Intervall mellom fødslar studert ved "Phase Type Distributions"

av

Leiv Magne Asperheim

Oppgåve for graden

Master i Statistikk
Finansteori og Forsikringsmatematikk

Universitetet i Bergen

Matematisk Institutt

31. mai 2011

Takk

Eg vil rette ein stor takk til rettleiaren min, Trygve S. Nilsen, for å ha vore tålmodig, hjelpsam og motiverande. Det har vore ei inspirerande oppgåve, som har vore spennande å jobbe med.
Trygve S. Nilsen og Dag Tjøstheim skal ha ein spesiell takk for gode forelesingar, noko som var med å bidra til valget mitt å ta master i statistikk.
Vidare vil eg takke familien min for god støtte.
Ein stor takk rettast mine medstudentar for fagleg hjelp, samt sosiale stunder både på og utanfor universitetet. Her kunne eg nevnt mange, men vil spesielt sei at Miriam, Sindre, Gunhild og Bjarte har vore gode støttespelarar både fagleg og sosialt.
Til slutt vil eg takke Catrine og Inghild for å ha lese gjennom oppgåva.

Innhald

1 Introduksjon

1.1 "Phase Type Distributions"

"Phase Type Distributions" eller ei phase type fordeling er fordelinga til ei hendelsestid T i ei Markovkjede med endeleg tilstandsrom. Hendelsestida T er ein tilfeldig variabel som beskriv tida til absorpsjon for ein Markovprosess med ein absorberande tilstand. Jamfør Aalen, Borgan og Gjessing [2] side 388-390.

1.2 Innleiing

Dette er ei oppgåve som omhandlar tida mellom første og andre barn studert ved "Phase Type Distributions", basert på ein artikkel av Odd O. Aalen [1], eksempel 7. Denne phase type fordelinga er i vårt tilfelle fordelinga til tida mellom første og andre barn.

Me vil altså studere tida mellom første og andre barn ved å sjå på den tilhøyrande phase type fordelinga for denne tida. Denne finn me ved å prøve å tilpasse ei phase type fordeling til det aktuelle datasettet me ser på. Eg vil i denne oppgåva bruke eit nyare datasett enn Aalen brukte i [1], for så å sjå på korleis dei resultata eg får samsvarar med resultata han fekk. Dette vil sei oss noko om det har blitt endringar i fødselsmønsteret i tida som har gått mellom dei to datasetta.

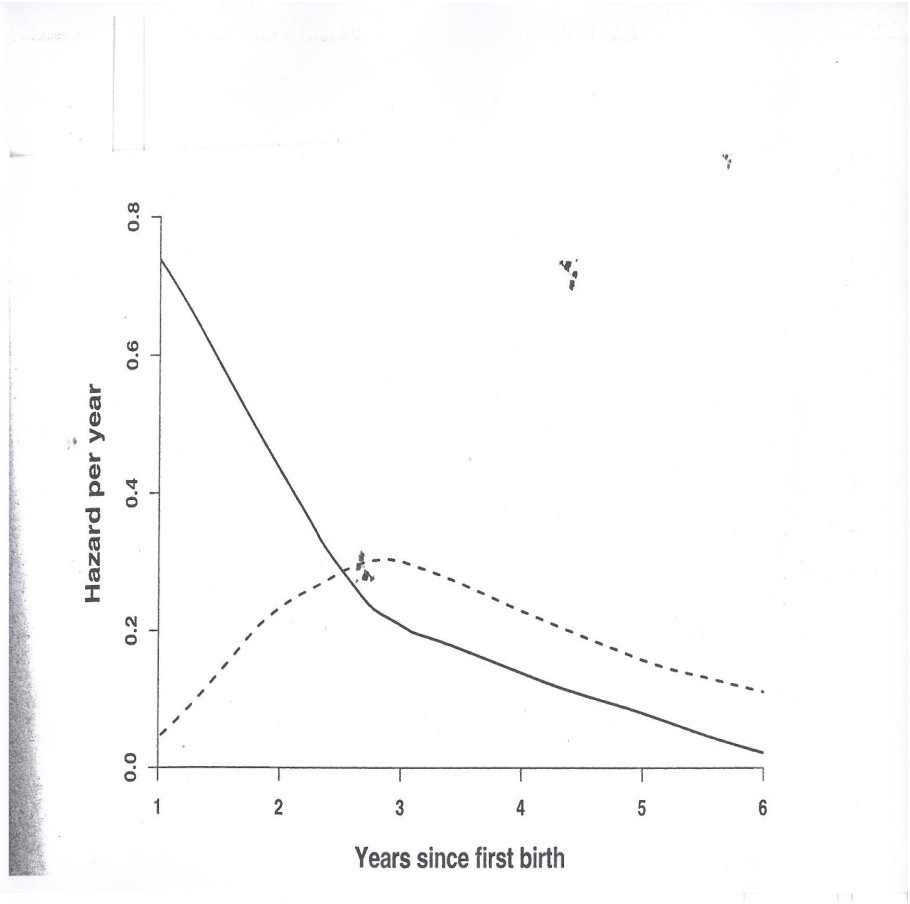
Medan Aalen i [1] brukte eit datasett med reelle data på 1779 mødre som hadde fått eit levande født førstebarn i perioden 1967-71, vil me bruke eit datasett med reelle data brukt i eksempel 1.1 i [2], side 8. Dette datasettet består av 53 558 observasjonar, der 53 296 av mødrene fekk eit levande første barn i perioden 1983-1997. Dei resterande 262 observasjonar der mødrene fekk eit dødt første barn er me kun interessert i når me skal undersøke hasardkurva for desse mødrene i Fig. 3.10 i [2], side 87. Denne figuren har eg tatt med i Figur 1.1 fordi eg vil samanlikne dei kurvene eg får med denne. Eg vil fra no av betegne dei to nemnte datasetta som henholdsvis "Aalens datasett" og "bokas datasett". Me ser altså at det er forskjell i tida desse datasetta er samla inn på, samt størrelsen på

dei.

Før eg byrjar å bruke Aalens metode for å tilpasse ei phase type fordeling [1], vil eg bruke glatting som ein alternativ metode for å finne ei fordeling til tida mellom første og andre barn. Dette går eg nærare inn på i kapittel 3.

Markovmodellen benytta i eksempel 7 i [1] for å tilpasse ei phase type fordeling vil eg fra no av betegne som "den dynamiske modell". Før eg byrjar å bruke "den dynamiske modell" på "bokas datasett" for å tilpasse ei phase type fordeling, vil eg imidlertid prøve "den dynamiske modell" på eit datasett eg veit er veldig likt "Aalens datasett". Dette fordi at "den dynamiske modell" ikkje var like problemfri å bruke når eg brukar den på "bokas datasett". Eg vil derfor undersøke om eg får samme resultat og at ting passar like godt som i [1] når eg brukar eit datasett som liknar "Aalens datasett". Av den årsak vil eg i kapittel 4 byrje med å prøve å tilpasse ei phase type fordeling på eit datasett eg har simulert til å vere så likt "Aalens datasett" som mogleg. Eg har simulert det fordi eg ikkje har tilgang på sjølve datasettet. Eg vil fra no av omtale dette datasettet som "det simulerte datasett". Når eg har gjort dette vil eg byrje med det som verkeleg er interessant, nemleg å sjå på kva som skjer når eg brukar "den dynamiske modell" på "bokas datasett" for å prøve å tilpasse ei phase type fordeling. Dette tar eg for meg i kapittel 5. I kapittel 6 vil eg introdusere min eigen Markovmodell og sjå korleis det går når eg prøver å tilpasse ei phase type fordeling ved hjelp av denne.

Til slutt vil eg bemerke at eg i oppgåva brukar ulike benemnelsar for fordelinga til tida mellom første og andre barn. Eg vil bruke phase type fordeling som benemnelse slik som Aalen har gjort det i [1], men eg vil og bruke hasardrate og estimert hendelsesrate som benemnelse der eg finn det naturleg. Tradisjonelt blir desse orda brukt om det samme, og det vil eg og gjere. Kan merke oss at det er den betinga fordelinga til tida mellom første og andre barn ved ei tid t, eg ser på, gitt at me har overlevd fram til denne tida t.



Figur 1.1: Estimerte hasardkurver for tid mellom første og andre barn. Desse tidene er henta fra "bokas datasett" og figuren er kopiert fra Fig. 3.10 i [2]. Heiltrekt linje: første barn døydde innan sitt første leveår, prikkete linje: første barn overlevde sitt første leveår.

2 Overlevelsesanalyse

For det følgjande kapittelet har eg henta informasjon fra [2] side 1-6. Overlevelsesanalyse er ein statistisk metode der ein er interessert i oppståinga av hendingar. Med hendingar meinar me hendingar i liva til individ som er av spesiell interesse i forskningsstudiar innanfor medisin, demografi, biologi, sosiologi, økonometri, osv. Eksempel på slike hendingar er død, hjerteinfarkt, forelskelse, bryllaup, skilsmisse og fødsel. I klassisk overelevelsesanalyse fokuserar ein på ei enkel hending for kvart individ, der oppståinga av hendinga blir beskrive ved hjelp av overlevelseskurver og hasardratar, og ved analysar av avhengigheiten mellom kovariatar med bruk av regresjonsmodellar. Viss ein samanføyer fleire hendingar med kvarandre, når dei oppstår for eit individ over tid, får me hendingshistoria. Ein kan til dømes vere interessert i å sjå på korleis menneske går gjennom ein sjukdom. Her kan følgjeleg sjukdommen ha fleire ulike tilstandar. Men hendingshistoriar er ikkje berre begrensa til menneske. Ei følgje av hendingar kan og skje med dyr, planter, celler, amalgamfyllingar, hofteproteser, lyspærer, bilar, osv. Ein kan sei at det kan skje med alt som forandrar, utviklar, eller forfell seg. Overlevelsesanalyse og hendingshistorieanalyse er brukt som eit verktøy på mange ulike område, der nokon døme er:

- Bevise eller motbevise nytten av medisinsk behandling for sjukdom.
- Forståing av riskfaktorar, og dermed hindre sjukdom.
- Evaluering av pålitelegheit for teknisk utstyr.
- Forståing av mekanismane i biologiske fenomen.
- Observasjon av sosiale fenomen som skilsmisse og arbeidsledigheit.

Det som er interessant i mi oppgåve er klassisk overlevelsesanalyse. Altså er me kun interessert i tida til ei enkel hending for kvart individ. Meir spesifikt kan me sei tida fra ei innleiande hending til ei hending me er interessert i skjer. Døme her kan vere:

- Tid fra fødsel til død.

- Tid fra første til andre fødsel.
- Tid fra sjukdom til død.
- Tid fra bryllaup til skilsmisse.

Som eit allmennt namn for desse tidene vil me bruke survival time eller overlevelsestid. Dette sjølv om endepunktet kan vere noko anna enn død.

Eit problem ein nesten alltid møter når ein studerar overlevelsestider er at sidan ein ventar på at ei hending skal skje, så vil ein når studien er over og analysen skal ta til finna ut at hendinga har skjedd for nokre individ, men ikkje for andre. I ein studie om tid mellom første og andre barnefødsel vil nokre av kvinnene ha fått sitt andre barn i tida studien pågjekk, medan andre ikkje har fått det. Dei vil kunne få sitt andre barn seinare, men det er ukjent for oss når me analyserer data. Dermed blir data me har samla inn bestående av ein blanding av fullstendige og ufullstendige observasjonar. Dette er årsaken til at me treng ein anna statistisk teori for desse observasjonane enn for observasjonar som til dømes måling av blodtrykk. Her er alle observasjonane fullstendige og me brukar godt utvikla metodar for kontinuerlig eller muligens diskret data. Dei ufullstendige observasjonane kallar me sensorerte overlevelsestider. Dette kan skje ved avslutninga av studien ved at nokre personar ikkje har opplevd hendinga me interesserer oss for, eller ved at nokon trekker seg fra studien eller forsvinn fra oppfølging under studien. Dei personane som ikkje har opplevd hendinga av interesse ved ei gitt tid t, og som heller ikkje har blitt sensorerte ved tid t, er antallet som er i risiko ved tid t.

Sjølv om dei vanlege statistiske metodane ikkje kan takle sensorerte overlevelsesdata, så er det nokså lett å analysere slike data. Det ein treng er dei riktige omgrepa, og det er to grunnleggjande omgrep som er tilstades under all teori om overlevelsesanalyse, nemleg overlevelsesfunksjonen og hasardraten. Hasardraten er den betinga sannsynsfordelinga til overlevelsestida for hendinga me ventar på. I mi oppgåve er det tida fra første til andre fødsel. Ein startar med ein viss mengde av individ ved tid null og ventar på at ei spesiell hending skal skje for desse. Overlevelsesfunksjonen $S(t)$ gjev da den forventa andel av individ som ikkje har opplevd hendinga ved tid t. Viss den tilfeldige variabelen T betegnar overlevelsestida, kan me formelt skriva

$$S(t) = P(T > t) \, . \tag{2.1}$$

Overlevelsesfunksjonen gjev altså sannsynet for at hendinga av interesse ikkje har skjedd ved tid t, noko som ikkje treng å ha med død å gjere. Ofte vil overlevelsesfunksjonen gå mot null når t blir større fordi etterkvart som tida går vil fleire og fleire oppleve hendinga av interesse. Men me kan og sjå på hendingar som ikkje nødvendigvis skjer med alle individa, og da vil overlevelsesfunksjonen minke mot ein positiv verdi når t går mot uendeleg. Eksempel på dette kan vere skilsmisser og testikkelkreft. Overlevelsesfunksjonen (2.1) gjev det ubetinga sannsynet for at hendinga av interesse ikkje har skjedd ved tid t. Hasardraten $\alpha(t)$ er derimot definert ved hjelp av betinga sannsyn. Ved å anta at T er

absolutt kontinuerlig, det vil sei at T har ein sannsynstettleik, kan me sjå på dei individa som ikkje har opplevd hendinga av interesse ved tid t, og sjå på kva sannsynet er for at dei opplever denne hendinga i eit lite tidsintervall $[t, t+dt)$. Da vil dette sannsynet vere lik $\alpha(t)dt$. Meir presist er hasardraten definert som ei grense på følgjande måte

$$\alpha(t) = \lim_{\triangle t \to 0} \frac{1}{\triangle\ t} P(t \leq T < t + \triangle\ t | T \geq t) \ . \tag{2.2}$$

Me kan her merka oss at medan overlevelseskurva vil starte i 1 og avta over tid, kan hasardraten vere ein kva som helst ikkje-negativ funksjon. Me kan og legge merke til at hasardraten er modellen sitt motstykke til hendelsesraten som er brukt mykje i epidemiologiske studiar. Fra sensorerte overlevelsesdata kan me lett estimere overlevelseskurva ved hjelp av Kaplan-Meier estimatoren. Sidan denne estimatoren ikkje er noko ein treng kunnskap om i denne oppgåva, vil me ikkje gå noko meir inn på den. Å estimere hasardraten er meir komplisert. Det me gjer først er å estimere den kumulative hasardraten

$$A(t) = \int_0^t \alpha(s) ds \tag{2.3}$$

ved Nelson-Aalen estimatoren. Dette er ein ikkje-parametrisk estimator som er brukt for å estimere den kumulative hasardraten $A(t)$ fra sensorerte overlevelsesdata. Denne er gitt ved (3.4) i [2] side 72, og er som følgjer

$$\widehat{A}(t) = \sum_{T_j \leq t} \frac{1}{Y(T_j)} \ . \tag{2.4}$$

Me kan da glatte aukingane i Nelson-Aalen estimatoren for å finne eit estimat av hasardraten, sjå (3.1). Det er 2 fundamentale matematiske samanhengar mellom overlevelsesfunksjonen og hasardraten. Den første er at ved (2.2) og (2.3) har me at

$$\begin{aligned}
\frac{d}{dt}A(t) = \alpha(t) &= \lim_{\triangle t \to 0} \frac{1}{\triangle\ t} P(t \leq T < t + \triangle\ t | T \geq t) \\
&= \lim_{\triangle t \to 0} \frac{1}{\triangle\ t} \frac{S(t) - S(t + \triangle\ t)}{S(t)} \\
&= -\frac{1}{S(t)} \lim_{\triangle t \to 0} \frac{S(t + \triangle\ t) - S(t)}{\triangle\ t} \\
&= -\frac{\frac{d}{dt}S(t)}{S(t)} \ . \\
\Rightarrow \alpha(t) = \frac{d}{dt}A(t) &= -\frac{\frac{d}{dt}S(t)}{S(t)} \ .
\end{aligned} \tag{2.5}$$

Så ved integrasjon og bruk av at S(0)=1 får me den andre samanhengen som er

$$\begin{aligned}
\int_0^t \alpha(s) ds &= -\log(S(t)) \ . \\
S(t) &= e^{-\int_0^t \alpha(s) ds} \ .
\end{aligned} \tag{2.6}$$

3

Hasardkurvene for datasetta funne ved glatting av Nelson-Aalen estimatoren

Før eg vil forsøke å tilpasse ei phase type fordeling til datasetta, vil eg i dette kapittelet studere datasetta brukt i [1] og [2] ved å glatte aukingane i Nelson-Aalen estimatoren (2.4). Merk at eg ikkje brukar det samme datasettet som Aalen, men eit datasett eg har simulert til å likne hans gitt i tillegg A. Som nemnt i innleiinga kallar eg dette datasettet for "det simulerte datasett". Simuleringa er gjort i ei programpakke kalla R [8], og det er i dette programmet at eg har gjort all programmering for denne oppgåva. For å forstå simuleringa anbefalar eg å vente med å sjå på tillegg A til me byrjar på kapittel 4. Glattinga av aukingane i Nelson-Aalen estimatoren vil følgjeleg gje meg hasardraten for overlevelsestida, der overlevelsestida er den tida som går mellom første og andre barn.

Me vil starte med å bruke ein ferdig funksjon i R kalla muhaz, jamfør [7]. Muhaz estimerar hasardraten fra høgre-sensorerte data ved å bruke kjernebaserte metodar. Muhaz treng kun informasjon om overlevelsestidene, og om tidene er sensorerte eller usensorerte. Når den har fått denne informasjonen gjev den tilbake ein hasardrate. Denne hasardraten er som nemnt funne ved hjelp av glatting av aukingane i Nelson-Aalen estimatoren. Hasardraten er ei betinga fordeling til overlevelsestida. Før eg gjev grafen til hasardraten me finn ved hjelp av muhaz, vil eg gje ei kort forklaring av korleis denne glattinga føregår.

Ein kan bruka fleire metodar for å glatte denne estimatoren, men muhaz brukar kjernefunksjonsglatting. Da blir hasardraten $\alpha(t)$ estimert av eit vekta gjennomsnitt av Nelson-Aalen aukingane $\Delta\widehat{A}(T_j)$ over intervallet [t-b,t+b]:

$$\widehat{\alpha}(t) = \frac{1}{b}\sum_{T_j} K(\frac{t-T_j}{b})\Delta\widehat{A}(T_j)\ . \tag{3.1}$$

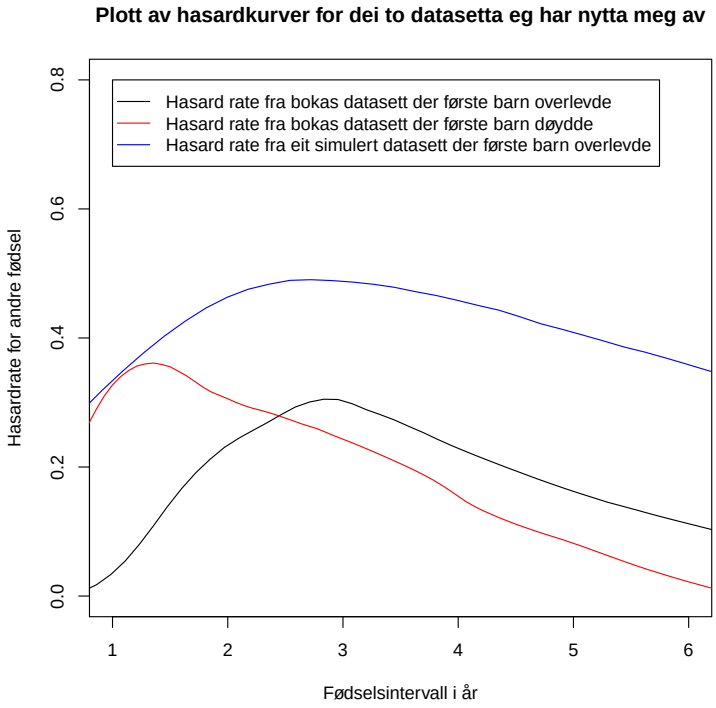
Her er b bandbreidde, medan kjernefunksjonen $K(x)$ er ein bunden funksjon som forsvinn utanfor [-1,1] og har integral 1. Eksempel på slike kjernefunksjonar er:

- Uniform kjerne: $K(x) = 1/2$

- Epanechnikov kjerne: $K(x) = 3(1 - x^2)/4$
- Biweight kjerne: $K(x) = 15(1 - x^2)^2/16$

Legg merke til at formlane gjeld for $|x| \leq 1$ og at alle dei 3 kjernefunksjonane er 0 når |x|>1, jamfør [2] side 85.

I denne oppgåva brukar me Epanechnikov som kjernefunksjon. Ved å bruke muhaz funksjonen på "det simulerte datasett" samt "bokas datasett" får eg da følgjande figur:



Figur 3.1: Eg har her samanlikna korleis hasardkurvene for dei 2 datasetta blir der første barn overlevde sitt første leveår, samt sett på korleis hasardkurva for "bokas datasett" blir der første barn døydde i sitt første leveår.

