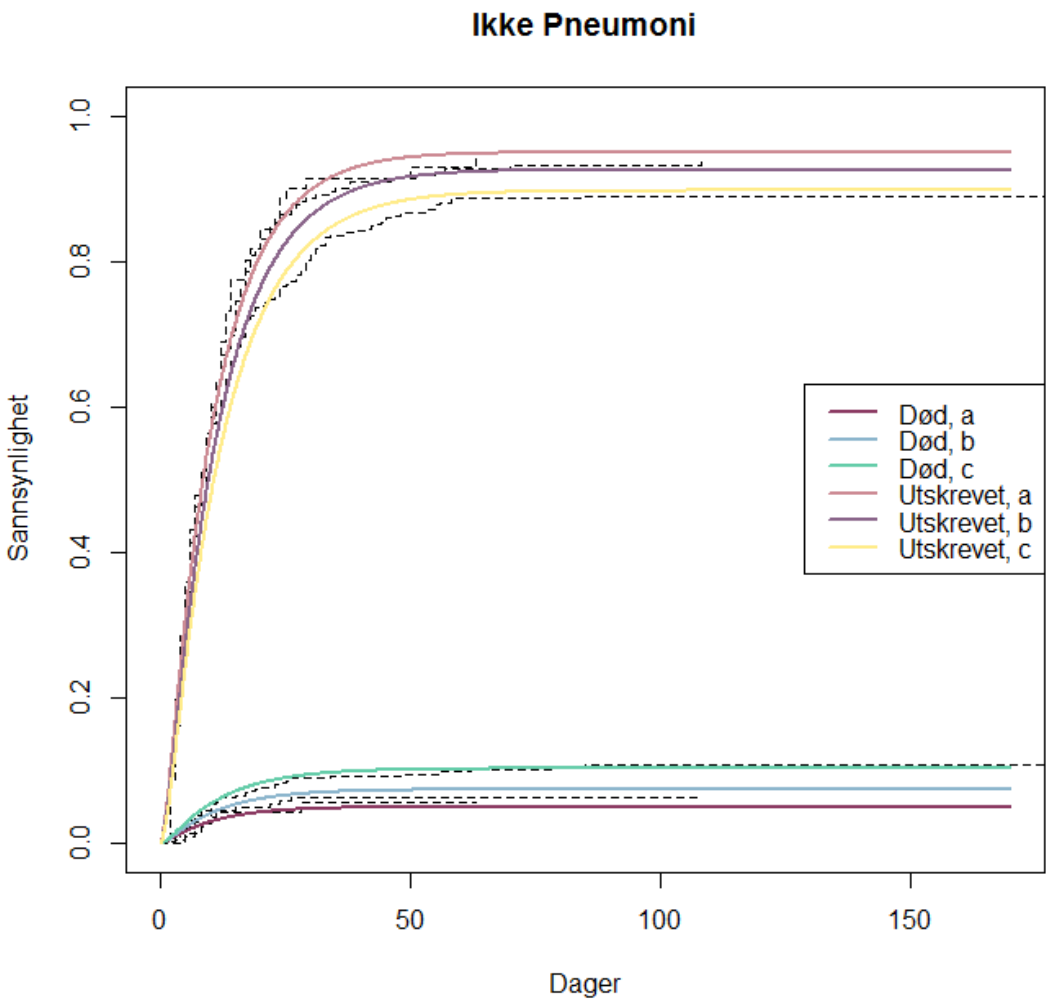
H_0	H_1	loglik H_0	loglik H_1	D	*df*	χ^2_{df} 95%
$\beta_1 = 0$	$\beta_1 \neq 0$	-2400.188	-2400.087	0.202	1	3.841
$\beta_2, \beta_3 = 0$	$\beta_2, \beta_3 \neq 0$	-2404.907	-2400.087	9.640	2	5.991