Programmeringen i R som har gjeve meg denne figuren følgjer av koden som er gitt i tillegg B. Ser utifra Figur 3.1 at hasardraten blir mindre i "bokas datasett" enn det den blir i "det simulerte datasett". Kan dermed sjå ut som det har blitt ei endring i fødselsmønsteret for dei mødrene som fekk eit levande første barn. Dette vil me få nærare svar på når me ser vidare på datasetta i kapittel 4, 5 og 6. Kan og legge merke til at hasardkurva for "bokas datasett", der første barn overlevde virkar å vere identisk med den prikkete hasardkurva som er i Figur 1.1. Sidan den også er funne ved muhaz på samme datasett, bør desse vere like og gjev meg ein bekreftelse på at muhaz funksjonen fungerer.

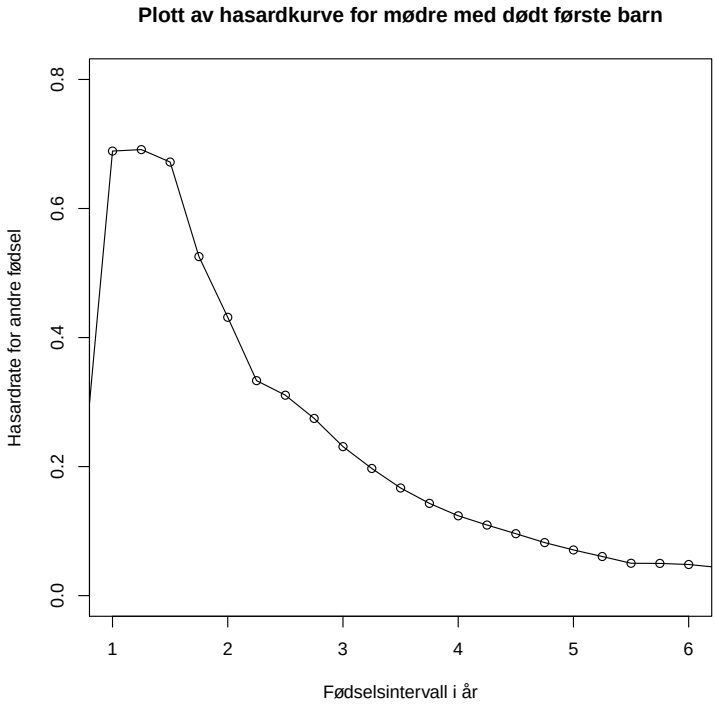
I Figur 3.1 er det også tatt med eit plott av korleis hasardraten ser ut for mødre fra "bokas datasett" der første barn døydde. Me ser at denne hasardkurva ikkje samsvarar med hasardkurva for samme datasett i Figur 1.1. Medan hasardkurva eg finn er ganske lik kurva i Figur 1.1 fra 3 år og utover, er dei totalt ulike mellom 0 til 3 år. Figur 1.1 får her ei kurve som avtek i ei nærast rett linje fra ein hasardrate på over 0.7 ved 1 år til 0.2 ved 3 år. Medan som ein ser av Figur 3.1 så får eg for tilsvarande tidsintervall ei kurve som stig til ein topp med hasardrate på litt over 0.3 ved 1,5 år og som synk ned til 0.2 ved 3 år. Sidan desse også burde samsvare med kvarandre har eg valgt å sjå litt nærare på kva som kan vere årsaken til desse ulike resultata.

Eg byrjar da med å lage min eigen funksjon som skal likne muhaz. Eg vil at denne funksjonen skal bruke kjernefunksjons glatting med Epanechnikov som kjerne slik som er blitt brukt i muhaz funksjonen. Vidare må me ta hensyn til at grensepunkta i nærleiken av 0 år vil få ein Epanechnikovkjerne som vektar observasjonane med ein kjerne som strekker seg utanfor området me ser på. Epanechnikovkjernen vår vil her vekta observasjonane som om det fantes eit område før tid 0 med i observasjonsmengden vår. Derfor innfører me ein boundarykjerne som også muhaz funksjonen nyttar seg av. Boundarykjernen til Epanechnikov blir

$$K_q(x) = \frac{12}{(1+q)^4}(x+1)\{x(1-2q) + \frac{3q^2-2q+1}{2}\} \, . \tag{3.2}$$

Jamfør Muller og Wang [5]. Denne funksjonen gjeld da på området $[-1, q]$, der q er avstanden fra $x = 0$ til grensepunktet for observasjonane. Eg har da alt som trengs for å lage min eigen muhaz liknande funksjon i R. Koden for dette ligg også i tillegg B og me får utifrå dette hasardkurva i Figur 3.2 på neste side. Eg har brukt forskjellig bandbreidde etter kor nær grensa me er i denne figuren. Me ser at denne hasardkurva er mykje meir lik den som er i Figur 1.1 enn kva den liknar den me fann ved å bruke muhaz funksjonen i Figur 3.1. Årsaken til desse forskjellane skuldas at bandbreidda som muhaz funksjonen har valgt har vore for stor nær grensepunktet for observasjonane, noko som har dratt toppen på kurva mykje lenger ned enn kva som er tilfelle med mindre bandbreidde. Det som blir ulempa med å minske bandbreidden er at kurva blir mindre glatt, noko ein kan sjå er tilfellet i Figur 3.2 nær grenseområdet ved tid 0.

Konklusjonen ein kan dra utifra dette når det gjeld mødre med dødt første barn, er at muhaz funksjonen gjev ein for liten hasardrate i byrjinga av intervallet. Difor er nok den hasardraten me finn ved vår eigen funksjon meir korrekt. Den liknar meir på hasardraten i Figur 1.1 og sidan me har mindre bandbreidde nær grenseområdet, vil ikkje mangelen på observasjonar nær tid null klare å dra kurva like langt ned som det som var tilfelle i muhaz funksjonen.



Figur 3.2: Har her brukt "bokas datasett" med bandbreidde på 0.3 år nær grensa, medan eg har auka bandbreidda til 2 år ellers.

4

Tilpassing av ein "Phase Type Distribution" ved å bruke "den dynamiske modell" på "det simulerte datasett"

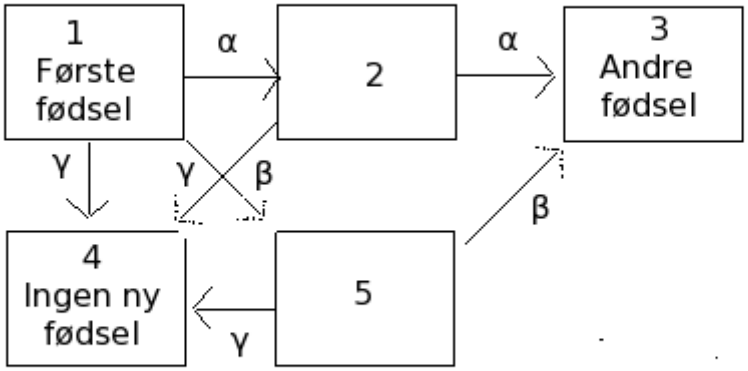
4.1 "Den dynamiske modell"

Aalen ville i artikkelen sin [1], prøve å finne ei phase type fordeling som han kunne tilpasse med datasettet han nytta seg av. Dette datasettet har eg i innleiinga definert som "Aalens datasett" som inneheld informasjon om tidene mellom første og andre barn. Aalen ville analysere denne tida som er av interesse i demografiske undersøkelsar. Datasettet er henta fra fødselsregisteret i Norge for 1779 kvinner som fekk eit levande første barn i perioden 1967-71 og som vart følgt opp fram til 1982. Me vil som nemnt bruke eit liknande datasett kalla "det simulerte datasett" fordi me ikkje har tilgang på "Aalens datasett". Først analyserte Aalen data ved aktuarmetoden. Det vil sei at ein etter første fødsel har observert hendelsesraten for andre fødsel med intervall på 6 månadar, der ein deler antall fødslar i eit intervall på antall som er i fare for å føde ved byrjinga av intervallet. Dette er punkta som kjem i Figur 4.3 i slutten av kapittelet og kallast den observerte hendelsesraten. Det som meinast med å tilpasse ei phase type fordeling er at fordelinga me finn skal passe så godt som mogleg med den observerte hendelsesraten.

Så kan me byrje å sjå nærare på "den dynamiske modell" som Aalen vidare introduserar. Dette er ein Markovmodell som består av fem tilstandar og er beskrive av Figur 4.1 på neste side. Sidan me i resultat fra Figur 3.1 har sett at hendelsesraten for "det simulerte datasett" har ein topp på rundt tre år etter første fødsel, kan me også få ein liknande topp for vår fordeling ved å ha eksponensielle steg i ein serie. For eit steg mellom første og andre barn vil tida vere eksponensielt fordelt og dermed vil fordelinga til tida vere monotont avtagande. Me ynskjer derimot ei fordeling som har ein topp rundt tre år etter første barn. Viss me velger å ha to steg mellom første og andre barn vil fordelinga til tida stige til ein topp for deretter å synke igjen. Dette er ei slik form me vil ha fordelinga vår på og derfor er det naturleg å ha minst to steg i modellen. Da vil me minimum ha tre tilstandar, der å gå fra første til andre tilstand betyr at paret vil forsøke å få eit nytt barn, medan å gå fra andre til tredje tilstand betyr at paret har fått eit nytt barn. Så

kan ein ta hensyn til at det eksisterar heterogenitet/skjørhet mellom ulike foreldrepar. Nokon vil fort få eit nytt barn, medan andre vil vente i fleire år. Dermed bør det vere minst 2 vegar ein kan gå gjennom tilstandane på, der ein er kjappare enn den andre. Til slutt må ein óg ta hensyn til at nokre kvinner aldri vil få eit andre barn. Dermed må me ha ein ekstra absorberande tilstand i tilstandsrommet.

Aalen har valgt å halde antall parametrar så lågt som mogleg slik at me kan identifisere desse. Eit individ startar altså i tilstand 1 og kan derfra gå til tilstand 3 gjennom tilstand 2 og 5. Intensitetane mellom dei to tilstandane i den øvre vegen å gå er satt lik α, medan i den nedre vegen er dei satt lik β. Fra alle dei transiente tilstandane er det ein moglegheit med intensitet γ som er å bli absorbert i tilstand 4. Å bli absorbert i tilstand 4 betyr at ingen andre fødsel vil skje. Me har da introdusert "den dynamiske modell", og vil i resten av kapittelet vise korleis me kan bruke denne til å finne ei phase type fordeling for datasettet me ser på. Eg understrekar igjen at eg brukar "det simulerte datasett" og ikkje "Aalens datasett". Basert på denne modellen vil eg i neste delkapittel utleie differensiallikningane for dei ulike tilstandane.



Figur 4.1: Markov modell for fødselsintervall med 5 tilstandar, i denne oppgåva kalla "den dynamiske modell".

4.2 Utleiing av differensiallikningar

Før me byrjar å utleie differensiallikningane for "den dynamiske modell" som er ein fem tilstands Markovmodell, vil eg gjennomgå litt nødvendig teori.

Ein stokastisk prosess er ei samling av tilfeldige variablar $X(t)$, der t er ein parameter fra ein mengde T som kan vere anten diskret eller kontinuerlig:

$$T = \begin{cases} \{0, 1, 2, ...\}, & \text{for diskret T,} \\ (0, \infty), & \text{for kontinuerlig T.} \end{cases}$$

Stokastiske prosessar er karakterisert av tilstandsrommet, som er dei moglege verdiane den tilfeldige variabelen $X(t)$ kan ta, ved indeksmengden T og ved avhengighetsrelasjonane mellom dei tilfeldige variablane $X(t)$. Jamfør Taylor og Karlin [9], side 5.

Ein Markovprosess $X(t)$ er ein type stokastisk prosess med eigenskapen at gitt verdien til $X(t)$, så er verdien til $X(s)$ for $s > t$ ikkje påverka av verdien til $X(u)$ for $u < t$. Med andre ord kan ein sei at sannsynet for framtidig oppførsel for prosessen, når den noverande verdien er kjent eksakt, ikkje blir påverka av tilleggsinformasjon angåande tidligare oppførsel. Formelt kan me sette opp Markoveigenskapen som

$$P(X(t+s) = j | X(s) = i, X(t_n) = i_n, ..., X(t_1) = i_1) = P(X(t+s) = j | X(s) = i) \ , \tag{4.1}$$

gitt tidene $0 < t_1 < t_2 < ... < t_n < s$ og $t > 0$. Jamfør [2] side 463 og Norberg [6] side 55. Det siste leddet i (4.1) er sannsynet for at $X(t+s)$ er i tilstand j gitt at $X(s)$ er i tilstand i som blir kalla for eitt stegs overgangssannsyn. Det er betegna som $P_{ij}(t)$. Altså er

$$P_{ij}(t) = P(X(t+s) = j | X(s) = i) \ . \tag{4.2}$$

Når eitt stegs overgangssannsynet kun er avhengig av separasjonen i tid t og ikkje er avhengig av startida s, seier me at Markovkjeda har stasjonære overgangssannsyn og kan kallast ei homogen Markovkjede. Formelt kan me sei at Markovkjeda er homogen og har stasjonære overgangssannsyn når

$$P(X(t+s) = j | X(s) = i) = P(X(t) = j | X(0) = i) \ . \tag{4.3}$$

Me ser i "den dynamiske modell" på ei Markovkjede i kontinuerlig tid; $X(t)$ $(t > 0)$ som er ein Markovprosess på tilstandane 1, 2, 3, 4 og 5. Tilstandsrommet S er endeleg i denne modellen. Me antek som vanlig at overgangssannsyna er stasjonære. Markoveigenskapen slår da fast at $P_{ij}(t)$ må tilfredsstilla

(a) $P_{ij}(t) \geq 0$,

(b) $\sum_{j=1}^{n} P_{ij}(t) = 1, \quad i, j = 1, 2..., n$

(c) $P_{ik}(s+t) = \sum_{j=1}^{n} P_{ij}(s) P_{jk}(t)$ for t,s$\geq$ 0 (Chapman-Kolmogorov likninga)

og me kan i tillegg sei at

(d)

$$\lim_{t \to 0+} P_{ij}(t) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Over er n antall tilstandar i modellen. Jamfør [9] side 394.

Chapman-Kolmogorov likninga eg no vil bruke for å utlede differensiallikningane for "den dynamiske modell" er jamført [9], side 356:

$$P_{ij}(t+h) = \sum_{k=1}^{n} P_{ik}(t)P_{kj}(h) \ . \tag{4.4}$$

Her er fortsatt n antall tilstandar i modellen ein ser på.

$X(t)$ er den tilfeldige variabelen som fortel kva tilstand me er i ved tid t. La oss anta at prosessen er i tilstand i ved tid t. Da vil q_{ij} (for $i \neq j$) måle kor fort tilstandsendringen fra i til j vil skje. Denne q_{ij} vil i "den dynamiske modell" vere ein av dei tre intensitetane α, β eller γ. Me får da i vårt tilfelle ein Markovprosess $X(t)$ der

(i) $P(X(t+h) = j|X(t) = i) = q_{ij}h + o(h)$ for $i \neq j$. $i, j = 1, 2, ..., n$,

(ii) $P(X(t+h) = i|X(t) = i) = 1 - \sum_{j=1, j\neq i}^{n} q_{ij}h + o(h)$,

(iii) $q_{ij} = 0$ viss det ikkje går an å hoppa direkte fra tilstand i til tilstand j.

Jamfør [9] side 395. Punkt (iii) seier oss altså at å hoppe meir enn eitt steg på tiden h vil ha sannsyn $o(h)$. Når $h \to 0$ vil $o(h)$ gå mot null, dermed tar eg ikkje med moglegheitane for meir enn eit hopp på tida h når eg skal bruke (4.4). Me kan da ved hjelp av desse punkta og bruk av (4.4) utleie differensiallikningane for tilstandane i "den dynamiske modell". Me har altså at $n = 5$ i "den dynamiske modell" og nyttar den samme nummerering som Aalen [1]. Sidan ein i vår modell alltid startar i tilstand 1, vil eg for enkelheits skuld benemne $P_{1j}(t) = P_j(t)$. Her er j ein av tilstandane i modellen vår, så $P_j(t)$ er da sannsynet for at ei mor er i tilstand j ved tid t. Startbetingelsane for "den dynamiske modell" er $P_1(0) = 1$ og $P_j(0) = 0$ for $j \neq 1$. Utleier først for tilstand 1 og så dei andre tilstandane.

$$\begin{aligned}
P_1(t+h) &= P_1(t)P(X(t+h) = 1|X(t) = 1) \\
&= P_1(t)(1 - (\alpha + \beta + \gamma)h + o(h)) \ , \\
\lim_{h\to 0} \frac{P_1(t+h) - P_1(t)}{h} &= \lim_{h\to 0} -(\alpha + \beta + \gamma)P_1(t) + \frac{o(h)}{h}P_1(t) \ , \\
\frac{d}{dt}P_1(t) &= -(\alpha + \beta + \gamma)P_1(t) \ .
\end{aligned} \tag{4.5}$$

$$\begin{aligned}
P_2(t+h) &= P_1(t)P(X(t+h) = 2|X(t) = 1) \\
&\quad + P_2(t)(X(t+h) = 2|X(t) = 2) \\
&= P_1(t)(\alpha h + o(h)) + P_2(t)(1 - (\alpha + \gamma)h + o(h)) \ , \\
\lim_{h\to 0} \frac{P_2(t+h) - P_2(t)}{h} &= \alpha P_1(t) - (\alpha + \gamma)P_2(t) \ , \\
\frac{d}{dt}P_2(t) &= \alpha P_1(t) - (\alpha + \gamma)P_2(t) \ .
\end{aligned} \tag{4.6}$$

$$
\begin{aligned}
P_3(t+h) &= P_2(t)P(X(t+h)=3|X(t)=2) \\
&\quad + P_3(t)P(X(t+h)=3|X(t)=3) \\
&\quad + P_5(t)P(X(t+h)=3|X(t)=5) \\
&= P_2(t)(\alpha h + o(h)) + P_3(t)(1) \\
&\quad + P_5(t)(\beta h + o(h)) \,, \\
\lim_{h\to 0} \frac{P_3(t+h)-P_3(t)}{h} &= \alpha P_2(t) + \beta P_5(t) \,, \\
\frac{d}{dt}P_3(t) &= \alpha P_2(t) + \beta P_5(t) \,. \qquad (4.7)
\end{aligned}
$$

$$
\begin{aligned}
P_4(t+h) &= P_1(t)P(X(t+h)=4|X(t)=1) \\
&\quad + P_2(t)P(X(t+h)=4|X(t)=2) \\
&\quad + P_4(t)P(X(t+h)=4|X(t)=4) \\
&\quad + P_5(t)P(X(t+h)=4|X(t)=5) \\
&= P_1(t)(\gamma h + o(h)) + P_2(t)(\gamma h + o(h)) \\
&\quad + P_4(t)(1) + P_5(t)(\gamma h + o(h)) \,, \\
\lim_{h\to 0} \frac{P_4(t+h)-P_4(t)}{h} &= \gamma(P_1(t) + P_2(t) + P_5(t)) \,, \\
\frac{d}{dt}P_4(t) &= \gamma(P_1(t) + P_2(t) + P_5(t)) \,. \qquad (4.8)
\end{aligned}
$$

$$
\begin{aligned}
P_5(t+h) &= P_1(t)P(X(t+h)=5|X(t)=1) \\
&\quad + P_5(t)P(X(t+h)=5|X(t)=5) \\
&= P_1(t)(\beta h + o(h)) + P_5(t)(1-(\beta+\gamma)h + o(h)) \,, \\
\lim_{h\to 0} \frac{P_5(t+h)-P_5(t)}{h} &= \beta P_1(t) - (\beta+\gamma)P_5(t) \,, \\
\frac{d}{dt}P_5(t) &= \beta P_1(t) - (\beta+\gamma)P_5(t) \,. \qquad (4.9)
\end{aligned}
$$

Eg har no utleidd differensiallikningar for alle dei 5 tilstandane. Ynskjer så å løyse desse i neste delkapittel.

4.3 Løysing av differensiallikningane

Eg startar med å løyse (4.5) fra forrige delkapittel:

$$\begin{aligned}
\frac{d}{dt}P_1(t) &= -(\alpha+\beta+\gamma)P_1(t)\ , \\
\frac{\frac{d}{dt}P_1(t)}{P_1(t)} &= -(\alpha+\beta+\gamma)\ , \\
\int_0^t \frac{d}{d\tau}\log(P_1(\tau))\ d\tau &= \int_0^t -(\alpha+\beta+\gamma)\ d\tau\ , \\
\log(P_1(t)) - \log(P_1(0)) &= -[(\alpha+\beta+\gamma)\tau]_0^t = -(\alpha+\beta+\gamma)t\ .
\end{aligned}$$

Sidan $\log(P_1(0)) = \log(1) = 0$ blir dette

$$P_1(t) = e^{-(\alpha+\beta+\gamma)t}\ . \tag{4.10}$$

Me løyser så (4.6):

$$\begin{aligned}
\frac{d}{dt}P_2(t) &= \alpha P_1(t) - (\alpha+\gamma)P_2(t)\ , \\
\frac{d}{dt}P_2(t) + (\alpha+\gamma)P_2(t) &= \alpha P_1(t)\ .
\end{aligned}$$

Eg multipliserer så heile uttrykket med integrerande faktor som er $e^{\int_0^t \alpha+\gamma\ d\tau} = e^{(\alpha+\gamma)t}$ og får

$$\begin{aligned}
(\frac{d}{dt}P_2(t) + (\alpha+\gamma)P_2(t))e^{(\alpha+\gamma)t} &= \alpha P_1(t)e^{(\alpha+\gamma)t}\ , \\
\int_0^t \frac{d}{d\tau}(P_2(\tau)e^{(\alpha+\gamma)\tau})\ d\tau &= \int_0^t \alpha e^{-(\alpha+\beta+\gamma)\tau}e^{(\alpha+\gamma)\tau}\ d\tau \\
&= \int_0^t \alpha e^{-\beta\tau}\ d\tau\ , \\
P_2(t)e^{(\alpha+\gamma)t} - P_2(0)e^{(\alpha+\gamma)0} &= [-\frac{\alpha}{\beta}e^{-\beta\tau}]_0^t \\
&= -\frac{\alpha}{\beta}(e^{-\beta t} - e^{-\beta 0}) \\
&= \frac{\alpha}{\beta}(1-e^{-\beta t})\ , \\
P_2(t) &= e^{-(\alpha+\gamma)t}\frac{\alpha}{\beta}(1-e^{-\beta t})\ , \\
P_2(t) &= \frac{\alpha}{\beta}e^{-(\alpha+\beta+\gamma)t}(e^{\beta t}-1)\ .
\end{aligned} \tag{4.11}$$

Me løyser vidare (4.9):

$$\frac{d}{dt}P_5(t) = \beta P_1(t) - (\beta+\gamma)P_5(t)\ ,$$
$$\frac{d}{dt}P_5(t) + (\beta+\gamma)P_5(t) = \beta P_1(t)\ .$$

Eg multipliserer også her heile uttrykket med integrerande faktor som er $e^{\int_0^t \beta+\gamma\ d\tau} = e^{(\beta+\gamma)t}$ og får

$$\begin{aligned}
(\frac{d}{dt}P_5(t) + (\beta+\gamma)P_5(t))e^{(\beta+\gamma)t} &= \beta P_1(t)e^{(\beta+\gamma)t}\ , \\
\int_0^t \frac{d}{d\tau}(P_5(\tau)e^{(\beta+\gamma)\tau})\ d\tau &= \int_0^t \beta e^{-(\alpha+\beta+\gamma)\tau}e^{(\beta+\gamma)\tau}\ d\tau \\
&= \int_0^t \beta e^{-\alpha\tau}\ d\tau\ , \\
P_5(t)e^{(\beta+\gamma)t} - P_5(0)e^{(\beta+\gamma)0} &= [-\frac{\beta}{\alpha}e^{-\alpha\tau}]_0^t \\
&= -\frac{\beta}{\alpha}(e^{-\alpha t}-1)\ , \\
P_5(t) &= -\frac{\beta}{\alpha}e^{-(\alpha+\beta+\gamma)t}(1-e^{\alpha t})\ , \\
P_5(t) &= \frac{\beta}{\alpha}e^{-(\alpha+\beta+\gamma)t}(e^{\alpha t}-1)\ . \qquad (4.12)
\end{aligned}$$

Eg ser at $P_5(t)$ er som $P_2(t)$, berre at α og β har blitt bytta med kvarandre. Dette er naturleg sidan einaste forskjell på sannsynet for å vere i tilstand 5 ved tid t, i høve til tilstand 2 ved tid t, er at α har blitt bytta ut med ein β intensitet.

Løyser så (4.7):

$$\begin{aligned}
\frac{d}{dt}P_3(t) &= \alpha P_2(t) + \beta P_5(t) \\
&= \alpha(\frac{\alpha}{\beta}e^{-(\alpha+\beta+\gamma)t}(e^{\beta t}-1)) + \beta(\frac{\beta}{\alpha}e^{-(\alpha+\beta+\gamma)t}(e^{\alpha t}-1)) \ , \\
\int_0^t \frac{d}{d\tau}P_3(\tau)\ d\tau &= \frac{\alpha^2}{\beta}[\int_0^t e^{-(\alpha+\gamma)\tau}\ d\tau - \int_0^t e^{-(\alpha+\beta+\gamma)\tau}\ d\tau] \\
&\quad + \frac{\beta^2}{\alpha}[\int_0^t e^{-(\beta+\gamma)\tau}\ d\tau - \int_0^t e^{-(\alpha+\beta+\gamma)\tau}\ d\tau] \ , \\
P_3(t) - P_3(0) &= \frac{\alpha^2}{\beta}([-\frac{1}{\alpha+\gamma}e^{-(\alpha+\gamma)\tau}]_0^t - [-\frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)\tau}]_0^t) \\
&\quad + \frac{\beta^2}{\alpha}([-\frac{1}{\beta+\gamma}e^{-(\beta+\gamma)\tau}]_0^t - [-\frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)\tau}]_0^t) \ , \\
P_3(t) &= \frac{\alpha^2\beta+\alpha\beta^2+\alpha^2\gamma+\beta^2\gamma}{(\alpha+\gamma)(\beta+\gamma)(\alpha+\beta+\gamma)} - \frac{\alpha^2}{\beta(\alpha+\gamma)}e^{-(\alpha+\gamma)t} \\
&\quad - \frac{\beta^2}{\alpha(\beta+\gamma)}e^{-(\beta+\gamma)t} + \frac{\alpha^3+\beta^3}{\alpha\beta(\alpha+\beta+\gamma)}e^{-(\alpha+\beta+\gamma)t} \ . \qquad (4.13)
\end{aligned}$$

Så til slutt løyser eg (4.8), sjølv om det eigentleg er er unødvendig sidan svaret ikkje trengs vidare.

$$\begin{aligned}
\frac{d}{dt}P_4(t) &= \gamma(P_1(t) + P_2(t) + P_5(t)) \\
&= \gamma(e^{-(\alpha+\beta+\gamma)t} + \frac{\alpha}{\beta}e^{-(\alpha+\beta+\gamma)t}(e^{\beta t}-1) + \frac{\beta}{\alpha}e^{-(\alpha+\beta+\gamma)t}(e^{\alpha t}-1)) \ , \\
\int_0^t \frac{d}{d\tau}P_4(\tau)\ d\tau &= \int_0^t \gamma e^{-(\alpha+\beta+\gamma)\tau}\ d\tau + \int_0^t \frac{\gamma\alpha}{\beta}e^{-(\alpha+\beta+\gamma)\tau}(e^{\beta\tau}-1)\ d\tau \\
&\quad + \int_0^t \frac{\gamma\beta}{\alpha}e^{-(\alpha+\beta+\gamma)\tau}(e^{\alpha\tau}-1)\ d\tau \ .
\end{aligned}$$

Eg deler dei tre integrala på høgre side av likhetsteiknet inn i I,II og III og løyser dei.

$$\begin{aligned}
I &= \gamma \int_0^t e^{-(\alpha+\beta+\gamma)\tau}\, d\tau = \gamma[-\frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)\tau}]_0^t \\
&= \frac{\gamma}{\alpha+\beta+\gamma}(1-e^{-(\alpha+\beta+\gamma)t})\ . \\
II &= \frac{\gamma\alpha}{\beta}(\int_0^t e^{-(\alpha+\gamma)\tau}\, d\tau - \int_0^t e^{-(\alpha+\beta+\gamma)\tau}\, d\tau) \\
&= \frac{\gamma\alpha}{\beta}([-\frac{1}{\alpha+\gamma}e^{-(\alpha+\gamma)\tau}]_0^t - [-\frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)\tau}]_0^t) \\
&= \frac{\gamma\alpha}{\beta}(-\frac{1}{\alpha+\gamma}e^{-(\alpha+\gamma)t} + \frac{1}{\alpha+\gamma} + \frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)t} - \frac{1}{\alpha+\beta+\gamma})\ . \\
III &= \frac{\gamma\beta}{\alpha}(\int_0^t e^{-(\beta+\gamma)\tau}\, d\tau - \int_0^t e^{-(\alpha+\beta+\gamma)\tau}\, d\tau) \\
&= \frac{\gamma\beta}{\alpha}([-\frac{1}{\beta+\gamma}e^{-(\beta+\gamma)\tau}]_0^t - [-\frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)\tau}]_0^t) \\
&= \frac{\gamma\beta}{\alpha}(-\frac{1}{\beta+\gamma}e^{-(\beta+\gamma)t} + \frac{1}{\beta+\gamma} + \frac{1}{\alpha+\beta+\gamma}e^{-(\alpha+\beta+\gamma)t} - \frac{1}{\alpha+\beta+\gamma})\ . \\
P_4(t) - P_4(0) &= I + II + III\ .
\end{aligned}$$

Dette blir etter litt utregning:

$$\begin{aligned}
P_4(t) = \frac{\gamma^3+2\gamma^2\alpha+2\gamma^2\beta+3\alpha\beta\gamma}{(\alpha+\gamma)(\beta+\gamma)(\alpha+\beta+\gamma)} - \frac{\gamma\alpha}{\beta(\alpha+\gamma)}e^{-(\alpha+\gamma)t} - \frac{\gamma\beta}{\alpha(\beta+\gamma)}e^{-(\beta+\gamma)t} \\
+ \frac{\gamma\beta^2+\gamma\alpha^2-\gamma\alpha\beta}{\alpha\beta(\alpha+\beta+\gamma)}e^{-(\alpha+\beta+\gamma)t}\ . \qquad (4.14)
\end{aligned}$$

4.4 Overlevelsesfunksjon og likelihood funksjon for "den dynamiske modell"

På grunnlag av dei løyste differensiallikningane våre kan me no sette opp overlevelsesfunksjonen (2.1) for denne modellen. Denne finn me enkelt og greitt ved å sette $S(t) = 1 - P_3(t)$. Sidan $P_3(t)$ er sannsynet for å ha fått eit andre barn ved tid t, altså at ein ikkje har overlevd, er $1 - P_3(t)$ sannsynet for at ein har overlevd til tid t som vil sei at ein ikkje har fått eit nytt barn. Dermed er $S(t) = 1 - P_3(t)$, så ved innsetting av (4.13) som me fann i forrige delkapittel får me at

$$\begin{aligned}
S(t) = 1 - \frac{\alpha^2\beta+\alpha\beta^2+\alpha^2\gamma+\beta^2\gamma}{(\alpha+\gamma)(\beta+\gamma)(\alpha+\beta+\gamma)} + \frac{\alpha^2}{\beta(\alpha+\gamma)}e^{-(\alpha+\gamma)t} \\
+ \frac{\beta^2}{\alpha(\beta+\gamma)}e^{-(\beta+\gamma)t} - \frac{\alpha^3+\beta^3}{\alpha\beta(\alpha+\beta+\gamma)}e^{-(\alpha+\beta+\gamma)t}\ . \qquad (4.15)
\end{aligned}$$

Eg har da funne ein overlevelsesfunksjon $S(t)$ for denne modellen som avhenger av tid og kva verdiar me velger for α, β og γ. Ved hjelp av $S(t)$ kan me no også finne ein

likelihood funksjon for "den dynamiske modell". La oss først byrje med å kalle denne for H slik som Aalen har gjort i [1]. Vidare deler me opp tida fra 0 år og opp til det største tidsintervallet mellom første og andre barnefødsel i 6 månaders intervall. Årsaken til at me gjer dette er at me da kan sette opp H som produktet av binomiske sannsyn over alle desse tidsintervalla. Intervall i vil da bestå av tida $[t_{i-1}, t_i)$, der for eksempel intervall 1 da består av $[t_0, t_1)$=[0,0.5). Tidsenheten er i år, slik at vidare blir t_2=1, t_3=1,5 osv. I intervall i vil da det binomiske sannsynet for at akkurat N_i personar skal føde i intervallet $[t_{i-1}, t_i)$, gitt at det er R_i personar i risiko for å føde ved starten av intervallet vere uttrykt som:

$$H(\alpha, \beta, \gamma) = \begin{pmatrix} R_i \\ N_i \end{pmatrix} (1 - P_i)^{R_i - N_i} (P_i)^{N_i} \ . \tag{4.16}$$

Her er P_i sannsynet for å føde i intervall i, gitt at ein ikkje har født fram til dette intervallet. Merk at denne P_i ikkje er den samme som den $P_j(t)$ me har snakka om tidlegare i kapittelet. Når eg no skal bruke overlevelsesfunksjonen vil eg for enkelheits skuld sette $S(t_i) = S_i$. Sannsynet for å ikkje føde i intervall i, gitt at ein ikkje har født fram til dette intervallet vil vere

$$\frac{S_i}{S_{i-1}} = \frac{P(T > t_i)}{P(T > t_{i-1})} = P(T > t_i | T > t_{i-1}) = 1 - P_i \ .$$

Dermed får me at

$$P_i = 1 - \frac{S_i}{S_{i-1}} \ .$$

Set da inn for P_i i (4.16) og tek produktet over alle tidsintervalla for å få dette nye uttrykket for likelihood funksjonen H:

$$H(\alpha, \beta, \gamma) = \prod_{i=1}^{i=N} \begin{pmatrix} R_i \\ N_i \end{pmatrix} (\frac{S_i}{S_{i-1}})^{R_i - N_i} (1 - \frac{S_i}{S_{i-1}})^{N_i} \ . \tag{4.17}$$

Her er N antallet intervall me treng etter kva største intervall mellom barnefødslar er i datasettet me brukar. I "Aalens datasett" er 28 intervallar nok. Altså er den største tida mellom barnefødslar der mellom 13,5 og 14 år.

4.5 Maximum Likelihood estimering av α, β og γ

Eg vil no utføre maximum likelihood estimering av α, β og γ for likelihood funksjonen (4.17) som me fann i forrige delkapittel på "det simulerte datasett". Dette gjer me fordi me vil bruke maximum likelihood estimata(MLE) av α, β og γ til å tilpasse ei phase type fordeling og sjå kor godt denne passar med data. MLE av α, β og γ vil vere dei intensitetane som passar best i "den dynamiske modell" vist i Figur 4.1 for det datasettet me ser på. Dette fordi at det er desse estimata som maksimerer sannsynet for at andre fødslane i datasettet vårt skjer akkurat når dei har skjedd. Sidan eg som nemnt ikkje har

"Aalens datasett" har eg simulert eit nytt datasett kalla "det simulerte datasett". Me har allerede brukt dette datasettet før i oppgåva, men har no fått tilstrekkelig informasjon slik at simuleringa kan forklarast. Eg simulerte "det simulerte datasett" med å bruke dei MLE Aalen fann for α, β og γ når han brukte sitt datasett i [1]. Desse er gitt i Tabell 4.1 saman med dei tilhøyrande standardavvika som og er henta fra [1].

Tabell 4.1: MLE for fødselsintervall fra "Aalens datasett"

Parametrar	α	β	γ
Estimat	0.190	0.822	0.0476
Standardavvik	0.058	0.015	0.0046

Eg har da brukt desse verdiane som intensitetar i "den dynamiske modell", og simulerer tidene som mødre brukar gjennom desse tilstandane. Eg treng da overgangssannsyna for "den dynamiske modell". Sannsynet for å hoppe til ein av dei forskjellige tilstandane i modellen gitt at det skjer eit hopp blir

$$P(\text{Hopp fra 1 til 2 i dt}|\text{Hopp i dt}) = \frac{P(\text{Hopp fra 1 til 2 i dt})}{P(\text{Hopp i dt})} = \frac{\alpha dt + o(dt)}{(\alpha+\beta+\gamma)dt + o(dt)}$$
$$= \frac{\alpha}{\alpha+\beta+\gamma}\frac{(1+\frac{1}{\alpha}\frac{o(dt)}{dt})}{(1+\frac{1}{\alpha+\beta+\gamma}\frac{o(dt)}{dt})} = \frac{\alpha}{\alpha+\beta+\gamma}(1+\frac{o(dt)}{dt}) ,$$

der $\frac{o(dt)}{dt} \to 0$. Så

$$P_{12} = P(\text{Hopp fra 1 til 2 i dt}|\text{Hopp i dt}) = \frac{\alpha}{\alpha+\beta+\gamma} .$$

På nett samme måten finn me dei andre overganssannsyna som da blir

$$P_{15} = \frac{\beta}{\alpha+\beta+\gamma} .$$
$$P_{14} = \frac{\gamma}{\alpha+\beta+\gamma} .$$
$$P_{23} = \frac{\alpha}{\alpha+\gamma} .$$
$$P_{24} = \frac{\gamma}{\alpha+\gamma} .$$
$$P_{53} = \frac{\beta}{\beta+\gamma} .$$
$$P_{54} = \frac{\gamma}{\beta+\gamma} .$$

Overgangssannsyna samt bruk av intensitetane i Tabell 4.1 gjev meg alt eg treng for å lage "det simulerte datasett" som no vil vere ganske likt "Aalens datasett". Dermed kan ein forvente å få liknande resultat som Aalen fekk i [1], når ein bruker "det simulerte datasett". (Sjå tillegg A for kode av simuleringa.) Vidare vil eg no utføre maximum

likelihood estimering av α, β og γ ved å bruke H på "det simulerte datasett". Sidan å maksimere H med hensyn på α, β og γ er det samme som å maksimere $\log H$ med hensyn på α, β og γ, brukar eg $\log H$ som er ein lettare funksjon å handtere. Eg startar da med å finne $\log H$ og får rett fram uttrykket under ved å ta logaritmen av (4.17):

$$\log H(\alpha,\beta,\gamma) = \sum_{i=1}^{n}(\log \begin{pmatrix} R_i \\ N_i \end{pmatrix} + (R_i - N_i)\log(\frac{S(i)}{S(i-1)}) + N_i \log(1 - \frac{S(i)}{S(i-1)})) \ . \tag{4.18}$$

Maksimerer så $\log H$ ved å bruke minimeringsmetodar i R på $-\log H$. Når ein minimerer -$\log H$ blir $\log H$ maksimert, og dermed får me MLE'ane til α, β og γ ved å gjere dette. (Sjå tillegg B for kode av maksimeringen av $\log H$ i R.)

Eg har køyrt koden i tillegg A 10 gongar for å få simulert 10 ulike datasett. Storleiken på kvar av desse valgte eg til 1000. Storleiken er mindre enn i "Aalens datasett", men det vil ikkje gjere resultatet annleis. Altså vil det ved tid 0 vere 1000 mødre som har fått eit barn som er i risiko for å få eit nytt barn. Etter dette har eg gjort maksimeringen av H i tillegg B med hensyn på kvar av desse 10 datasetta. Så tok eg gjennomsnittet av dei 10 MLE'ane eg fann for kvar parameter. Dette fordi at eg ville prøve å kvitte meg med noko av den naturlige variasjonen som vil bli for kvar simulering, samtidig som dette gjev meg ein moglegheit til å finne det empiriske standardavviket for kvar parameter. Da fekk eg gjennomsnittleg MLE for α, β og γ gitt i Tabell 4.2.

Tabell 4.2: Gjennomsnittleg MLE for fødselsintervall fra "det simulerte datasett"

Parametrar	α	β	γ
Estimat	0.175	0.8153	0.0448
Standardavvik 1	0.0487	0.0224	0.0036
Standardavvik 2	0.0516	0.0249	0.0038

Altså får eg omtrent dei samme estimata som Aalen fekk når han maksimerte H med sitt datasett. Dermed tyder det på at dei metodane eg har brukt i R fungerer godt, og at "det simulerte datasett" er liknande "Aalens datasett".

I Tabell 4.2 ser me at det er oppgitt to ulike standardavvik for dei gjennomsnittlege MLE'ane me har funne. Me ser at det er godt samsvar mellom dei to ulike estimata for standardavvika. Dette tyder på at begge metodane for å finne desse estimata fungerer godt. Standardavvik 1 har eg funne ved å bruke formelen for det empiriske standardavviket s som er

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \ ,$$

der x_i er ein av dei 10 MLE'ane me fann for anten α, β eller γ. Jamfør [3] side 82. $\overline{x}$ er den tilhøyrande gjennomsnittlege MLE'en.

Standardavvik 2 er funne ved å bruke log-likelihood funksjonen $\log H$. Dette er den metoden me skal bruke for å finne standardavvik i resten av oppgåva. Derfor skal eg forklare denne metoden nærare. Ein fin eigenskap ved log-likelihood funksjonen er at ein funksjon av den andre deriverte av log-likelihood funksjonen kan bli brukt til å estimere variansen til fordelinga for datasettet me ser på. Spesifikt må me ta inversen av den negative forventningsverdien til den andre deriverte av log-likelihood funksjonen. Matematisk kan dette uttrykkes som

$$I(\theta)^{-1} = (-E(\frac{\partial^2 LL}{\partial\theta\partial\theta^T}))^{-1} ,$$

der θ er parameteren eller vektoren av parametrar og $I(\theta)$ er informasjonsmatrisa. Kvadratrota av dei diagonale elementa av denne matrisa er standardavvika. Jamfør Lynch [4], side 39-40. Eg har brukt numerisk estimerte andre deriverte av log-likelihood funksjonen. Ved å bruke denne formelen i R har eg funne standardavvik 2, samt standardavvik i resten av oppgåva. Sidan denne metoden er veldig rett fram å bruke har eg ikkje lagt til R koden for dette.