10.3 Sammenligne med `cuminc`

Funksjonen `cuminc` er ikke-parametrisk og tegner kumulative insidensfunksjoner fra Aalen-Johansen estimatoren, for de ulike gruppene separat. De tilsvarende funksjonene fra phase-type modellene med kovariater er funnet som beskrevet tidligere ved å først finne likelihoodfunksjonen og deretter maksimere ved hjelp av `optim`. For kovariatene alder og kjønn gir modell 1 og 2 svært like resultater. For modell 2 er disse vist i figurene 10.1 og 10.2. Phase-type modellene med kovariater klarer å tilpasses bra i grove trekk, men klarer ikke følge kurvene nøyaktig. Fra figur 10.1 ser ut til at kvinner har større sjanse for å dø uten pneumoni enn menn. Figur 10.2 gir forventet resultat på den måten at de yngste har lavest sannsynlighet for å dø og størst sannsynlighet for å bli utskrevet.



Figur 10.1: Modell 2. De stiplede svarte kurvene er kumulative insidensfunksjoner fra `cuminc`, for gruppene 'Mann' og 'Kvinne'. De fargede kurvene er tilsvarende kumulative insidensfunksjoner fra phase-type modell 1, med kovariat kjønn.



Figur 10.2: Modell 2. De stiplede svarte kurvene er kumulative insidensfunksjoner fra `cuminc`, for gruppene a, b og c, der a tilsvarer pasienter i alder under 30, b tilsvarer alder mellom 30 og 60, og c tilsvarer alder over 60. De fargede kurvene er tilsvarende kumulative insidensfunksjoner fra phase-type modell 1, med kovariat alder.

Kapittel 11

Videre arbeid

Noe av det vanskeligste med en oppgave som dette er å klare å avgrense den. Det vil alltid være mer som er interessant å undersøke og ting det er mulig å jobbe videre med. I dette kapittelet blir det listet opp noen av de områdene som kan være aktuelt å jobbe videre med.

11.1 Flere phase-type modeller

I denne oppgaven er det fokusert på fire phase-type modeller av typen coxiske fordelinger, med to konkurrerende risikoer. Det er mange flere som hadde vært spennende å se nærmere på. Spesielt med tanke på semi-konkurrerende risikoer. I kapittel 9 avsnitt 9.2 ble det sett at phase-type modellene klarte å estimere formen på overgangssannsynligheten P_{01}, men at den ikke klarte å følge kurven i toppen. Det er mulig det finnes andre phase-type modeller som hadde klart dette bedre.

11.2 Direkte identifiserbarhet

Som det ble vist i kapittel 7 er det mulig å finne flere optimale løsninger for noen av phase-type modellene. For modell 2 ble det vist at det er to mulige løsninger. Det ble i forbindelse med dette forsøkt å finne en metode numerisk som gjør det mulig å velge hvilken man vil ha av dem. De to løsningene svarer til at enten $a > b$ eller $b > a$, der $a = k + l_1 + m_1$ og $b = l_2 + m_2$. Det vil si at enten er total rate ut fra tilstand 0 større enn total rate ut fra tilstand 1 eller omvendt. Dette er ofte noe som er kjent ut fra datasettet som analyseres. I og med at phase-type modellene er ment å være intuitive vil man helst ha den løsningen som stemmer

logisk. For å prøve å løse dette problemet ble det forsøkt å omparametrisere slik at $(k, l_1, l_2, m_1, m_2) \to (k, l_1, l_2, b, \gamma)$, der $\gamma = a - b$ for $a > b$ og $\gamma = b - a$ for $b > a$.

Dette fungerte bra med simulerte data, men for datasett 1 ble det problemer med at likelihoodfunksjonen ikke kunne optimeres i `optim`. Problemet kommer av at i likelihoodfunksjonen blir $f_j(t)$ noen ganger negativ og logaritmen til denne kan derfor ikke evalueres. Dette kan komme av følgende. Likelihoodfunksjonen beregnes som en sum av logaritmen av de tre funksjonene

$$\begin{aligned} f_2(t) &= \mathbf{p}e^{\mathbf{Q}t}\mathbf{L}\mathbf{v}_1, \text{ for } C = 2 \\ f_3(t) &= \mathbf{p}e^{\mathbf{Q}t}\mathbf{L}\mathbf{v}_2, \text{ for } C = 3 \\ 1 - F(t) &= \mathbf{p}e^{\mathbf{Q}t}\mathbf{I}_q, \text{ for sensurerte data.} \end{aligned}$$