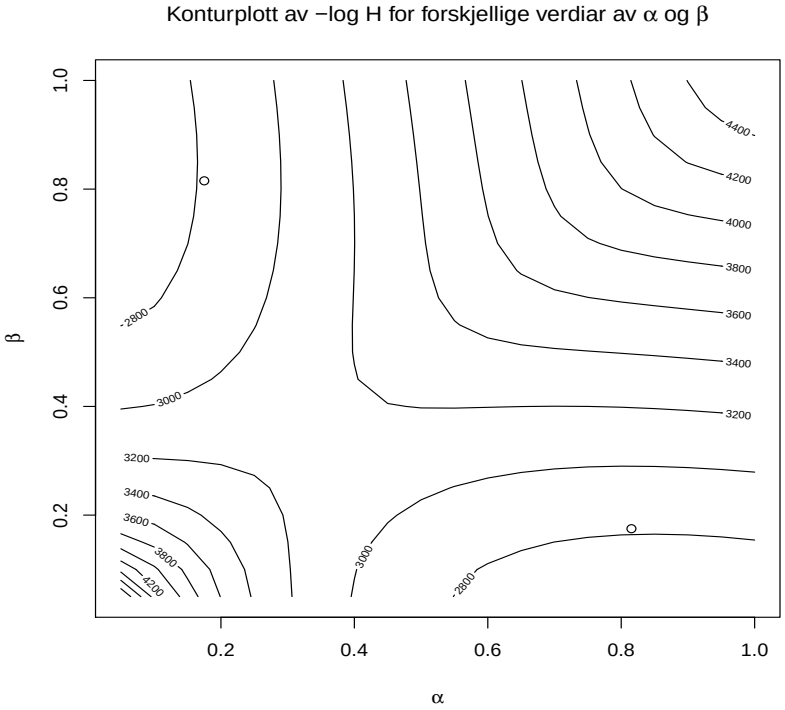
For å sjå grafisk på kva verdiar som gjev størst H, plottar eg eit konturplott av $-\log H$. Der $-\log H$ er minst, vil da H vere størst. Sjå Figur 4.2. Ser av denne figuren at det er to små områder der H er maksimert. Det eine området ligg omtrent ved dei verdiane me har funne for intensitetane α og β ved maximum likelihood estimering(sjå punkta i figuren), samt eit område til der verdiane for α og β er bytta om. Ein ser tydeleg fra dette plottet at funksjonen er symmetrisk om α og β. Har satt γ konstant lik 0.0448 i denne figuren som er estimatet me fann for γ i Tabell 4.2.

No vil me vidare sjå på om den estimerte hendelsesraten som blir funnen og den observerte hendelsesraten ser ut til å passe saman. Finner først den observerte hendelsesraten ved å dele antall fødslar i eit intervall på 6 månadar på antall som er i risiko for å føde ved byrjinga av intervallet. Dette er gjort i tillegg B. Så når me har den observerte hendelsesraten vil me estimere hendelsesraten me får fra modellen. Dette gjer me ved å bruke relasjon (2.5), der me set inn (4.15) og dens deriverte slik at me får den estimerte hasardraten $\widehat{\alpha}(t)$:

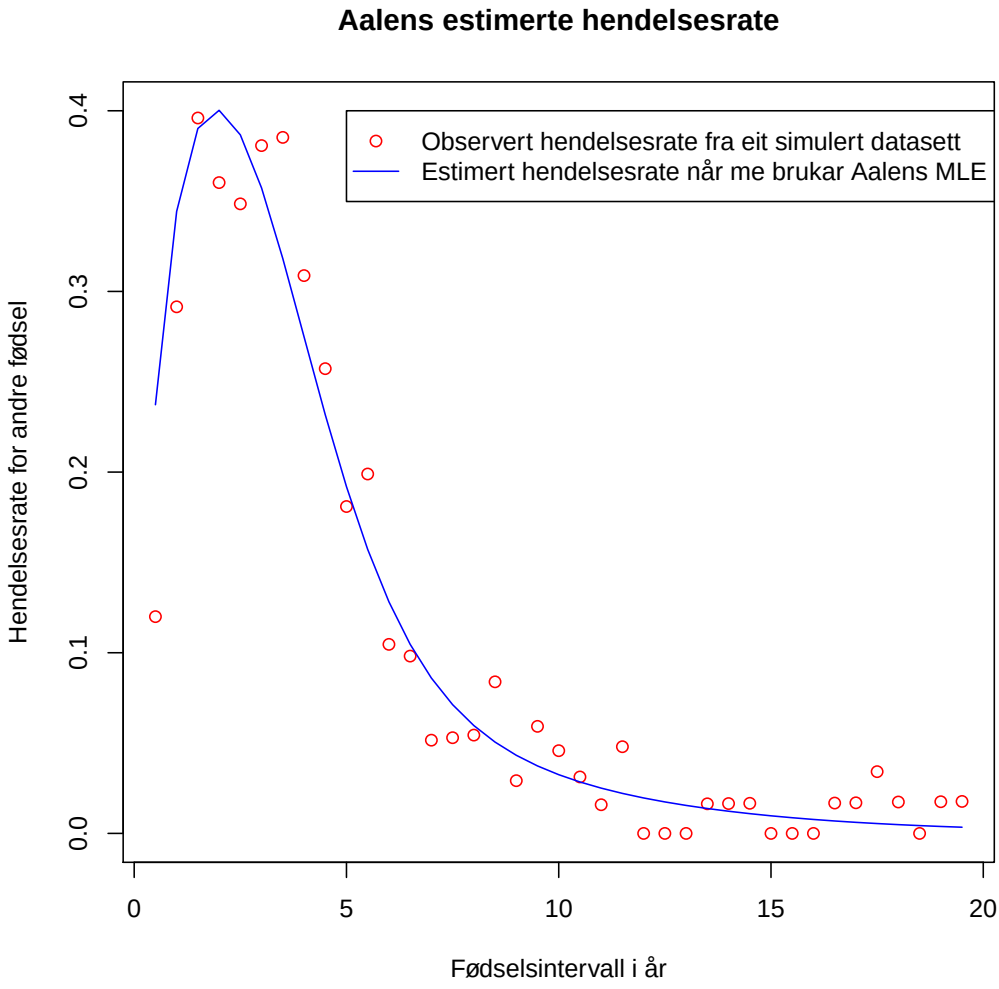
$$\begin{aligned}\widehat{\alpha}(t) &= -\frac{\frac{d}{dt}S(t)}{S(t)} \\ &= -\frac{-\frac{\alpha^2}{\beta}e^{-(\alpha+\gamma)t} - \frac{\beta^2}{\alpha}e^{-(\beta+\gamma)t} + \frac{\alpha^3+\beta^3}{\alpha\beta}e^{-(\alpha+\beta+\gamma)t}}{1 - \frac{\alpha^2\beta+\alpha\beta^2+\alpha^2\gamma+\beta^2\gamma}{(\alpha+\gamma)(\beta+\gamma)(\alpha+\beta+\gamma)} + \frac{\alpha^2}{\beta(\alpha+\gamma)}e^{-(\alpha+\gamma)t} + \frac{\beta^2}{\alpha(\beta+\gamma)}e^{-(\beta+\gamma)t} - \frac{\alpha^3+\beta^3}{\alpha\beta(\alpha+\beta+\gamma)}e^{-(\alpha+\beta+\gamma)t}} .\end{aligned} \tag{4.19}$$

Overlevelsesfunksjonen brukar da Aalen sine MLE for α, β og γ, samt at den får inn ein vektor med forskjellige tider. Me får da den estimerte hendelsesraten når me gjev inn ein tidvektor og MLE'ane fra Tabell 4.1 i (4.19). Plottar denne estimerte hendelsesraten

som ei kurve og ser korleis den passar med den observerte hendelsesraten. Sjå Figur 4.3 på neste side. Ser av denne grafen at den estimerte hendelsesraten passar godt med den observerte. Altså virkar "den dynamiske modell" å vere ein god modell for "det simulerte datasett". Eg vil her nemne at både dei observerte hendelsesrate punkta og den estimerte hendelsesrate kurva i Figur 4.3 er dobbelt så stor som det Aalen fekk i Figur 12 i [1]. Her er det Aalen som har gjort ein liten feil, da han skulle ha gitt hendelsesratane sine per år og ikkje per halve år slik som det er gjort i figuren hans. Når me no veit at ratane i figuren hans skulle vore dobbelt så store ser me at Figur 4.3 passar godt. Me fann omtrent samme MLE som Aalen fekk i [1] og derfor har eg brukt dei MLE han fann gitt i Tabell 4.1 til å putte inn i (4.19), for så å plotte denne $\widehat{\alpha}(t)$ for forskjellige tider. Det at eg fekk omtrent samme estimat som Aalen tyder på at eg har korrekte framgangsmetodar, og gjev oss også eit inntrykk av at estimeringen av MLE'ane er god. Med denne tryggheten kan me da gå vidare til neste kapittel der me vil sjå på korleis "den dynamiske modell" vil passe på "bokas datasett".



Figur 4.2: Dei 2 punkta i plottet er punkta for dei gjennomsnittlege MLE'ane til α og β. Me ser at desse 2 punkta har havna så vidt utanfor minimumsområdet for -log H. Dette er fordi konturplottet er laga for eit av dei simulerte datasetta, og dermed vil ikkje punktet for dei gjennomsnittlege MLE'ane passe heilt perfekt for akkurat dette simulerte datasettet.



Figur 4.3: Me ser av figuren at den estimerte hendelsesraten Aalen fann passar meget godt med punkta for den observerte hendelsesraten. Den observerte hendelsesraten er her laga med bruk av eitt av dei simulerte datasetta. Kan sjå vekk i fra korleis kurva ser ut før tid lik 1 år. Det er berre ei fortsettelse av kurva som ikkje vil vere korrekt.

5

Tilpassing av ein "Phase Type Distribution" ved å bruke "den dynamiske modell" på "bokas datasett"

I dette kapittelet vil eg nytte meg av "bokas datasett" som er nemnt i innleiinga. Altså vil eg no sjå på korleis det fungerer å bruke "den dynamiske modell" på dette datasettet, samt om den phase type fordelinga me finn vil passe godt med dei observerte hendelsesratane. Eg har i kapittel 4 funne funksjonane for likelihood (4.17), overlevelse (4.15) og hasard (4.19). Me kan difor berre gå rett igang med maximum likelihood estimering av likelihood funksjonen H med bruk av "bokas datasett". (For kode av dette sjå tillegg C.)

Eg fann da ved bruk av forskjellige minimeringsmetodar to resultat for kva MLE av α, β og γ kunne bli. Desse resultata er gjevne i Tabell 5.1.

Tabell 5.1: 2 forskjellige MLE for fødselsintervall fra "bokas datasett"

Parametrar	α	β	γ
Estimat	0.259	0.259	0.0234
Standardavvik	0.0098	0.0098	0.0019
Estimat	0.364	0.0000	0.0311
Standardavvik	0.0024	0.0057	0.0034

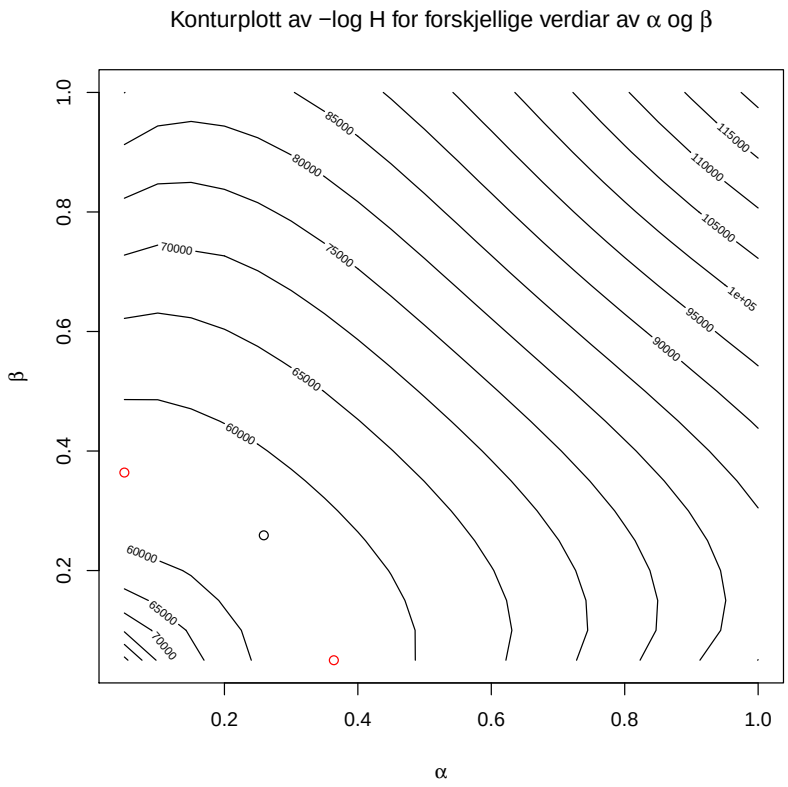
For dei to estimata i Tabell 5.1, blei $\log H$ litt større i det siste estimatet. Sidan me vil ha $\log H$ størst mogleg, er da det siste resultatet litt betre enn det første. Men det er så liten forskjell på desse 2 funksjonsverdiane at me kan ta med oss begge resultata vidare. Eg får altså ut heilt andre MLE for "bokas datasett" i forhold til det ein fekk fra "Aalens datasett". Altså kan det virke som om fødselsmønsteret for mødre som allereie har eit barn har endra seg mellom datasetta. For det siste resultatet har det ingen betyding om me byttar om på α og β. Altså kunne me like gjerne satt resultatet som $\alpha = 0$, $\beta = 0.364$ og $\gamma = 0.0311$. Dette gjev akkurat den samme funksjonsverdien for $\log H$. Eg kunne og sjølvsagt ha bytta om α og β for det første resultatet vårt, sidan desse intensitetane er like. Me ser utifra Tabell 5.1 at det eine resultatet velger samme

intensitet for begge vegar å nå tilstand 3 på, medan det andre resultatet kuttar den eine av vegane å nå tilstand 3. "Den dynamiske modell" er laga for at dei to vegane å nå tilstand 3 på skal framstille ein kjapp og ein treigare måte å nå andre fødsel. Når me da får resultat som anten ignorerer den eine vegen å gå på, eller set begge vegane med lik intensitet, tyder dette på at me ikkje klarar å skilje mødrene for å ta hensyn til heterogenitet.

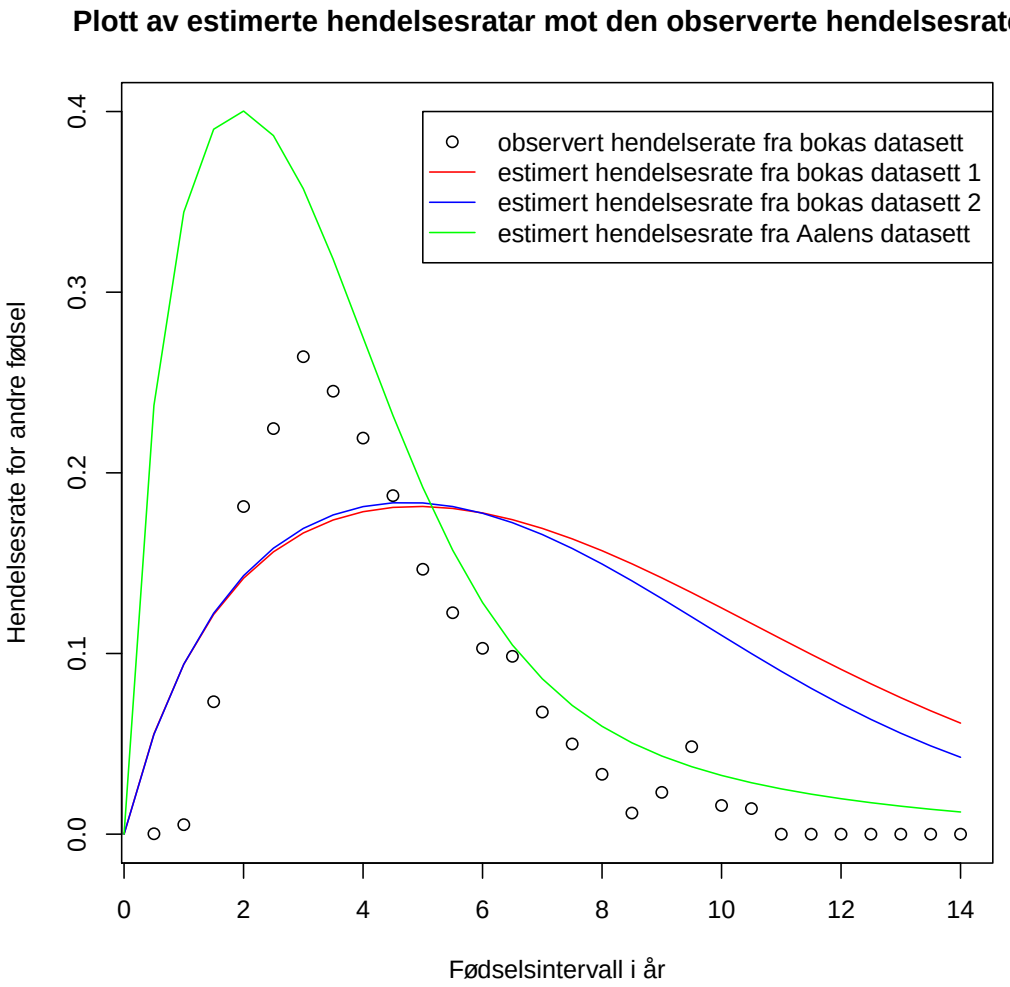
Fordi eg får 2 punkt som gjev maksimum av $\log H$ har eg i Figur 5.1 neste side laga eit konturplott av -$\log H$. (Sjå tillegg C.) Dette konturplottet er laga med forskjellige verdiar av α og β når $\gamma = 0.0234$. At me for enkelheits skuld set $\gamma = 0.0234$ er fordi at dette er det eine av estimata me fann for γ i Tabell 5.1. Me ser utifra Figur 5.1 at det er eit ganske stort område der -$\log H$ er på sitt minste, og at verdiane av α og β som me har funne ved våre minimeringar er innanfor dette område. Det virkar som at -$\log H$ er konstant i dette "dalsøkket" fra (0,0.4) til (0.4,0). Altså ser det ut som det ikkje berre treng vere eitt, men at fleire punkt kan vere minimum for -$\log H$. Me hugsar at dette er det samme som maksimum av H. Dette stemmer godt med at me har funne fleire forskjellige maksimumspunkt for H. Det er altså ikkje eitt punkt $-\log H$ konvergerar mot slik som i "det simulerte datasett". Me ser og utifra konturplottet vårt at likelihooden er symmetrisk i α og β.

Me kan vidare sjå på korleis vår estimerte hendelsesrate med dei parametrane me har funne vil passa med den observerte hendelsesraten. Måten å finne desse to på er akkurat den samme som i forrige kapittel, og koden er lagt til i tillegg C. Plottar så desse mot kvarandre og får Figur 5.2 på side 32. Har óg tatt med den estimerte hendelsesraten som er funne fra "Aalens datasett". Her ser eg at verken denne eller dei estimerte hendelsesratar fra "bokas datasett" passar godt med den observerte hendelsesraten. At den estimerte hendelserate fra "Aalens datasett" ikkje passar her var forventa sidan MLE'ane ikkje stemmer for dette datasettet. At våre to estimerte hendelsesrate kurver fra "bokas datasett" som er veldig like, heller ikkje passar godt, tyder på at "den dynamiske modell" ikkje passar med "bokas datasett" slik som den gjorde for "Aalens datasett".

Fordi "den dynamiske modell" ikkje klarte å tilpasse ei phase type fordeling med "bokas datasett", vil eg i neste kapittel forsøke å lage ein ny modell. Eg har forsøkt å lage fleire modellar der eg både har hatt ein og to vegar å nå tilstanden for andre fødsel på. Den modellen som gav best resultat er den eg no vil gjennomgå.



Figur 5.1: Maksimum av H er der me finn minimumsområdet i dette konturplottet. γ er satt lik 0.0234. Dei røde punkta er dei 2 symmetriske punkta me får fra siste resultat for estimat av α og β i Tabell 5.1. Det svarte punktet er punktet ein får fra det første resultatet i Tabell 5.1. Me ser at alle desse punkta er i minimumsområdet for -log H.



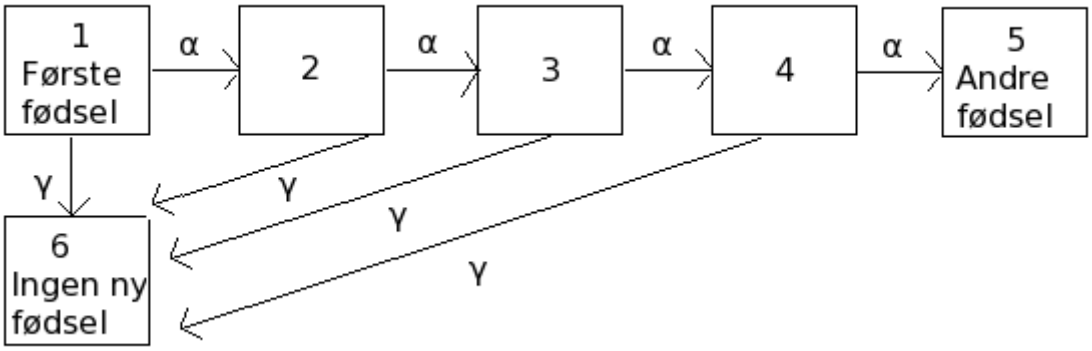
Figur 5.2: Den røde linja er hasardkurva for α og β lik 0.259 og γ lik 0.0234. Den blåe linja er hasardkurva for α lik 0.364, β lik 0 og γ lik 0.0311. Den grøne linja er funne ved å bruke dei estimata Aalen fann for "Aalens datasett". Kan her også sjå vekk i fra korleis kurvene ser ut før tid lik 1 år. Det er berre ei fortsettelse av kurvene som ikkje vil vere korrekt.

6

Tilpassing av ein "Phase Type Distribution" ved å bruke "den alternative dynamiske modell" på "bokas datasett"

6.1 "Den alternative dynamiske modell"

Etter å ha prøvd ut fleire nye modellar for å tilpasse ei phase type fordeling med den observerte hendelsesraten, fann eg at ein 6 tilstandsmodell med kun ein veg å nå tilstand 5 på passa godt. (Tilstand 5 er no den tilstanden som representerar andre fødsel.) Eg velger å kalle denne modellen for "den alternative dynamiske modell". Modellen ser da slik ut:



Figur 6.1: Modell for fødselsintervall med 6 tilstandar, kalla "den alternative dynamiske modell".

Det er fleire årsaker til at eg har valgt akkurat denne 6-tilstandsmodellen. Ein er at resultata i kapittel 5 viste oss at me ikkje greier å skilje mødrene i to grupper slik som Aalen klarte. Dermed er det ikkje noko hensikt i å ha to vegar å nå andre fødsel på. Dette hadde det kun vore behov for viss me klarte å skilje to grupper av mødre fra kvarandre, der den eine gruppa brukte lenger tid enn den andre på å få eit nytt barn. Ein annan årsak er at

eg vil bruke fleire steg mellom første og andre barn enn det eg fekk i forrige modell. Når me lagar fleire steg mellom første og andre fødsel, der alle stega har intensitet α, må α auke for å kompensere for dei ekstra stega. Årsaken til at eg vil oppnå fleire slike steg er at dette vil gje meg ei kurve som får større og smalare topp, noko som utifra Figur 5.2 kan sjå ut til å passe betre med den observerte hendelsesraten. Me kan definere antall steg mellom første og andre fødsel som k.

Forklaringa på at ein større k gjev ein større, samt smalare topp er følgjande; den ubetinga fordelinga ($f(t)$) til tida mellom første og andre barn vil vere ganske lik ei gammafordeling med parameter $\Gamma(k, \alpha) = \Gamma(4, \alpha)$, fordi det er summen av 4 eksponensielle steg i serie med intensitet α. Me veit vidare at forventinga til denne gammafordelinga er lik $\mu = \frac{k}{\alpha}$ og at variansen er $\sigma^2 = \frac{k}{\alpha^2} = \frac{1}{k}(\frac{k}{\alpha})^2 = \frac{1}{k}\mu^2$. Når me da aukar antall eksponesielle steg k, så må også α aukast for å behalde den samme μ. Me ser da at variansen vår σ^2 vil ha den samme μ'en som følgje av dette, medan $\frac{1}{k}$ vil ha minka fordi k har auka. Dette gjev oss altså at kurva vil ha samme forventing, men mindre varians. Dette er noko som fører til ei høgare og smalare kurve. Det er nettopp dette me vil oppnå for kurva til den estimerte hendelsesraten vår sidan den forrige kurva vår var for låg og brei. Sidan den estimerte hendelsesraten er funne ved (2.5) der $f(t) = -\frac{d}{dt}S(t)$, ser me at ein topp i $f(t)$ vil gjenspegla seg i den betinga fordelinga til tida mellom første og andre barn ($\alpha(t)$) dersom $S(t)$ er passe glatt. Dermed håpar me no at det å auka antallet eksponensielle steg har gjeve oss ein modell som betre klarar å tilpasse ei phase type fordeling til data. Modellen har framleis med ein tilstand 6 som representerar det å aldri få eit nytt barn. Da representerar altså tilstand 1 fortsatt å få sitt første barn, tilstand 2, 3 og 4 er tilstandar ein må innom før andre fødsel, og tilstand 5 representerar det å ha fått eit andre barn. Fra alle tilstandane før tilstand 5 er det alltid ein moglegheit å havne i tilstand 6, altså å aldri få eit nytt barn. Eg kan da byrje å utleie differensiallikningane for "den alternative dynamiske modell" i neste delkapittel.

6.2 Utleiing av differensiallikningar

Eg utleier differensiallikningane for denne modellen ved hjelp av Chapman-Kolmogorovs framover-likning (4.4) akkurat slik som me gjorde det i kapittel 4. Har i tillegg at $P_1(0) = 1$ og at $P_k(0) = 0$ for k=2, 3, 4, 5 og 6. Eg startar med tilstand 1:

$$\begin{aligned} P_1(t+h) &= P_1(t)P(X(t+h) = 1|X(t) = 1) \\ &= P_1(t)(1 - (\alpha+\gamma)h + o(h)) \ , \\ \lim_{h\to 0} \frac{P_1(t+h) - P_1(t)}{h} &= -(\alpha+\gamma)P_1(t) \ , \\ \frac{d}{dt}P_1(t) &= -(\alpha+\gamma)P_1(t) \ . \end{aligned} \tag{6.1}$$

For tilstand 2 er situasjonen akkurat den samme som den var for tilstand 2 i "den dynamiske modell", så her setter eg berre rett opp differensiallikninga me fann i kapittel

4:

$$\frac{d}{dt}P_2(t) = \alpha P_1(t) - (\alpha+\gamma)P_2(t)\ . \tag{6.2}$$

Eg utleiier vidare likningane til tilstand 3, 4 og 5.

$$\begin{aligned}
P_3(t+h) &= P_2(t)P(X(t+h)=3|X(t)=2) \\
&\quad + P_3(t)P(X(t+h)=3|X(t)=3) \\
&= P_2(t)(\alpha h + o(h)) + P_3(t)(1-(\alpha+\gamma)h + o(h))\ , \\
\lim_{h\to 0}\frac{P_3(t+h)-P_3(t)}{h} &= \alpha P_2(t) - (\alpha+\gamma)P_3(t)\ , \\
\frac{d}{dt}P_3(t) &= \alpha P_2(t) - (\alpha+\gamma)P_3(t)\ .
\end{aligned} \tag{6.3}$$

$$\begin{aligned}
P_4(t+h) &= P_3(t)P(X(t+h)=4|X(t)=3) \\
&\quad + P_4(t)P(X(t+h)=4|X(t)=4) \\
&= P_3(t)(\alpha h + o(h)) + P_4(t)(1-(\alpha+\gamma)h + o(h))\ , \\
\lim_{h\to 0}\frac{P_4(t+h)-P_4(t)}{h} &= \alpha P_3(t) - (\alpha+\gamma)P_4(t)\ , \\
\frac{d}{dt}P_4(t) &= \alpha P_3(t) - (\alpha+\gamma)P_4(t)\ .
\end{aligned} \tag{6.4}$$

$$\begin{aligned}
P_5(t+h) &= P_4(t)P(X(t+h)=5|X(t)=4) \\
&\quad + P_5(t)P(X(t+h)=5|X(t)=5) \\
&= P_4(t)(\alpha h + o(h)) + P_5(t)(1)\ , \\
\lim_{h\to 0}\frac{P_5(t+h)-P_5(t)}{h} &= \alpha P_4(t)\ , \\
\frac{d}{dt}P_5(t) &= \alpha P_4(t)\ .
\end{aligned} \tag{6.5}$$

Sidan me ikkje treng likningen for tilstand 6 for å finne overlevelsesfunksjonen, er det ikkje nokon årsak til å ta den med. Eg har derfor hoppa over denne og byrjar vidare å løyse differensiallikningane i neste del.

6.3 Løysing av differensiallikningane

Eg startar med å løyse (6.1):

$$\begin{aligned}
\frac{d}{dt}P_1(t) &= -(\alpha+\gamma)P_1(t)\ , \\
\frac{\frac{d}{dt}P_1(t)}{P_1(t)} &= -(\alpha+\gamma)\ , \\
\int_0^t \frac{d}{d\tau}\log(P_1(\tau))\ d\tau &= \int_0^t -(\alpha+\gamma)\ d\tau\ , \\
\log(P_1(t)) - \log(P_1(0)) &= -[(\alpha+\gamma)\tau]_0^t = -(\alpha+\gamma)t\ .
\end{aligned}$$

Sidan $\log P_1(0) = \log(1) = 0$ blir dette

$$P_1(t) = e^{-(\alpha+\gamma)t}\ . \tag{6.6}$$

Eg har da funne løysinga av (6.1). Held deretter fram med å løyse dei 4 siste differensiallikningane, altså løyser me (6.2), (6.3), (6.4) og (6.5).

$$\begin{aligned}
\frac{d}{dt}P_2(t) &= \alpha P_1(t) - (\alpha+\gamma)P_2(t)\ , \\
\frac{d}{dt}P_2(t) + (\alpha+\gamma)P_2(t) &= \alpha P_1(t)\ .
\end{aligned}$$

Eg multipliserer så heile uttrykket med integrerande faktor som er $e^{\int_0^t \alpha+\gamma\ d\tau} = e^{(\alpha+\gamma)t}$.

$$\begin{aligned}
(\frac{d}{dt}P_2(t) + (\alpha+\gamma)P_2(t))e^{(\alpha+\gamma)t} &= \alpha P_1(t)e^{(\alpha+\gamma)t}\ , \\
\int_0^t \frac{d}{d\tau}(P_2(\tau)e^{(\alpha+\gamma)\tau})\ d\tau &= \int_0^t \alpha e^{-(\alpha+\gamma)\tau}e^{(\alpha+\gamma)\tau}\ d\tau \\
&= \int_0^t \alpha\ d\tau\ , \\
P_2(t)e^{(\alpha+\gamma)t} - P_2(0)e^{(\alpha+\gamma)0} &= [\alpha\tau]_0^t = \alpha t\ , \\
P_2(t) &= \alpha t e^{-(\alpha+\gamma)t}\ .
\end{aligned} \tag{6.7}$$

$$\begin{aligned}
\frac{d}{dt}P_3(t) &= \alpha P_2(t) - (\alpha+\gamma)P_3(t)\ , \\
\frac{d}{dt}P_3(t) + (\alpha+\gamma)P_3(t) &= \alpha P_2(t)\ .
\end{aligned}$$

Eg multipliserer med integrerande faktor som fortsatt er $e^{(\alpha+\gamma)t}$.

$$
\begin{aligned}
(\frac{d}{dt}P_3(t) + (\alpha+\gamma)P_3(t))e^{(\alpha+\gamma)t} &= \alpha P_2(t)e^{(\alpha+\gamma)t} \ , \\
\int_0^t \frac{d}{d\tau}(P_3(\tau)e^{(\alpha+\gamma)\tau})\ d\tau &= \int_0^t \alpha(\alpha\tau e^{-(\alpha+\gamma)\tau})e^{(\alpha+\gamma)\tau}\ d\tau \\
&= \int_0^t \alpha^2\tau\ d\tau \ , \\
P_3(t)e^{(\alpha+\gamma)t} - P_3(0)e^{(\alpha+\gamma)0} &= [\frac{\alpha^2\tau^2}{2}]_0^t = \frac{(\alpha t)^2}{2} \ , \\
P_3(t) &= \frac{(\alpha t)^2}{2}e^{-(\alpha+\gamma)t} \ . \qquad (6.8)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{dt}P_4(t) &= \alpha P_3(t) - (\alpha+\gamma)P_4(t) \ , \\
\frac{d}{dt}P_4(t) + (\alpha+\gamma)P_4(t) &= \alpha P_3(t) \ .
\end{aligned}
$$

Eg multipliserer med $e^{(\alpha+\gamma)t}$ på begge sider og får:

$$
\begin{aligned}
(\frac{d}{dt}P_4(t) + (\alpha+\gamma)P_4(t))e^{(\alpha+\gamma)t} &= \alpha P_3(t)e^{(\alpha+\gamma)t} \ , \\
\int_0^t \frac{d}{d\tau}(P_4(\tau)e^{(\alpha+\gamma)\tau})\ d\tau &= \int_0^t \alpha\frac{(\alpha\tau)^2}{2}e^{-(\alpha+\gamma)\tau})e^{(\alpha+\gamma)\tau}\ d\tau \\
&= \int_0^t \frac{\alpha^3\tau^2}{2}\ d\tau \ , \\
P_4(t)e^{(\alpha+\gamma)t} - P_4(0)e^{(\alpha+\gamma)0} &= [\frac{\alpha^3\tau^3}{6}]_0^t = \frac{(\alpha t)^3}{6} \ , \\
P_4(t) &= \frac{\alpha^3}{6}t^3e^{-(\alpha+\gamma)t} \ . \qquad (6.9)
\end{aligned}
$$

$$
\begin{aligned}
\frac{d}{dt}P_5(t) &= \alpha P_4(t) = \alpha\frac{\alpha^3}{6}t^3e^{-(\alpha+\gamma)t} \ , \\
\int_0^t \frac{d}{d\tau}P_5(\tau)\ d\tau &= \int_0^t \frac{\alpha^4}{6}\tau^3e^{-(\alpha+\gamma)\tau}\ d\tau \ , \\
P_5(t) - P_5(0) &= \frac{\alpha^4}{6}\int_0^t \tau^3e^{-(\alpha+\gamma)\tau}\ d\tau \ , \\
\frac{6}{\alpha^4}P_5(t) &= \int_0^t \tau^3e^{-(\alpha+\gamma)\tau}\ d\tau \ . \qquad (I)
\end{aligned}
$$

Eg brukar delvis integrasjon på integralet på høgre side av likning (I) der me set $\tau^3 = u$ og $e^{-(\alpha+\gamma)\tau} = v'$. Formelen me brukar er: $\int uv'\ d\tau = uv - \int u'v\ d\tau$.

$$
\begin{aligned}
u &= \tau^3 \qquad & u' &= 3\tau^2 \ , \\
v' &= e^{-(\alpha+\gamma)\tau} \qquad & v &= -\frac{1}{\alpha+\gamma}e^{-(\alpha+\gamma)\tau} \ ,
\end{aligned}
$$

$$I = uv - \int u'v \, d\tau = [-\tau^3 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)\tau}]_0^t + \frac{3}{\alpha+\gamma} \int_0^t \tau^2 e^{-(\alpha+\gamma)\tau} \, d\tau \; .$$

Her er $v' = \frac{d}{dt}v$ og $u' = \frac{d}{dt}u$. Eg utfører så delvis integrasjon på det gjenstående integralet i siste ledd over, der me no velger $\tau^2 = u$ og $e^{-(\alpha+\gamma)\tau} = v'$.

$$\begin{aligned} u &= \tau^2 \qquad & u' &= 2\tau \; , \\ v' &= e^{-(\alpha+\gamma)\tau} \qquad & v &= -\frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)\tau} \; , \end{aligned}$$

$$\begin{aligned} I &= -\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{uv - \int u'v \, d\tau\} \\ &= -\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{[-\tau^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)\tau}]_0^t + \frac{2}{\alpha+\gamma} \int_0^t \tau e^{-(\alpha+\gamma)\tau} \, d\tau\} \; . \end{aligned}$$

Integralet i siste ledd over løyser me igjen ved delvis integrasjon og får da:

$$\begin{aligned} u &= \tau \qquad & u' &= 1 \; , \\ v' &= e^{-(\alpha+\gamma)\tau} \qquad & v &= -\frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)\tau} \; , \end{aligned}$$

$$\begin{aligned} I &= -\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{-t^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{2}{\alpha+\gamma} \{[-\frac{\tau}{\alpha+\gamma} e^{-(\alpha+\gamma)\tau}]_0^t \\ &+ \frac{1}{\alpha+\gamma} \int_0^t e^{-(\alpha+\gamma)\tau} \, d\tau\}\} \\ &= -\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{-t^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{2}{\alpha+\gamma} \{-\frac{t}{\alpha+\gamma} e^{-(\alpha+\gamma)t} \\ &- \frac{1}{(\alpha+\gamma)^2} [e^{-(\alpha+\gamma)\tau}]_0^t\}\} \\ &= -\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{-t^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{2}{\alpha+\gamma} \{-\frac{t}{\alpha+\gamma} e^{-(\alpha+\gamma)t} \\ &- \frac{1}{(\alpha+\gamma)^2} (e^{-(\alpha+\gamma)t} - 1)\}\} \; . \end{aligned}$$

Dette medfører at $P_5(t)$ blir ved å sette inn det me har funne i likning (I):

$$\begin{aligned} \frac{6}{\alpha^4} P_5(t) = &-\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{-t^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)t} \\ &+ \frac{2}{\alpha+\gamma} \{-\frac{t}{\alpha+\gamma} e^{-(\alpha+\gamma)t} - \frac{1}{(\alpha+\gamma)^2} (e^{-(\alpha+\gamma)t} - 1)\}\} \; , \end{aligned}$$

$$\begin{aligned} P_5(t) = &\frac{\alpha^4}{6} \{-\frac{t^3}{\alpha+\gamma} e^{-(\alpha+\gamma)t} + \frac{3}{\alpha+\gamma} \{-t^2 \frac{1}{\alpha+\gamma} e^{-(\alpha+\gamma)t} \\ &+ \frac{2}{\alpha+\gamma} \{-\frac{t}{\alpha+\gamma} e^{-(\alpha+\gamma)t} - \frac{1}{(\alpha+\gamma)^2} (e^{-(\alpha+\gamma)t} - 1)\}\}\} \; . \end{aligned} \tag{6.10}$$

Ved å gjere litt om på likning (6.10) får eg eit uttrykk for $P_5(t)$ som ser slik ut:

$$P_5(t) = \frac{\alpha^4}{2(\alpha+\gamma)}\{-\frac{t^3}{3}e^{-(\alpha+\gamma)t} - \frac{t^2}{\alpha+\gamma}e^{-(\alpha+\gamma)t} - \frac{2t}{(\alpha+\gamma)^2}e^{-(\alpha+\gamma)t} - \frac{2}{(\alpha+\gamma)^3}e^{-(\alpha+\gamma)t} + \frac{2}{(\alpha+\gamma)^3}\} \,. \tag{6.11}$$

Eg har da løyst alle differensiallikningane me utleidde i forrige delkapittel og funne $P_5(t)$ som er det me har behov for å ta med oss vidare. Ser me nærare på den deriverte av $P_5(t)$ som er det samme som sannsynsfordelinga $f_5(t)$ til tilstand 5, får me at uttrykket ser slik ut:

$$f_5(t) = \frac{d}{dt}P_5(t) = \frac{\alpha^4}{6}t^3e^{-(\alpha+\gamma)t} \,. \tag{6.12}$$

Når me ser på integralet av $f_5(t)$ får me

$$\int_0^\infty f_5(t)\,dt = \int_0^\infty \frac{\alpha^4}{6}t^3e^{-(\alpha+\gamma)t}\,dt$$
$$= (\frac{\alpha}{\alpha+\gamma})^4 \overbrace{\int_0^\infty \frac{(\alpha+\gamma)^4}{\Gamma(4)}t^3e^{-(\alpha+\gamma)t}\,dt}^{I} \,.$$

Der me kan sjå at I er integralet av ein $\Gamma(4, \alpha+\gamma)$ fordeling og er derfor lik 1. Dermed får me at

$$\int_0^\infty f_5(t)\,dt = (\frac{\alpha}{\alpha+\gamma})^4 \,.$$

Sidan integralet av $f_5(t)$ blir lik $(\frac{\alpha}{\alpha+\gamma})^4$ som er mindre enn 1, vil det sei at $f_5(t)$ er ein defekt fordeling. Dette fordi integralet av ein fordeling over heile området sitt skal vere lik 1. Årsaken til at me får ein defekt fordeling er at nokon av dei mødrene me følgjer gjennom tilstandane vil hoppe ned til tilstand 6 istadenfor å følgje vegen til tilstand 5. Dermed kan ikkje integralet av sannsynsfordelinga for å nå tilstand 5 bli lik 1, nettopp fordi det ikkje er alle som vil komme til denne tilstanden.