De elementene som er avgjørende i disse funksjonene er matrisene $\mathbf{L}$ og $e^{\mathbf{Q}t}$. Matrisen $e^{\mathbf{Q}t}$ vil alltid ha positive elementer da den er på formen

$$e^{Qt} = \text{matrixExp} \begin{bmatrix} a & c \\ 0 & b \end{bmatrix} = \begin{bmatrix} e^a & \frac{c(e^a - e^b)}{a-b} \\ 0 & e^b \end{bmatrix}$$

Elementene e^a og e^b er selvfølgelig positive og $\frac{c(e^a - e^b)}{a-b}$ er positiv fordi

$$a > b \iff e^a > e^b.$$

Dette fører til at problemene må oppstå i L-matrisen. Den er originalt på formen

$$\mathbf{L} = \begin{bmatrix} l_1 & m_1 \\ l_2 & m_2 \end{bmatrix},$$

slik at med de nye parameterne blir den

$$\mathbf{L} = \begin{cases} \begin{bmatrix} l_1 & b + \gamma - l_1 - k \\ l_2 & b - l_2 \end{bmatrix} & \text{hvis } a > b \\ \begin{bmatrix} l_1 & b - \gamma - l_1 - k \\ l_2 & b - l_2 \end{bmatrix} & \text{hvis } b > a \end{cases}$$

I `optim` kan det garanteres at alle parameterne hver for seg er positive, ved å sette nedre grense. Det gjør at L med de gamle parameterne er garantert positiv. Med de nye parameterne derimot kan man garantere at b, γ, l_1 og k er positive, men ikke at $b + \gamma - l_1 - k$ er positiv. Dette kan være noe av det som er grunnen til at `optim` krasjer. Dette blir derfor et numerisk problem. Utfordringen videre er å finne parametere som det ikke oppstår problemer med.

11.3 Generell identifiserbarhet

I kapittel 7 ble det vist betingelser for identifiserbarhet for modell 2 og 3. Likevel er det slik nå at dette må gjøres helt fra starten dersom man finner en ny modell man ønsker å bruke. Det som hadde vært spennende videre hadde vært å finne en generell regel for hvor mange løsninger en modell har og hvilke restriksjoner de må oppfylle.

11.4 Kovariater

I denne oppgaven ble det bare sett på kovariater for de to enkleste modellene. Videre er det også mulig å se på kovariater i de andre modellene. Det kan også være andre måter å ta med kovariater i phase-type modellene.

Kapittel 12

Konklusjon

I denne oppgaven har phase-type modeller blitt tilpasset reelle konkurrerende risikodatasett. Det har blitt beskrevet fire phase-type modeller, av typen coxiske modeller, av ulik dimensjon og oppbygning.

Modellene har blitt testet ved å simulere data fra dem for så å estimere parameterne. Dette avdekket informasjon om at det kunne være mer enn én mulig løsning for flere av modellene. Ved å studere identifiserbarhet nærmere, ble det vist at for modell 2 er det to mulige løsninger. Modell 3 har i utgangspunktet seks løsninger, men visse restriksjoner må oppfylles for at de skal være mulige. Generelt vil det være en løsning for alle mulige permutasjoner av egenverdiene til $\mathbf{Q}$-matrisen til en modell, dersom de i tillegg oppfyller visse restriksjoner som forhindrer motsigelser.