Etter å ha sett på dette fekk eg ein idé om at det kanskje er ein lettare måte å finne $P_5(t)$ på, utan å måtte utleie og løyse fleire differensiallikningar for å få eit resultat. For å finne $P_5(t)$ på ein annan måte startar eg med å sjå på Laplacetransformasjonen av ein $W \sim \Gamma(4, \alpha+\gamma)$. Her er tettleiken til W: $f(w) = \frac{(\alpha+\gamma)^4}{\Gamma(4)}w^3e^{-(\alpha+\gamma)w}$ som me får bruk for i Laplacetransformasjonen av W:

$$\mathcal{L}_W(u) = E[e^{-uW}] = \int_0^\infty e^{-uw}\frac{(\alpha+\gamma)^4}{\Gamma(4)}w^3e^{-(\alpha+\gamma)w}\,dw$$
$$= \frac{(\alpha+\gamma)^4}{(\alpha+\gamma+u)^4}\overbrace{\int_0^\infty \frac{(\alpha+\gamma+u)^4}{\Gamma(4)}w^3e^{-(\alpha+\gamma+u)w}\,dw}^{1} = (\frac{\alpha+\gamma}{\alpha+\gamma+u})^4 \,. \tag{6.13}$$

Me ser at det siste integralet vårt fra 0 til ∞ vil bli lik 1 sidan det er ei sannsynsfordeling for $\Gamma(4, \alpha+\gamma+u)$ som er inni integralet. Eg har da funne eit uttrykk for Laplacetransformasjonen til ein varabel $W \sim \Gamma(4, \alpha+\gamma)$ i (6.13). Eg ser så på ventetidene i "den alternative dynamiske modell" og definerar dei, samt ein indikatorfunksjon I:

$$V_i = \begin{cases} T_i, & \text{viss hopp til høgre mot andre fødsel,} \\ \infty, & \text{hopp ned til aldri å få eit nytt barn.} \end{cases}$$

$$I = \begin{cases} 1, & \text{viss hopp til høgre mot andre fødsel,} \\ 0, & \text{hopp ned til aldri å få eit nytt barn.} \end{cases}$$

Viss me hoppar eit steg nærare andre fødsel, vil V_i vere lik ventetida i tilstand i (T_i), medan viss me hoppar til tilstand 6 vil V_i vere lik ∞ som vil sei å aldri få eit nytt barn. Me ser no vidare på Laplacetransformasjonen av V_i:

$$\begin{aligned} \mathcal{L}_{V_i}(u) &= E[e^{-uV_i}] = E(E[e^{-uV_i}|I]) \\ &= P(I=0)E[e^{-uV_i}|I=0] + P(I=1)E[e^{-uV_i}|I=1] \\ &= \frac{\gamma}{\alpha+\gamma}e^{-\infty} + \frac{\alpha}{\alpha+\gamma}E[e^{-uT_i}] \\ &= \frac{\alpha}{\alpha+\gamma}E[e^{-uT_i}]\,. \end{aligned} \tag{6.14}$$

Ventetida i tilstand i, T_i, er eksponensielt fordelt med parameter $(\alpha+\gamma)$. Altså er $T_i \sim \exp(\alpha+\gamma)$ og me kan finne eit uttrykk for $E[e^{-uT_i}]$ ved hjelp av at me no veit at $f_{T_i}(t) = (\alpha+\gamma)e^{-(\alpha+\gamma)t}$.

$$\begin{aligned} E[e^{-uT_i}] &= \int_0^\infty e^{-ut}(\alpha+\gamma)e^{-(\alpha+\gamma)t}\,dt \\ &= (\alpha+\gamma)\int_0^\infty e^{-(\alpha+\gamma+u)t}\,dt \\ &= (\alpha+\gamma)[-\frac{1}{\alpha+\gamma+u}e^{-(\alpha+\gamma+u)t}]_0^\infty \\ &= (\alpha+\gamma)(-\frac{1}{\alpha+\gamma+u}e^{-\infty} + \frac{1}{\alpha+\gamma+u}e^{0}) \\ &= \frac{\alpha+\gamma}{\alpha+\gamma+u}\,. \end{aligned}$$

Eg setter da dette uttrykket for $E[e^{-uT_i}]$ inn i (6.14) og får

$$\mathcal{L}_{V_i}(u) = \frac{\alpha}{\alpha+\gamma}\frac{\alpha+\gamma}{\alpha+\gamma+u}\,. \qquad i = 1,2,3,4 \tag{6.15}$$

Eg kallar så den samla tida mellom første og andre fødsel for V, som er summen av ventetidene i tilstandane 1,2,3 og 4. Altså er $V = V_1 + V_2 + V_3 + V_4$. Får da at Laplacetransformasjonen for V blir

$$\mathcal{L}_V(u) = E[e^{-u\sum_{i=1}^4 V_i}] \overset{*}{=} (E[e^{-uV_i}])^4 = (\frac{\alpha}{\alpha+\gamma})^4(\frac{\alpha+\gamma}{\alpha+\gamma+u})^4\,. \tag{6.16}$$

Årsaken til at likhetsteiknet med * over gjeld er at V_i'ane er uavhengige. Hadde me ikkje hatt leddet $(\frac{\alpha}{\alpha+\gamma})^4$ med på høgre side av likhetsteiknet i (6.16), ville me kjent igjen $\mathcal{L}_V(u)$ som Laplacetransformasjonen til ein $\Gamma(4, \alpha+\gamma)$ fordelt variabel. Me ser det ved å sjå tilbake på (6.13). Dette kan me no bruka til å finne $f_5(t)$ for modellen vår ved først å sette opp Laplacetransformasjonen for V som

$$\mathcal{L}_V(u) = E[e^{-uV}] = \int_0^\infty e^{-uv} f(v)\, dv\ .$$

Altså får me fra (6.16) at

$$\int_0^\infty e^{-uv} f(v)\, dv = (\frac{\alpha}{\alpha+\gamma})^4 (\frac{\alpha+\gamma}{\alpha+\gamma+u})^4\ ,$$

$$\int_0^\infty e^{-uv} \overbrace{f(v)(\frac{\alpha+\gamma}{\alpha})^4}^{g(v)\sim\Gamma(4,\alpha+\gamma)}\ dv = (\frac{\alpha+\gamma}{\alpha+\gamma+u})^4\ .$$

Me veit at $g(v) = f(v)(\frac{\alpha+\gamma}{\alpha})^4$ er sannsynsfordelinga til ein $\Gamma(4, \alpha+\gamma)$ fordi me kjenner igjen leddet på høgre side av likningen over som Laplacetransformasjonen til ein $\Gamma(4, \alpha+\gamma)$ fordelt variabel. Kjenner det igjen ved å sjå på (6.13). Ved å sette inn fordelinga til $g(v)$ får me da eit uttrykk for fordelinga til f(v) som er sannsynsfordelinga til tilstand 5 og før i oppgåva blitt notert som $f_5(t)$.

$$\frac{(\alpha+\gamma)^4}{\Gamma(4)} v^3 e^{-(\alpha+\gamma)v} = f(v)(\frac{\alpha+\gamma}{\alpha})^4\ ,$$

$$f(v) = \frac{\alpha^4}{6} v^3 e^{-(\alpha+\gamma)v}\ . \tag{6.17}$$

Når me da brukar notasjonen tidlegare i oppgåva der me brukar t istadenfor v som notasjon for tid, blir (6.17)

$$f_5(t) = \frac{\alpha^4}{6} t^3 e^{-(\alpha+\gamma)t}\ .$$

Dette kjenner me igjen som likning (7.12), og dermed treng me kun å integrere $f_5(t)$ så har me funne $P_5(t)$ utan å utleie og løyse differensiallikningar for modellen. Sidan me no veit at

$$\begin{aligned} P_5(t) &= \frac{\alpha^4}{2(\alpha+\gamma)}\{-\frac{t^3}{3}e^{-(\alpha+\gamma)t} - \frac{t^2}{\alpha+\gamma}e^{-(\alpha+\gamma)t} - \frac{2t}{(\alpha+\gamma)^2}e^{-(\alpha+\gamma)t} \\ &\quad - \frac{2}{(\alpha+\gamma)^3}e^{-(\alpha+\gamma)t} + \frac{2}{(\alpha+\gamma)^3}\} \\ &= \int_0^t f_5(t)\, dt \end{aligned}$$

treng me ikkje å løyse integralet. Hadde eg derimot ikkje funne dette fra før hadde me lett funne $P_5(t)$ ved delvis integrasjon av $f_5(t)$.

Sidan det er $P_5(t)$ som representerar sannsynet for at ein ny fødsel har skjedd ved tid t, blir $1 - P_5(t)$ sannsynet for at ein ny fødsel ikkje har skjedd ved tid t. Dermed blir overlevelsesfunksjonen $S(t)$ i dette tilfellet lik $1 - P_5(t)$, og ved innsetting av likning (6.11) får me:

$$S(t) = 1 - \frac{\alpha^4}{2(\alpha+\gamma)}\{-\frac{t^3}{3}e^{-(\alpha+\gamma)t} - \frac{t^2}{\alpha+\gamma}e^{-(\alpha+\gamma)t} - \frac{2t}{(\alpha+\gamma)^2}e^{-(\alpha+\gamma)t} - \frac{2}{(\alpha+\gamma)^3}e^{-(\alpha+\gamma)t} + \frac{2}{(\alpha+\gamma)^3}\}\,. \tag{6.18}$$

Eg har altså funne likninga for overlevelsesfunksjonen som gjeld for denne modellen, medan likelihood funksjonen H og dens logaritme $\log H$ er dei samme og gitt av likningane (4.17-4.18). Merk berre at me no brukar vår nye $S(t)$ i (4.17-4.18).

6.4 Maximum Likelihood estimering av α og γ

Me har no funne alt som trengs for å byrje maximum likelihood estimeringa av H. Dette har eg gjort i R og koden ligg i tillegg E. MLE'ane eg finn for α og γ er gitt i Tabell 6.1.

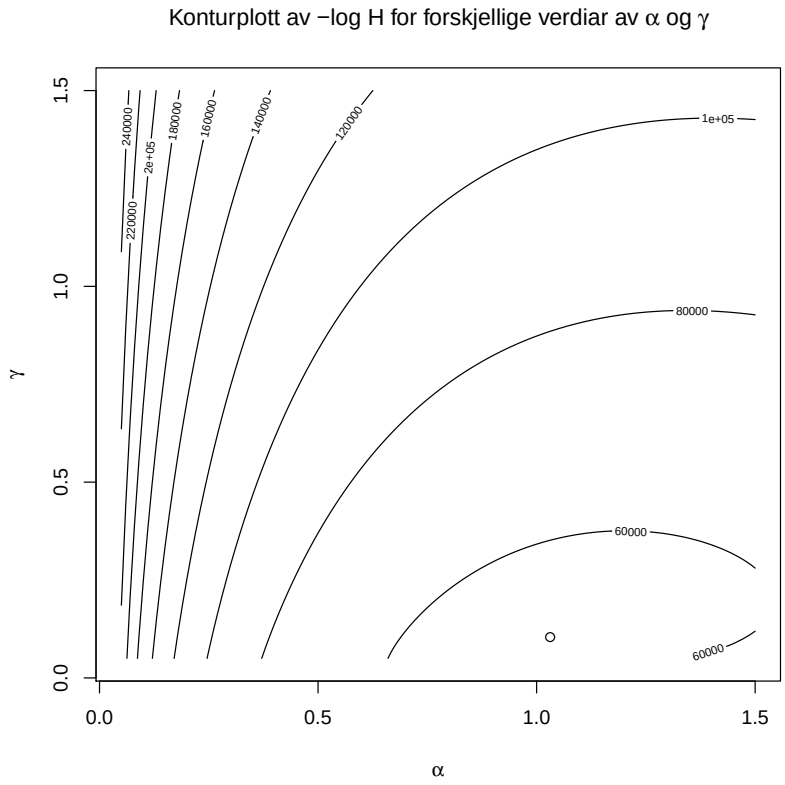
Tabell 6.1: MLE for fødselsintervall fra "bokas datasett"

Parametrar	α	γ
Estimat	1.031	0.1043
Standardavvik	0.0045	0.0020

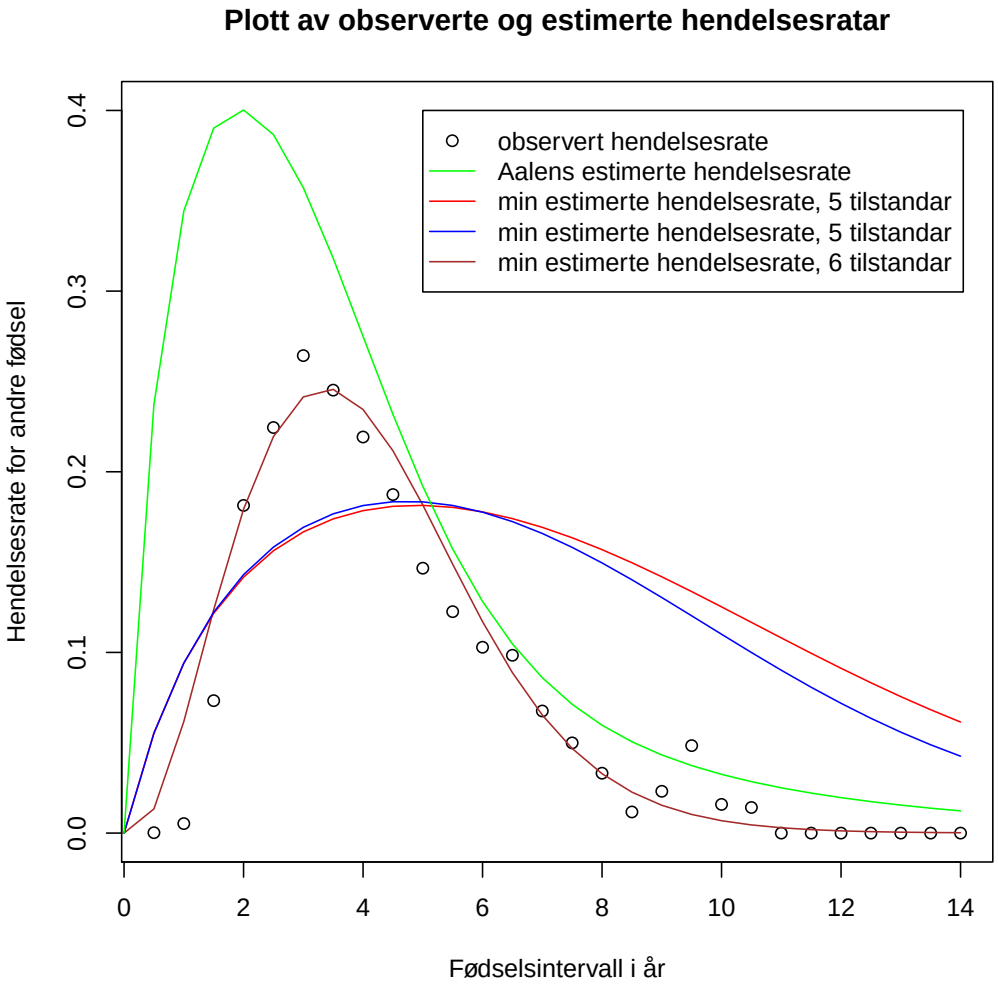
Konturplottet for -log H ser me i Figur 6.2 på neste side.(Sjå tillegg E.) Ser at maksimumspunktet me har funne i Tabell 6.1 er innanfor minimumsområdet i konturplottet. Brukar så desse MLE'ane, (2.5) samt overlevelsesfunksjonen (6.18) saman med sin deriverte for å finne hasardraten for tida mellom første og andre barn. Denne kallar eg her for $\eta(t)$ og vil da sjå slik ut:

$$\begin{aligned}\eta(t) &= -\frac{\frac{d}{dt}S(t)}{S(t)} \\ &= -\frac{-\frac{\alpha^4}{6}t^3e^{-(\alpha+\gamma)t}}{1 - \frac{\alpha^4}{2(\alpha+\gamma)}\{-\frac{t^3}{3}e^{-(\alpha+\gamma)t} - \frac{t^2}{\alpha+\gamma}e^{-(\alpha+\gamma)t} - \frac{2t}{(\alpha+\gamma)^2}e^{-(\alpha+\gamma)t} - \frac{2}{(\alpha+\gamma)^3}e^{-(\alpha+\gamma)t} + \frac{2}{(\alpha+\gamma)^3}\}}\,.\end{aligned} \tag{6.19}$$

Eg må bytte α og γ med dei estimata me har funne i Tabell 6.1 for å få denne estimerte hendelsesraten eller phase type fordelinga som me óg kallar den. Eg vil så sjekke korleis denne passar med den observerte hendelsesraten. Den observerte hendelsesraten finn me på akkurat samme måte som forklart tidlegare og R-koden ligg i tillegg E. Plottar så desse mot kvarandre (i tillegg E) og får Figur 6.3 på side 44. Eg har i denne figuren óg tatt med Aalen sin estimerte hendelsesrate, samt dei to eg fann når eg brukte "den dynamiske modell" på mitt datasett. Eg ser av figuren at "den alternative dynamiske modell" gjev ein estimert hendelsesrate som passar mykje betre med den observerte hendelsesraten.



Figur 6.2: Maksimum av H er der me finn minimumsområdet i dette konturplottet. Ser at punktet for dei MLE'ane eg har funne er innanfor dette området.



Figur 6.3: I denne figuren er punkta den observerte hendelsesraten fra "bokas datasett", den grøne linja er den estimerte hendelsesraten ein får med å bruke Aalen sine MLE fra [1], altså omtrent samme linja som ved å bruke "den dynamiske modell" på "det simulerte datasett". Den raude og den blåe linja er dei to estimerte hendelsesratane eg fann med "den dynamiske modell" på "bokas datasett", og til slutt er den brune linja den estimerte hendelsesraten eg fann med "den alternative dynamiske modell" på "bokas datasett". Ser fortsatt vekk i fra korleis kurvene ser ut før tid lik 1 år.

7 Konklusjon

Formålet med denne oppgåva var å studere tida mellom første og andre barn for "bokas datasett", samt å sjå på om det har blitt nokon endringar i fødselsmønsteret for kvinner fra den tida "Aalens datasett" vart samla inn, og til den tida "bokas datasett" vart samla inn.

Det første eg gjorde var å sjekke at metodane mine fungerte slik som dei skulle i kapittel 4. Dette ved å bruke "den dynamiske modell" på "det simulerte datasett". Når eg da fekk liknande resultat som Aalen i [1], tok eg dette som ein bekreftelse på at mine utrekningar og metodar stemte. Det å simulere eit datasett med gitte verdiar for α, β og γ, for så å sjå om eg får tilbake dei verdiane eg starta med, gjev meg også eit inntrykk av kor god estimeringen er. Startparametrane mine var $\alpha = 0.19$, $\beta = 0.822$ og $\gamma = 0.0476$ når eg simulerte datasettet, sjå Tabell 4.1. Gjennomsnittlege estimerte parametrar basert på 10 simulerte datasett vart $\alpha = 0.175$, $\beta = 0.8153$ og $\gamma = 0.0448$ sjå Tabell 4.2. Dette må seiast å stemme godt med startparametrane for simuleringa som Aalen fann i Tabell 4.1, og dermed virkar estimeringen min å vere god.

Når eg no kjente meg trygg på at eg gjekk fram på rett måte, byrja eg å sjå på "bokas datasett". Resultata eg her fekk med "den dynamiske modell" var totalt ulike fra resultata fra "det simulerte datasett". Altså har det tydeleg skjedd endringar i fødselsmønsteret for mødre som fødte sitt første barn i perioden 1967-71, til mødre som fødte sitt første barn i perioden 1983-1997. Maksimeringa av H konvergerte ikkje mot ein bestemt verdi lenger, og eg fekk fleire moglege løysingar av kva α, β og γ kunne bli. Det som kom fram av desse løysingane for α, β og γ, var at "den dynamiske modell" ikkje klarar å skilje mødrene i to grupper lenger. Eine løysinga ville sende alle mødrene gjennom ein av dei to vegane å nå andre fødsel på, medan andre løysinga ville sende like mange gjennom kvar av vegane for å nå andre fødsel. Dei to vegane å nå andre fødsel er laga for å framstille at det er forskjell på når mødrene vil få sitt andre barn. Dette er tydelegvis ikkje like lett å få fram lenger, fordi det har vorte endringar i datasetta me ser på.

Eg har og i kapittel 5 laga to kurver av den estimerte hendelsesraten me finn for tida mellom første og andre barn, og plotta desse mot den observerte hendelsesraten. Me ser at dei estimerte hendelsesratane ikkje passar med den observerte hendelsesraten i det heile tatt, medan me for "Aalens datasett" fekk ei fordeling som følgde den observerte hendelsesraten godt. Altså har ikkje "den dynamiske modell" vore god for å finne ei phase type fordeling som passar med den observerte hendelsesraten i "bokas datasett".

Fordi at "den dynamiske modell" passa dårleg, laga eg i kapittel 6 min eigen modell for å prøve finne ei phase type fordeling som passar godt for "bokas datasett". Eg kutta ut den eine av vegane å nå andre fødsel på sidan me ikkje klarte å skilje mødrene i to grupper i "bokas datasett". Vidare dobla eg antall steg mellom første og andre barn. Dette resulterte i at eg fekk ein del større intensitet for både α og γ enn det eg fekk i dei ulike resultata fra "den dynamiske modell". At intensiteten for å gå eit steg nærmare andre fødsel måtte aukast når me aukar antall steg mellom fødslane er naturleg. Dersom intensiteten var den samme som i forrige modell ville det ta mykje lenger tid å komme seg gjennom tilstandane. Når eg da plotta fordelinga for tida mellom første og andre barn funne ved bruk av "den alternative dynamiske modell", fekk eg ei kurve som passa mykje betre med den observerte hendelsesraten enn det eg fekk fra "den dynamiske modell". Altså virkar "den alternative dynamiske modell" å vere god for "bokas datasett". Det som er ulempa med min modell, er at me ikkje får framstillt heterogenitet slik som Aalen så fint fekk det i "den dynamiske modell". Det blir også vansklegare å intuitivt forklare fasane mellom første og andre fødsel.

Skal eg nevne noko forslag til vidare arbeid med denne oppgåva må det vere å forsøke å finne ein annan Markovmodell for å tilpasse ei phase type fordeling med den observerte hendelsesraten. Det å finne ein modell som har to vegar å nå andre fødsel på, samtidig som den godt klarar å tilpasse ei phase type fordeling med data ville vore interessant. Da hadde me hatt ein modell som prøvde å framstille heterogeniteten i datasettet slik som Aalen hadde i "den dynamiske modell".

Konklusjonen for denne oppgåva er altså at fødselsmønsteret har endra seg, noko som ikkje er urimeleg sidan det er opptil 30 års forskjell på når datasetta blei samla inn. Hendelsesraten er blitt mindre i "bokas datasett" enn det den var i "Aalens datasett". Dette tyder på at ikkje like mange får eit andre barn i "bokas datasett". Sjølve tidene som går mellom første og andre barn har derimot ikkje endra seg spesielt. Me ser at både den estimerte hendelsesrate kurva fra "Aalens datasett" og fra "bokas datasett" har ein topp ved omtrent samme tid.

Litteratur

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

[2] Odd O. Aalen, Ørnulf Borgan, and Håkon K. Gjessing. *Survival and event history analysis: a process point of view*. Springer, New York, 2008.

[3] Robert V. Hogg and Elliot A. Tanis. *Probability and Statistical Inference*. Pearson-/Prentice Hall, Upper Saddle River, N.J, 2006.

[4] Scott M. Lynch. *Introduction to applied Bayesian statistics and estimation for social scientists*. Springer, New York, 2007.

[5] Hans-Georg Muller and Jane-Ling Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):pp. 61–76, 1994.

[6] Ragnar Norberg. *Basic Life Insurance Mathematics*. Lecture notes, draft version, Copenhagen, 1998.

[7] S original by Kenneth Hess and R port by R. Gentleman. *muhaz: Hazard Function Estimation in Survival Analysis*, 2010. R package version 1.2.5.

[8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[9] Howard M. Taylor and Samuel Karlin. *An introduction to stochastic modeling*. Academic Press, San Diego, Calif., 1998. 3rd ed.

A

Simulering av "det simulerte datasett" i R

Me ynskjer her å bruke R til å simulere eit datasett (kalla "det simulerte datasett") til å bli så likt "Aalens datasett" som mogleg. For å få til dette brukar me dei MLE for α, β og γ som han fekk i [1] som intensitetar når me simulerer. Desse intensitetane er gitt i Tabell 4.1. Me startar med å definere dei 3 intensitetane α, β og γ i R:

```
alpha=0.19; beta=0.822; gamma=0.0476;
```

For enklare notasjon lagar eg og ein variabel som er summen av dei 3 intensitetane:

```
lambda<-alpha+beta+gamma
```

Me kan da starte med simuleringa. Me vil no lage oss eit datasett på 1000 observasjonar. Me vil altså finne samla tid ei mor er i dei forskjellige tilstandane i "den dynamiske modell", før ho kjem seg til tilstand 3. Dette er det samme som tida mellom første og andre barn. Sidan alle byrjar i tilstand 1, simulerer me først tida dei er i denne tilstanden. Denne tida er eksponensielt fordelt med parameter lambda, og me finn derfor desse 1000 tidene i tilstand 1 ved å lage 1000 tider fra denna fordelinga i vektoren `tid1`:

```
tid1<-rexp(1000,lambda)
```

For så å avgjere korleis veg dei 1000 mødrene vil ta etter dette lagar me 1000 variablar mellom 0 og 1 fra den uniforme fordeling:

```
u<-runif(1000)
```

Me simulerer da vegen dei skal gå vidare ved å velge at alle dei uniformt fordelte variablane `u[i]`$\leq \frac{\alpha}{\lambda}$, representerar å gå fra tilstand 1 til tilstand 2. Her er `u` ein vektor med alle dei uniforme variablane, der `u[i]` representerar variabel nr i og i=1,2,.....,1000. Vidare representerar da $\frac{\alpha}{\lambda} <$`u[i]`$\leq \frac{\alpha+\beta}{\lambda}$ at ein går til tilstand 5, medan `u[i]`$> \frac{\alpha+\beta}{\lambda}$ betyr at ein har blitt absorbert i tilstand 4. Grensene me har valgt for kva den uniforme variabelen representerar er henta fra overgangssannsyna me fann i kapittel 4 for "den dynamiske modell". Kan da sjå på koden for simulering av dette første steget:

```
tilstand2=0;tilstand3=0;tilstand4=0;tilstand5=0;
for(i in 1:1000){
 if(u[i]<=alpha/landa2){
   tilstand2=tilstand2+1
 }
 if(u[i]>alpha/landa2 && u[i]<=(alpha+beta)/landa2){
   tilstand5=tilstand5+1
 }
 if(u[i]>(alpha+beta)/landa2){
 tilstand4=tilstand4+1
 }
}
```

Her har me definert variablar for dei ulike tilstandane, der desse blir oppdatert til å inneholde kor mange som har gått til kvar tilstand. Når me no veit kor mange som har gått til kvar tilstand, finn me tidene dei er i den nye tilstanden. Her vil tidene i tilstand 2 vere eksponensielt fordelt med parameter alpha+gamma, tidene i tilstand 5 eksponensielt fordelt med parameter beta+gamma og tidene i tilstand 4 set me veldig høgt. Dette fordi ein er blitt absorbert i tilstand 4, noko som vil sei at ein aldri kjem seg ut av tilstand 4. Eg har valgt å sette tidene her lik 1000. Kan da sjå på koden for desse tidene:

```
tid2<-rexp(tilstand2,(alpha+gamma))
tid5<-rexp(tilstand5,(beta+gamma))
tid4<-rep(1000,tilstand4)
```

No har me altså funne alle tidene i tilstand 1, 2 og 5. Kan da halde fram med å simulere korleis dei som er i tilstand 2 no vil fortsette. Lagar igjen uniforme variablar mellom 0 og 1 som skal bestemme den vidare vegen mødrene i tilstand 2 vil gå. I tilstand 2 er overgangssannsynet til tilstand 3 lik $\frac{\alpha}{\alpha+\gamma}$ og overgangssannsynet til tilstand 4 er lik $\frac{\gamma}{\alpha+\gamma}$. Av den årsak velger me at dei uniforme variablane `u[i]`$\leq \frac{\alpha}{\alpha+\gamma}$ representerar at mor nr i går til tilstand 3. Dei uniforme variablane `u[i]`$>\frac{\alpha}{\alpha+\gamma}$ representerar dermed at mor nr i går til tilstand 4.

```
tilstand3fra2=0; tilstand3fra5=0; tilstand4fra2=0; tilstand4fra5=0;
u<-runif(tilstand2)
for(i in 1:tilstand2){
 if(u[i]<=alpha/(alpha+gamma)){
   tilstand3fra2=tilstand3fra2+1
 }
 if(u[i]>alpha/(alpha+gamma)){
   tilstand4fra2=tilstand4fra2+1
 }
}
```

I første linje av koden over har me definert 4 variablar. Desse fortel oss kor mange som gjekk til tilstand 3 og 4 fra tilstand 2, og kor mange som gjekk til tilstand 3 og 4 fra

tilstand 5. Fortset så med å simulere korleis dei i tilstand 5 vil fortsette. Her gjer me akkurat det samme som me gjorde for dei som var i tilstand 2, berre at me no har nokre andre overgangssannsyn. No er overgangssannsynet fra tilstand 5 til 3 lik $\frac{\beta}{\beta+\gamma}$, og fra tilstand 5 til 4 er den lik $\frac{\gamma}{\beta+\gamma}$.

```
u<-runif(tilstand5)
for(i in 1:tilstand5){
 if(u[i]<=beta/(beta+gamma)){
   tilstand3fra5=tilstand3fra5+1
 }
 if(u[i]>beta/(beta+gamma)){
   tilstand4fra5=tilstand4fra5+1
 }
}
```

Da har me simulert alle dei 1000 mødrene sine ulike vegar å gå gjennom tilstandane på. Dermed gjenstår det berre å legge saman tidene i kvar tilstand for ei mor som går gjennom denne prosessen. Dette gjer me på følgjande måte:

```
T<-rep(0,1000)
for(i in 1:tilstand3fra2){
 T[i]=tid1[i]+tid2[i]
}
for(i in (tilstand3fra2+1):(tilstand3fra2+tilstand4fra2)){
 T[i]=tid1[i]+tid2[i]+1000
}
for(i in (tilstand3fra2+tilstand4fra2+1):
(tilstand3fra2+tilstand4fra2+tilstand3fra5)){
 T[i]=tid1[i]+tid5[i-(tilstand3fra2+tilstand4fra2)]
}
for(i in (tilstand3fra2+tilstand4fra2+tilstand3fra5+1):
(tilstand3fra2+tilstand4fra2+tilstand3fra5+tilstand4fra5)){
 T[i]=tid1[i]+tid5[i-(tilstand3fra2+tilstand4fra2)]+1000
}
for(i in (tilstand3fra2+tilstand4fra2+tilstand3fra5+tilstand4fra5+1):
(tilstand3fra2+tilstand4fra2+tilstand3fra5+tilstand4fra5+tilstand4)){
 T[i]=tid1[i]+1000
}
```

Eg har no funne 1000 tider mellom første og andre fødsel i vektoren `T`, så med dette har me simulert ferdig eit datasett som er liknande "Aalens datasett".

B

Hasardkurver funne ved glatting av Nelson-Aalen aukingane i R

Ynskjer her å vise korleis eg finn hasardkurver for datasetta eg ser på i R ved hjelp av glatting av Nelson-Aalen estimatoren. Eg startar med å lese inn alle data fra fila `second_births.txt` i objektet `secbirth`. Fila `second_births.txt` inneheld "bokas datasett" som eg vil studere i mi oppgåve. Så gjer eg tidene fra `secbirth` om fra å vise antall dagar til antall år i vektoren `aar`. Etter å ha gjort dette brukar eg funksjonen `muhaz` i R til å finne hasardraten for mødrene som fekk eit levande første barn, samt dei som fekk eit dødt første barn. Funksjonen `muhaz` har eg forklart korleis fungerer i kapittel 3.

```
secbirth <- read.table("second_births.txt",header=T)
attach(secbirth)
aar <- time/365
hazfunc <- muhaz(aar[death==0],status[death==0])
hazfunc2<-muhaz(aar[death==1],status[death==1])
```

Eg har da funne hasardraten for mødrene i "bokas datasett" med levande og dødt første barn i henholdsvis `hazfunc` og `hazfunc2`. Vil så finne hasardraten for dei simulerte overlevelsestidene eg fann i tillegg A. Dette er for mødre med levande første barn. Har da ein vektor `T` som inneheld desse overlevelsestidene. Byrjar med å sortere dei i stigande rekkefølgje i vektoren `ordnaT`, før eg fjernar dei sensorerte tidene og får vektoren `usensorertT`. Får da hasardraten for desse overlevelsestidene ved igjen å bruke `muhaz` funksjonen i R på desse usensorerte overlevelsestidene.

```
ordnaT<-T[order(T)]
usensorertT<-subset(ordnaT,ordnaT<1000)
hazfunc3 <- muhaz(usensorertT)
```

Dermed har eg også funne hasardraten for mitt simulerte datasett i objektet hazfunc3. Når me no har funne desse hasardratane kan me plotte desse inn i figur 3.1 i kapittel 3 ved å bruke følgjande kode:

```
plot(hazfunc,xlim=c(1,6),ylim=c(0,0.8),main="Plott av hasardkurver for
dei to datasetta eg har nytta meg av")
lines(hazfunc2,col="red")
lines(hazfunc3,col="blue")
legend(1.5,0.8,c("Hasard rate fra ‘‘bokas datasett’’ der første barn
overlevde","Hasard rate fra ‘‘bokas datasett’’ der første barn døydde",
"Hasard rate fra eit simulert datasett der første barn overlevde"),
col=c("black","red","blue"),lty=c(1,1,1),pch=c(-1,-1,-1))
```

Vil så undersøke nærare hasardkurva der første barn døyr i løpet av sitt første leveår. Dette fordi at hasardkurva for dette i Figur 3.1 ikkje stemmer overeins med hasardkurva for akkurat samme data i Figur 1.1. Lagar da eit objekt kalla `deads` i koden som følgjer der eg har med kun dei mødrene med dødt første barn, og sorterar slik at eg får overlevelsestidene i stigande rekkefølgje. Desse tidene er i vektoren `aar2` der dei er oppgitt med år som tidsenhet.

```
dead <- subset(secbirth,death==1)
detach(secbirth)

attach(dead)
deads <- dead[order(time),]
detach(dead)

attach(deads)
aar2 <- time/365
```

Fortset så med å laga eit objekt kalla `NAest` som inneheld Nelson-Aalen estimatorane for dei observerte hendelsestidene. Her vil hendelsestidene vere gitt som `time` og Nelson-Aalen estimatoren som `hazard` i objektet `NAest`.

```
NAest <- basehaz(coxph(Surv(aar2,status==1)~1))
detach(deads)
attach(NAest)
```

Når me no har funne Nelson-Aalen for ulike tider, kan me starte å glatte aukene i denne estimatoren for å få hasardraten. Lagar først Epanechnikov kjernefunksjonen `Epanech` samt den tilhøyrande boundary funksjonen `Boundary` som me brukar på randområda nær tid 0:

```
Epanech <- function(tid,T,b){
 o<-(tid-T)/b
 if(abs(o)<=1) {(3/4)*(1-(o^2))}
 else {0}
}

Boundary<- function(tid,T,b){
```

```
  o<-(tid-T)/b
  q<-tid/b
  if(abs(o)<=1) {(12/((1+q)^4))*(o+1)*(o*(1-(2*q))+((3*(q^2)-(2*q)+1)/2))}
  else {0}
 }
```

Lagar så ei dobbel løkke som gjev oss heile summasjonsleddet for $\widehat{\alpha}(t)$ gitt av (4.1) for kvar t. Desse summane for kvar t lagrar me i ein vektor som me kallar `teller`. Me har her to av desse doble løkkene der den første brukar mindre bandbreidde og tar for seg punkta nær randområdet. Den andre doble løkka brukar større bandbreidde og er for punkta utanfor randområdet. Koden er som følgjer:

```
teller <- seq(0,0,length.out=49)
b = 0.3
tid <- seq(0,12,by=.25)#Length 49.
for(i in 1:9){
  t=tid[i]
  for(j in 1:length(time)){
    if(t>=b){
      if((t-b)<=time[j] && time[j]<=(t+b)){
if(j==1){teller[i]=teller[i]+((Epanech(t,time[j],b)*(hazard[j])))}
else{teller[i]=teller[i]+((Epanech(t,time[j],b)*(hazard[j]-
        hazard[j-1])))}
      }}

    else{
      if((t-b)<=time[j] && time[j]<=(t+b)){
if(j==1){teller[i]=teller[i]+((Boundary(t,time[j],b)*(hazard[j])))}
else{teller[i]=teller[i]+((Boundary(t,time[j],b)*(hazard[j]-
        hazard[j-1])))}
      }}
    j=j+1
  }
  i=i+1
}

c=2
for(i in 10:49){
  t=tid[i]
  for(j in 1:length(time)){
    if(t>=c){
      if((t-c)<=time[j] && time[j]<=(t+c)){
if(j==1){teller[i]=teller[i]+((Epanech(t,time[j],c)*(hazard[j])))}
else{teller[i]=teller[i]+((Epanech(t,time[j],c)*(hazard[j]-
```

```
        hazard[j-1]))))}
      }
    }

    else{
      if((t-c)<=time[j] && time[j]<=(t+c)){
if(j==1){teller[i]=teller[i]+((Boundary(t,time[j],c)*(hazard[j])))}
else{teller[i]=teller[i]+((Boundary(t,time[j],c)*(hazard[j]-
        hazard[j-1])))}
      }}
    j=j+1
  }
  i=i+1
}
```

Manglar da kun å dele desse summane på båndbreidden `b` som me ser av (3.1). Når me gjer dette får me hasardraten vår i vektoren `esthaz` og plottar denne inn i Figur 3.2 i kapittel 3. Koden er:

```
esthaz<-seq(0,0,length.out=49)
for(i in 1:9){
 esthaz[i] <- teller[i]/b
}
for(i in 10:49){
 esthaz[i] <- teller[i]/c
}

plot(tid,esthaz,main="Plott av hasardkurve for mødre med dødt
første barn",ylim=range(0,.8),xlim=c(1,6),xlab="Fødselsintervall
i år",ylab="Hasardrate for andre fødsel")

lines(tid,esthaz)
```

C

Maximum likelihood estimering av α, β og γ med bruk av "den dynamiske modell" på "det simulerte datasett"

Som nemnt i kapittel 4 ynskjer eg no å maksimere likelihood funksjonen H (4.17) ved å bruke minimeringsmetodar i R på -log H, der log H er gitt av (4.18). Minimerer eg -log H, maksimerer eg log H og dermed også H. Det første eg byrjar med er å ordne vektoren med dei simulerte tidene mine, vektoren `T` slik at den blir i stigande rekkefølgje fra minst til størst. Denne ordna vektoren kan me kalle `ordnaT`. Deretter lagar eg ei delmengde av `ordnaT`, der eg berre har dei som ikkje går til tilstand 4 inkludert. Kan kalla denne for `usensorertT`. Koden for dette blir:

```
ordnaT<-T[order(T)]
usensorertT<-subset(T1,T1<1000)
```

Definerar så vektorane `antfod` og `tid`. Desse gjer me så store som det trengs i forhold til antall intervall ein treng for å få med største tid i `usensorertT`. Antall fødslar i kvart intervall skal da bli satt inn i `antfod`, medan `tid` inneheld alle tidene med 6 månaders intervall fra 0 og opp til endetida for siste intervall. Koden for desse vektorane er som følgjer:

```
antfod<-rep(0,2*(ceiling(usensorertT[length(usensorertT)]*2)/2))
tid<-seq(0,ceiling(usensorertT[length(usensorertT)]*2)/2,by=0.5)
```

Oppdaterer så kor mange som blir født i kvart intervall i `antfod`:

```
for(j in 1:length(tid)){
 for(i in 1:length(usensorertT)){
  if(usensorertT[i]>=tid[j] && usensorertT[i]<tid[j+1]){
   antfod[j]=antfod[j]+1
}}}
```

Eg kan da laga vektoren `antatrisk` som seier kor mange som er i riskiko for å føde ved byrjinga av kvart intervall:

```
vec<-seq(1000,1000,length.out=length(antfod))
antatrisk<-vec-c(0,cumsum(antfod[1:length(antfod)-1]))
```

Me har da alt som trengs for å byrje minimeringen av -log H. Eg lagar funksjonen -log H i R og prøver nokre minimeringsmetodar på denne:

```
-logH <- function(u){
 alpha=u[1]; beta=u[2]; gamma=u[3];
 Survivalv<-1-(((alpha^2)*beta)+(alpha*(beta^2))+((alpha^2)*gamma)+
 ((beta^2)*gamma))/((alpha+gamma)*(beta+gamma)*(alpha+beta+gamma))+
 ((alpha^2)/(beta*(alpha+gamma)))*exp(-(alpha+gamma)*tid)+
 ((beta^2)/(alpha*(beta+gamma)))*exp(-(beta+gamma)*tid)-
 (((alpha^3)+(beta^3))/(alpha*beta*(alpha+beta+gamma)))*
 exp(-(alpha+beta+gamma)*tid)
 L<-seq(0,0,length.out=length(tid))
 for(i in 2:length(tid)){
  L[i]<-(antatrisk[i-1]-antfod[i-1])*log(Survivalv[i]/Survivalv[i-1])+
  (antfod[i-1])*log((1-(Survivalv[i]/Survivalv[i-1])))
 }
 Res <- -(sum(L))
 Res
}
```

Her har eg brukt både `nlm()` og `nlminb()`, der begge fungerte fint. Eg føler ikkje at det er noko behov for å ta med begge sidan dei gjev omtrent samme svar. Eg har derfor koden for `nlminb()` saman med ein vektor med startparametrar under:

```
param<-c(0.15,0.7,0.1)
nlminb(param,-logH,control=list(trace=5,abs.tol=1e-20),lower=c(0,0,0),
upper=c(1,1,1))
```

Har valgt desse parametrane i ein nærleik av det Aalen fekk, sidan eg trur det er i denne nærleiken minimum av -log H ligg. For 10 gjennomkøyringar av simuleringskoden i tillegg A slik at eg får 10 datasett, blir det følgjande resultat av `nlminb()`:

1 simulering: 0.21249490 0.83213260 0.0402109
2 simulering: 0.24998829 0.79755673 0.05152676
3 simulering: 0.17037511 0.81245528 0.04535608
4 simulering: 0.23490117 0.78861723 0.04307491
5 simulering: 0.13564575 0.78833745 0.04996157
6 simulering: 0.17858036 0.82231279 0.04443347
7 simulering: 0.16725421 0.86342578 0.04485773
8 simulering: 0.10641288 0.82151887 0.04511927
9 simulering: 0.18329134 0.82048083 0.04191603
10simulering: 0.10954574 0.80612460 0.04142815

Her er $\widehat{\alpha}$, $\widehat{\beta}$ og $\widehat{\gamma}$ gjeven i henholdsvis 1, 2 og 3 kolonne. Desse verdiane er MLE'ane av α, β og γ. Eg tek gjennomsnittet og standardavviket av desse 10 verdiane for kvar parameter:

```
alph<-c(0.213,0.250,0.170,0.235,0.136,0.179,0.167,0.106,0.183,0.110)
bet<-c(0.832,0.798,0.812,0.789,0.788,0.822,0.863,0.822,0.821,0.806)
gam<-c(0.0402,0.0515,0.0454,0.0431,0.0500,0.0444,0.0449,0.0451,0.0419,
0.0414)
alphabar<-mean(alph)0.175
sdalpha<-sd(alph)0.0487
betabar<-mean(bet)0.8153
sdbeta<-sd(bet)0.0224
gammabar<-mean(gam)0.0448
sdgamma<-sd(gam)0.0036
```

Me får da at gjennomsnittet av MLE over 10 simuleringar blir $\widehat{\alpha} = 0.175$, $\widehat{\beta} = 0.8153$ og $\widehat{\gamma} = 0.0448$. Detta er omtrent dei samme MLE for α, β og γ som Aalen fann. Desse gjennomsnitta har eg gitt i Tabell 4.2. Eg finn også standardavvika til $\widehat{\alpha}$, $\widehat{\beta}$ og $\widehat{\gamma}$ som er henholdsvis 0.0487, 0.0224 og 0.0036. Desse er også gjevne i Tabell 4.2. Desse lignar Aalen sine gjevne i Tabell 4.1. Altså ser det ut som minimeringsmetodane i R fungerer godt sidan eg får nesten likt resultat som Aalen fekk.