Datasett 1 om pneumoni ved innleggelse ble tilpasset phase-type modellene på to ulike måter. Først ble de tilpasset ved å sortere de med og uten pneumoni og tilpasse to separate modeller. Her viste det seg at modell 1 tilpasset helt greit, men hadde en del å gå på spesielt for de med pneumoni. Modell 2 tilpasset mye bedre, men det ble ikke mye forbedring med modell 3 og 4. En modifisert versjon av modell 4, med gammafordelte overgangstider før den opprinnelige modell 2, ble også prøvd. Denne viste seg å bli tilpasset svært bra. Den andre måten phase-type modellene ble tilpasset var ved å la de med og uten pneumoni starte i ulike tilstander. Dette ble vist for modell 2, 3 og 4. Her viste det seg at modell 2 gav helt grei tilpasning, men modell 3 og 4 ble en del bedre. Modell 3 og 4 gav noe ulike kumulative insidensfunksjoner, men det var vanskelig å avgjøre hvilke som var best. Fordelen med tilpasning type 2 var at det bare ble estimert ett sett med parametere. Dette var en stor fordel med denne typen.

Datasett 2 om sykehuservervet pneumoni inneholder informasjon om tidspunkt for ervervet pneumoni, i tillegg til tidspunkt for død eller utskrivelse. Dette gjorde at det ble et semi-konkurrerende risikoproblem. Datasettet ble først analysert på samme måte som datasett 1, uten å ta hensyn til den ekstra informasjonen.

Her ble det vist modell 2 og 3 med og uten pneumoni hver for seg. Modell 2 ble tilpasset greit, mens modell 3 ble tilpasset svært bra. Det var dermed ulikt for de to datasettene hvilken modell som ble tilpasset bra. For å ta hensyn til tidspunktene for sykehuservervet pneumoni, ble phase-type modellene modifisert til flertilstandsmodeller. Dette ble gjort for modell 2 og 3. Overgangssannsynlighetene ble estimert og sammenlignet med Aalen-Johansen estimatoren fra R-pakken `etm`. Det viste seg at modell 2 og 3 gav veldig like resultater i denne sammenhengen.

Til slutt ble det sett på kovariater i modell 1 og 2 for datasettet om pneumoni ved innleggelse. Her var det informasjon om alder og kjønn på pasientene. Kovariater i modellene ble testet på tre ulike måter. Det viste seg at modellene kunne tilpasse informasjonen til kovariatene i grove trekk, men klarte ikke å tilpasse nøyaktig.

Bibliografi

[Aalen, 1995] Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22(4):pp. 447–463.

[Ansell and Phillips, 1994] Ansell, J. and Phillips, J. (1994). *Practical Methods for Reliability Data Analysis*. Oxford science publications. Clarendon Press.

[Beyersmann, 2012] Beyersmann, J., A. A. S. M. (2012). *Competing Risks and Multistate Models with R*. Springer New York.

[Bladt, 2005] Bladt, M. (2005). A review on phase-type distributions and their use in risk theory.

[Cox, 1972] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):pp. 187–220.

[Crowder, 2001] Crowder, M. J. (2001). *Classical competing risks*. CRC Press.

[Fine and Gray, 1999] Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 22(4):pp. 447–463.

[Fine et al., 2001] Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4):907–919.

[Kalbfleisch and Prentice, 2011] Kalbfleisch, J. and Prentice, R. (2011). *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley.

[Kjølen, 2014] Kjølen, S. H. (2014). *Phase-type modeller for konkurrerende risikoer, en sammenligning med tradisjonelle metoder*. Institutt for matematiske fag, Norges teknisk-naturvitenskapelige universitet.

[Lindqvist, 2013] Lindqvist, B. H. (2013). Phase-type distributions for competing risks, in proceedings of the 59th isi world statistics congress. pages 25–30.

[Meeker and Escobar, 2014] Meeker, W. and Escobar, L. (2014). *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley.

[Neuts, 1981] Neuts, M. (1981). *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Algorithmic Approach. Dover Publications.

[Peng et al., 2008] Peng, L., Jiang, H., Chappell, R., and Fine, J. (2008). An overview of the semi–competing risks problem. *Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*, pages 177–192.

[Prentice et al., 1978] Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., J., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):pp. 541–554.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Ross, 2010] Ross, S. M. (2010). *Introduction to probability models*. Academic Press, 10th edition edition.

[Slud and Suntornchost, 2014] Slud, E. V. and Suntornchost, J. (2014). Parametric survival densities from phase-type models. *Lifetime data analysis*, 20(3):459–480.