For å sjå litt nærmare på kva verdiar av α og β som gjev størst H vil eg lage eit konturplott av -log H. Eg har satt γ lik 0.0448 i dette plottet. Koden for dette er:

```
alphav<-seq(0.05,1,by=0.05)
betav<-seq(0.05,1,by=0.05)
gammav<-rep(0.0448,20)
Hmatrix<-matrix(nrow=20,ncol=20)
for(i in 1:20){
 for(j in 1:20){
  Hmatrix[i,j]<--logH(c(alphav[i],betav[j],0))
}}
contour(alphav,betav,Hmatrix,xlab=expression(alpha),
ylab=expression(beta),main=expression(paste("Konturplott
av -log H for forskjellige verdiar av ",alpha," og ",beta,sep="")))
points(0.175,0.8153)
points(0.8153,0.175)
```

No vil eg vidare sjekke ut om "den dynamiske modell" passar godt med datasettet eg har simulert. Dette gjere eg ved å plotte den observerte hendelsesraten mot den estimerte hendelsesraten eg får ved bruk av Aalen sine MLE. Viss den estimerte hendelsesraten passar godt med den observerte hendelsesrate, kan me sei at "den dynamiske modell" passar godt med "det simulerte datasett". Me kan byrje med å finne den observerte

hendelsesraten. Dette er veldig enkelt no når eg har funne antall fødslar og antall i risiko for kvart intervall i vektorane `antatrisk` og `antfod` over. Deler berre vektoren `antfod` på vektoren `antatrisk` som gjev meg den observerte hendelsesraten pr heile år når eg deler på 0.5:

```
incrat=(antfod/antatrisk)/0.5
```

Eg lagar vidare den estimerte hendelsesraten per år ved å bruke dei MLE'ane for α, β og γ som Aalen fann i Tabell 4.1, `tid` vektoren, (2.5) og (4.15):

```
S<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 1-((((alpha^2)*beta)+(alpha*(beta^2))+((alpha^2)*gamma)+
 ((beta^2)*gamma))/((alpha+gamma)*(beta+gamma)*(alpha+beta+gamma)))+
 (((alpha^2)/(beta*(alpha+gamma)))*exp(-(alpha+gamma)*tid))+
 (((beta^2)/(alpha*(beta+gamma)))*exp(-(beta+gamma)*tid))-((((alpha^3)+
 (beta^3))/(alpha*beta*(alpha+beta+gamma)))*
 exp(-(alpha+beta+gamma)*tid))
}

dS<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 (-(alpha^2/beta)*exp(-(alpha+gamma)*tid))-((beta^2/alpha)*
 exp(-(beta+gamma)*tid))+(((alpha^3+beta^3)/(alpha*beta))*
 exp(-(alpha+beta+gamma)*tid))
}

w<-c(0.19,0.822,0.0476)
haz <- -(dS(w,tid)/S(w,tid))

plot(tid[-1],incrat,col='red',ylim=c(0,0.4),xlab="Fødselsintervall i år"
,ylab="Hendelsesrate for andre fødsel",main="Aalens estimerte
hendelsesrate")
lines(tid[-1],haz2[-1],col='blue')
legend(5,0.4,c("Observert hendelsesrate fra eit simulert datasett",
"Estimert hendelsesrate når me brukar Aalens MLE"),
col=c("red","blue"),lty=c(-1,1),pch=c(1,-1))
```

Funksjonen `dS` over er berre den deriverte av overlevelsesfunksjonen (4.15). Å derivere denne er rett fram, og derfor ikkje vist i oppgåva. Hasardraten per halve år er no i vektoren `haz`. Siste avsnitt i koden over er det som gjev grafen som er vist i Figur 4.3 i kapittel 4.

D

Maximum likelihood estimering av α, β og γ med bruk av "den dynamiske modell" på "bokas datasett"

Eg ynskjer her igjen å maksimere likelihood funksjonen H, som nemnt i kapittel 5. Som sagt før så kan dette gjerast ved å minimere -log H med minimeringsmetodar i R. Har funne log H i (4.18). Me brukar no "bokas datasett" som ligg i fila `second_births.txt`. Startar med å få R til å lese inn datasettet for så å sjekke dimensjonen på det:

```
 secbirth <- read.table("second_births.txt",header=T)
 dim(secbirth)
[1] 53558     5
```

Me ser altså at det er 53 558 førstefødande kvinner med i dette datasettet. Her er både kvinner der det første barnet dør innan sitt første leveår, og kvinner der det første barn overlever talt med. Vidare er det 5 kolonner med informasjon om desse kvinnene. Fra første til femte kolonne er dette henholdsvis informasjon om alder på mora ved første fødsel, kjønn på første barn, om det første barnet døydde i sitt første leveår, tid mellom første og andre fødsel og informasjon om det er ei sensorert eller usensorert tid. I dette tilfellet er me kun interessert i dei mødrene som har fått eit levande første barn. Derfor lagar eg ei delmengde av datasettet `secbirth` som eg kallar `alive`, der kun dei med eit levande første barn tel med:

```
 alive <- subset(secbirth,death==0)
 attach(alive)
 dim(alive)
[1] 53296     5
 live <- alive[order(time),]
 detach(alive)
 attach(live)
 time[53296]
[1] 5070
 unsclive<-subset(live,status==1)
```

```
detach(live)
attach(unsclive)
```

Me ser at det er 53 296 førstefødande som får eit barn som overlev sitt første leveår. Dermed er størrelsen på datasettet vårt no 53 296. Vidare ordnar eg datasettet slik at me får mødrene plassert i stigande rekkefølgje etter tidene mellom første og andre barn. Da vil mora med minst tid mellom første og andre barn vere først i datasettet. Det ordna datasettet kallar eg `live`. Her sjekkar eg kva den største tida mellom første og andre barn er fordi eg vil bruke denne informasjonen til å lage ein tidsvektor etterpå. Me ser at denne tida er 5070 dagar som er lik 13.89 år. Difor treng ikkje tidvektoren min å gå lenger enn til 14 år for å dekke alle tidene i datasettet. Til slutt vil eg ha ei delmengde av datasettet `live` der me fjernar alle dei sensorerte tidene. Altså får me i datasettet `unsclive` berre med dei som faktisk fekk eit andre barn medan data blei samla inn. Dei sensorerte tidene er tider som oppstår når ei mor ikkje blir oppfølgt lenger, utan å ha fått eit nytt barn. Da blir tida frå ho får sitt første barn og til oppfølginga av ho sluttar, det som blir skrive ned i datasettet. Byrjar så med å lage vektoren `antfod` som inneheld kor mange som får sitt andre barn i kvart tidsintervall. Tidsintervalla er som før på 6 månadar og startar ved tid 0.

```
tid<-seq(0,14,by=.5)
n<-dim(unsclive)[1]
aar <- unsclive$time/365
antfod<-seq(0,0,length.out=29)
for(j in 1:(length(tid)-1)){
 for(i in 1:n){
   if(aar[i]>=tid[j] && aar[i]<tid[j+1]){
     antfod[j+1]<-antfod[j+1]+1
}}}
```

Eg vil så finne antall som er i risiko for å føde ved byrjinga av kvart intervall i ein vektor `antatrisk`. Da må me for kvart nye intervall fjerne dei mødrene som har fått eit nytt barn eller blitt sensorert fra antall i risiko. Dermed må eg tilbake til å bruke datasettet `live`, der dei sensorerte tidene også er inkludert. Eg lagar vektoren `antscfod` som inneheld alle som får eit andre barn eller blir sensorert i kvart tidsintervall. Eg tek så alle i risiko ved tid 0 og trekker fra alle som er falle vekk fram til det tidsintervallet me er interessert i. Når me gjer dette for alle tidsintervalla finn me `antatrisk`. Koden er som følgjer:

```
detach(unsclive)
attach(live)
n<-dim(live)[1]
aar <- live$time/365
antscfod<-seq(0,0,length.out=29)
for(j in 1:(length(tid)-1)){
 for(i in 1:n){
   if(aar[i]>=tid[j] && aar[i]<tid[j+1]){
```

```
      antscfod[j+1]<-antscfod[j+1]+1
 }}}

nvec<-seq(n,n,length.out=29)
antatrisk<-nvec-c(0,cumsum(antscfod[2:length(antscfod)]))
```

Eg lagar så funksjonen -log H i R, der me har log H fra (4.18).

```
 -logH <- function(u){
  alpha=u[1]; beta=u[2]; gamma=u[3];
  Survivalv<-1-(((alpha^2)*beta)+(alpha*(beta^2))+((alpha^2)*gamma)+
  ((beta^2)*gamma))/((alpha+gamma)*(beta+gamma)*(alpha+beta+gamma))+
  ((alpha^2)/(beta*(alpha+gamma)))*exp(-(alpha+gamma)*tid)+((beta^2)/
  (alpha*(beta+gamma)))*exp(-(beta+gamma)*tid)-(((alpha^3)+(beta^3))/
  (alpha*beta*(alpha+beta+gamma)))*exp(-(alpha+beta+gamma)*tid)
  L<-seq(0,0,length.out=29)
  for(i in 2:29){
    L[i]<-(antatrisk[i-1]-antfod[i])*log(Survivalv[i]/Survivalv[i-1])+
    (antfod[i])*log((1-(Survivalv[i]/Survivalv[i-1])))
  }
  Res <- -(sum(L))
  Res
 }
```

Eg kan da starte minimeringen av -log H. Har her brukt funksjonane `nlm`, `nlminb` og `optim` som er minimeringsfunksjonar i R. Eg lagar ein vektor `param` med dei startverdiane for α, β og γ som eg vil at minimeringsmetodane skal byrje minimeringen fra.

```
 param<-c(0.4,0.5,0.02)
 typsize=rep(1, length(param))
 nlm(-logH,param,typsize=rep(1, length(param)),stepmax = max(0.1 *
 sqrt(sum((param/typsize)^2)), 1e-2))
 $minimum
[1] 57768.9
 $estimate
[1] 0.25917487 0.25917854 0.02340819
 $gradient
[1] -0.004511094  0.005391485  0.001164153
 $code
[1] 1
 $iterations
[1] 19
```

Vektoren `typesize` er eit estimat av størrelsen på kvar parameter i minimum. `stepmax` er største steg nlm får ta i minimeringen, og det siste talet i funksjonen er minste steglengde

funksjonen får ta. `nlm` funksjonen minimerer H ved å utføre ein type Newton algoritme. Me får da at minimumsverdien til -log H er 57768.9. Her er $\alpha = 0.259$, $\beta = 0.259$ og $\gamma = 0.0234$. Ser at me får kode 1 som output av funksjonen. Dette betyr at relativ gradient er nær null, noko som tyder på at svaret me får sannsynlegvis er løysinga.

Fordi me får 2 ulike løysingar av α, β og γ ved bruk av forskjellige minimeringsmetodar, vil eg no fortsette med minimeringsfunksjonen `optim`:

```
 optim(param,-logH,method="BFGS")
 $par
[1] 3.640402e-01 1.773647e-05 3.105331e-02
 $value
[1] 57689.96
 $counts
function gradient
      97       21
 $convergence
[1] 0
 $message
NULL
```

Metoden BFGS er ein kvasi-Newton metode. Denne brukar funksjonsverdiar og gradientar til å bygge opp eit bilde av overflata som skal bli optimalisert. Får her ein minimumsverdi for -log H på 57689.96, der $\alpha = 0.364$, $\beta = 0$ og $\gamma = 0.0311$. Når eg brukte metoden CG istadenfor BFGS får eg samme svar, berre at α og β har bytta verdi. Altså tyder det på at det ikkje har noko å sei kven av vegane som får α eller β som intensitet. Eg får i output konvergens 0 som tyder på at metoden har fungert godt.

Fordi eg får fleire minimumspunkt for -log H lagar eg eit konturplott for å sjekke om dette kan stemme. I dette plottet lar eg $\gamma = 0.0234$ og ser på forskjellige verdiar av α og γ. Plottet er vist i Figur 5.1 og koden er som følgjer:

```
 alphav<-seq(0.05,1,by=0.05)
 betav<-seq(0.05,1,by=0.05)
 gammav<-rep(0.0234,20)
 Hmatrix<-matrix(nrow=20,ncol=20)
 for(i in 1:20){
  for(j in 1:20){
    Hmatrix[i,j]<--logH(c(alphav[i],betav[j],0))
 }}
 contour(alphav,betav,Hmatrix,xlab=expression(alpha),
 ylab=expression(beta),main=expression(paste("Konturplott
 av -log H for forskjellige verdiar av ",alpha," og ",beta,sep="")))
 points(0.259,0.259)
 points(0.364,0,col="red")
```

```
points(0,0.364,col="red")
```

No vil eg lage eit plott med dei estimerte hendelsesratane mot den observerte hendelsesraten. Eg tar og med Aalens estimerte hendelsesrate i plottet. Eg lagar da tre vektorar som inneheld mine MLE og Aalens MLE, og brukar desse samt likning (2.5) og (4.15) til å finne dei estimerte hendelsesratane. Finn til slutt den observerte hendelsesraten på samme måte som i tillegg C og plottar alle desse i samme graf. Dette er vist i Figur 5.2 og koden er:

```
w<-c(0.25917754,0.25917751,0.02340923)
k<-c(0.364,0.000017,0.031)
p<-c(0.190,0.822,0.0476)

S<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 1-((((alpha^2)*beta)+(alpha*(beta^2))+((alpha^2)*gamma)+((beta^2)
 *gamma))/((alpha+gamma)*(beta+gamma)*(alpha+beta+gamma)))+
 (((alpha^2)/(beta*(alpha+gamma)))*exp(-(alpha+gamma)*tid))+
 (((beta^2)/(alpha*(beta+gamma)))*exp(-(beta+gamma)*tid))-
 ((((alpha^3)+(beta^3))/(alpha*beta*(alpha+beta+gamma)))*
 exp(-(alpha+beta+gamma)*tid))
}

dS<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 (-(alpha^2/beta)*exp(-(alpha+gamma)*tid))-((beta^2/alpha)*exp(-(beta+
 gamma)*tid))+(((alpha^3+beta^3)/(alpha*beta))*exp(-(alpha+beta+gamma)
 *tid))
}

haz1<- -(dS(w,tid)/S(w,tid))
haz2<- -(dS(p,tid)/S(p,tid))
haz3<- -(dS(k,tid)/S(k,tid))

incrat<-seq(0,0,length.out=28)
for(i in 1:length(antatrisk)-1){
 incrat[i]=2*(antfod[i+1]/antatrisk[i])
}

plot(tid[-1],incrat,ylim=c(0,0.4),xlab="Fødselsintervall i år",
ylab="Hendelsesrate for andre fødsel",main="Plott av estimerte
hendelsesratar mot den observerte hendelsesraten")
lines(tid,haz1,col="red")
lines(tid,haz3,col="blue")
```

```
lines(tid,haz2,col="green")
legend(5,0.4,c("observert hendelserate fra boka sitt datasett",
"estimert hendelsesrate fra bokas datasett 1",
"estimert hendelsesrate fra bokas datasett 2",
"estimert hendelsesrate fra Aalens datasett"),col=c("black","red",
"blue","green") ,lty=c(-1,1,1,1),pch=c(1,-1,-1,-1))
```

E

Maximum likelihood estimering av α og γ med bruk av "den alternative dynamiske modell" på "bokas datasett"

Skal igjen maksimere H slik som me gjorde i tillegg D med hensyn på boka sitt datasett. Me leser inn datasettet i R, og finn antall fødslar og antall i risiko for kvart intervall i vektorane `antfod` og `antatrisk`. Sidan koden for dette er akkurat den samme som i tillegg D forklarar eg ikkje koden ein gang til. Sjå tillegg D for forklaring.

```
secbirth <- read.table("second_births.txt",header=T)
alive <- subset(secbirth,death==0)
attach(alive)
live <- alive[order(time),]
detach(alive)
attach(live)
unsclive<-subset(live,status==1)
detach(live)
attach(unsclive)

tid<-seq(0,14,by=.5) #lengde29
n<-dim(unsclive)[1]
aar <- unsclive$time/365
antfod<-seq(0,0,length.out=29)
for(j in 1:(length(tid)-1)){
 for(i in 1:n){
   if(aar[i]>=tid[j] && aar[i]<tid[j+1]){
     antfod[j+1]<-antfod[j+1]+1
}}}

detach(unsclive)
attach(live)#dim er 53296 5
n<-dim(live)[1]
```

```
aar <- live$time/365
antscfod<-seq(0,0,length.out=29)
for(j in 1:(length(tid)-1)){
 for(i in 1:n){
   if(aar[i]>=tid[j] && aar[i]<tid[j+1]){
     antscfod[j+1]<-antscfod[j+1]+1
}}}

nvec<-seq(n,n,length.out=29)
antatrisk<-nvec-c(0,cumsum(antscfod[2:length(antscfod)]))
```

Eg lagar så funksjonen -log H i R, der me har log H fra (4.18) og brukar overlevelsesfunksjonen fra (6.18):

```
-logH <- function(u){
 alpha=u[1]; gamma=u[2];
 ag=alpha+gamma;
 Survivalv<-1-((alpha^4/(2*ag))*(-(tid^3/3)*exp(-ag*tid)-(tid^2/ag)*
 exp(-ag*tid)-((2*tid)/(ag^2))*exp(-ag*tid)-(2/(ag^3))*exp(-ag*tid)+
 (2/(ag^3))))
 L<-seq(0,0,length.out=29)
 for(i in 2:29){
   L[i]<-(antatrisk[i-1]-antfod[i])*log(Survivalv[i]/Survivalv[i-1])+
   (antfod[i])*log((1-(Survivalv[i]/Survivalv[i-1])))
 }
 Res <- -(sum(L))
 Res
}
```

Minimerar så -log H sidan dette gjev oss dei parametrane som maksimerer H. Brukar minimeringsmetoden `nlminb` for dette og koden følgjer under saman med output:

```
nlminb(parameter,-logH,lower=c(0,0))
$par
[1] 1.0305605 0.1043273
$objective
[1] 54980.93
$convergence
[1] 0
$message
[1] "relative convergence (4)"
$iterations
[1] 11
$evaluations
function gradient
      18       34
```

Me får at minimumsverdien av -log H er 54980.93 der parametrane α og β da er lik henholdsvis 1.031 og 0.1043. Får konvergens 0 som tyder på suksessfull konvergens, noko som gjev oss ein indikasjon på at dette sannsynlegvis er løysinga. Eg vil så lage eit konturplott for -log H ved følgjande kode:

```
alphav<-seq(0.05,1.5,by=0.01)
gammav<-seq(0.05,1.5,by=0.01)
-logH(c(alphav[1],gammav[1]))
Hmatrix<-matrix(nrow=146,ncol=146)
for(i in 1:146){
 for(j in 1:146){
   Hmatrix[i,j]<--logH(c(alphav[i],gammav[j]))
}}

contour(alphav,gammav,Hmatrix,xlab=expression(alpha),
ylab=expression(gamma),main=expression(paste("Konturplott
av -log H for forskjellige verdiar av ",alpha," og ",gamma,sep="")))
points(1.031,0.1043)
```

Brukar så desse parametrane som er MLE av α og γ til å finne den estimerte hendelseraten. Relasjon (2.5) gjev oss denne hendelsesraten/hasarden og me lagar derfor funksjonane for $S(t)$ og $\frac{d}{dt}S(t)$. Setter desse inn i (2.5) saman med tidvektoren `tid`, noko som gjev oss den estimerte hendelsesraten for alle tidene i tid. Koden for dette er som følgjer:

```
S<-function(v,tid){
 alpha=v[1]; gamma=v[2]; ag=alpha+gamma;
 1-((alpha^4/(2*ag))*(-(tid^3/3)*exp(-ag*tid)-(tid^2/ag)*exp(-ag*tid)
 -((2*tid)/(ag^2))*exp(-ag*tid)-(2/(ag^3))*exp(-ag*tid)+(2/(ag^3))))
}

dS<-function(v,tid){
 alpha=v[1]; gamma=v[2]; ag=alpha+gamma;
 -(alpha^4/6)*(tid^3)*exp(-ag*tid)
}

a<-c(1.031,0.1043)
haz<- -(dS(a,tid)/S(a,tid))
```

Funksjonen dS over er her den deriverte av overlevelsesfunksjonen $S(t)$. Vil så plotte denne estimerte hendelsesraten me har funne i vektoren `haz` mot den observerte hendelsesraten. Lagar den observerte som før og plottar dei mot kvarandre. Tar og med Aalen sin estimerte hendelsesrate, samt dei to estimerte hendelsesratane eg fann i forrige kapittel i figuren. Koden er:

```
incrat<-seq(0,0,length.out=28)
```

```
for(i in 1:length(antatrisk)-1){
 incrat[i]=2*(antfod[i+1]/antatrisk[i])
}

S1<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 1-((((alpha^2)*beta)+(alpha*(beta^2))+((alpha^2)*gamma)+((beta^2)*
 gamma))/((alpha+gamma)*(beta+gamma)*(alpha+beta+gamma)))+
 (((alpha^2)/(beta*(alpha+gamma)))*exp(-(alpha+gamma)*tid))+
 (((beta^2)/(alpha*(beta+gamma)))*exp(-(beta+gamma)*tid))-
 ((((alpha^3)+(beta^3))/(alpha*beta*(alpha+beta+gamma)))*
 exp(-(alpha+beta+gamma)*tid))
}

dS1<-function(v,tid){
 alpha=v[1]; beta=v[2]; gamma=v[3];
 (-(alpha^2/beta)*exp(-(alpha+gamma)*tid))-((beta^2/alpha)*
 exp(-(beta+gamma)*tid))+(((alpha^3+beta^3)/(alpha*beta))*
 exp(-(alpha+beta+gamma)*tid))
}
w<-c(0.25917754,0.25917751,0.02340923)
k<-c(0.364,0.000017,0.031)
n<-c(0.19,0.822,0.0476)
haz1<- -(dS1(w,tid)/S1(w,tid))
haz2<- -(dS1(k,tid)/S1(k,tid))
haz3<- -(dS1(n,tid)/S1(n,tid))

plot(tid[-1],incrat,ylim=c(0,0.4),main="Plott av observerte
og estimerte hendelsesratar",xlab="Fødselsintervall i år",
ylab="Hendelsesrate for andre fødsel")
lines(tid,haz,col="brown")
lines(tid,haz1,col="red")
lines(tid,haz2,col="blue")
lines(tid,haz3,col="green")
legend(5,0.4,c("observert hendelsesrate","Aalens estimerte
hendelsesrate","min estimerte hendelsesrate, 5 tilstandar",
"min estimerte hendelsesrate,5 tilstandar","min estimerte
hendelsesrate, 6 tilstandar"),col=c("black","green","red",
"blue","brown"),lty=c(-1,1,1,1,1),pch=c(1,-1,-1,-1,-1))
```

Dette gjev Figur 6.3. `S1` og `dS1` er her henholdsvis overlevelsesfunksjonen fra "den dynamiske modell" og dens deriverte.