Frailty-modeller i overlevelsesanalyse

Specialeprojekt - Forår 2018

Mikkel Findinge

Aalborg Universitet
Institut for Matematiske Fag

AALBORG UNIVERSITET

STUDENTERRAPPORT

Titel:
Frailty-modeller i overlevelsesanalyse

Tema:
Frailty-modeller

Projektperiode:
Speciale 2018

Deltagere:
Mikkel Findinge

Vejleder:
Rasmus Waagepetersen

Oplagstal: 2

Sidetal: 60

Afleveringsdato:
8. juni 2018

Abstract:

Dette projekt omhandler frailty-modeller i overlevelsesanalyse. Formålet er at opstille den grundlæggende teori inden for overlevelsesanalyse. Denne teori udvides dernæst til den en-dimensionelle og slutteligt den delte frailty-model. Frailty-variable er latente variable, hvorfor disse kun observeres indirekte gennem data. Dette motiverer brugen af EM-algoritmen til at estimere parametrene i en delt frailty-model. Foruden EM-algoritmen betragtes også PPL-metoden. Disse estimationsmetoder sammenlignes i et mindre simulationsstudie. I denne sammenligning betragtes også den parametriske likelihood-funktion samt den partielle likelihood-funktion, der anvendes til at tilpasse proportional hazard-modeller med og uden specificerede baseline hazard-funktioner. Slutteligt gives et eksempel på anvendelse af frailty-modeller. Dette sker ved brug af datasættet i `R` kaldet `rats`.

Indhold

Forord

Aalborg Universitet, 2018

Der refereres til figurer, tabeller, sætninger, definitioner osv. med tre tegn uden parentes: det første for kapitel, det næste for afsnittet og det sidste for hvilket nummer i afsnittet, der er tale om. Ydermere henvises til ligninger med to tal i parentes. Det første tal angiver kapitlet, og det andet tal angiver det nummer, som ligningen har i kapitlet. Kilder henvises til med forfatter(-ens/-nes) efternavn komma året for udgivelsen omkranset af kantede parenteser. Eventuelle sidetal angives til sidst i henvisningen, hvis der citeres direkte fra kilden. I litteraturlisten står kilderne i alfabetisk rækkefølge efter efternavn.

Mikkel Findinge
<mfindi13@student.aau.dk>

1. Indledning

Overlevelsesanalyse er et statistisk område, der søger at undersøge, hvornår en eller flere begivenheder af interesse indtræffer. Herudover ønskes også forståelse af, hvorfor en begivenhed hænder, når den gør. Til dette kan parametriske modeller hjælpe. Nærværende projekt beskæftiger sig med proportional hazard-modellen og udvidelser af denne. Den udvidelse, der betragtes er frailty-modellen, som indeholder en stokastisk variabel, der skal tage højde eventuel uobserverede variable. Ved introduktion af latente variable bliver estimationsprocessen mindre ligetil. Derfor præsenteres også nogle metoder til at estimere parametrene i en frailty-model.

I Kapitel 2 opbygges den generelle teori bag overlevelsesanalyse. Herunder præsenteres overlevelsesfunktionen, hazard- og den kumulative hazard-funktion samt begrebet censorering. I samme kapitel introduceres proportional hazard-modellen, som er den type parametriske model, der arbejdes med i projektet. Endvidere betragtes Cox proportional hazard-modellen, som ingen antagelser gør om formen af baseline hazard-funktionen i en proportional hazard-model. Derfor udledes den partielle likelihood-funktion, som anvendes til at estimere parametrene i en Cox proportional hazard-model.

Kapitel 3 beskæftiger sig med frailty-modeller, som indeholder stokastiske variable kaldet frailty eller frailty-variable. For disse frailty-variable antages en gamma-fordeling, da denne har nogle pæne egenskaber. I kapitlet behandles også teoretiske eksempler for at vise nogle af disse egenskaber.

Introduktionen af frailty-variable betyder, at de kendte likelihood-funktioner ikke er tilstrækkelige. Derfor har Kapitel 4 til formål at præsentere og behandle metoder, der kan tage højde for frailty. Metoderne, der betragtes, er EM-algoritmen og penalised partial likelihood-metoden (PPL-metoden). EM-algoritmen sættes op til at estimere parametrene i en frailty-model, hvor frailty-variablene er gamma-fordelte. Dette er også tilfældet for PPL-metoden, dog betragtes denne også i forhold til log-normal-fordelte frailty-variable. Metoderne sammenlignes i et mindre simulationsstudie, hvor både en parametrisk proportional hazard-model og en Cox proportional hazard-model også inkluderes.

Rapporten afrundes i Kapitel 5 ved at diskutere nogle af resultaterne i projektet. Derudover gives nogle afsluttende forslag til, hvad der kan arbejdes videre på.

2. Overlevelsesanalyse

I nærværende kapitel vil grundlæggende begreber, definitioner og sætninger inden for overlevelsesanalyse blive præsenteret. Dette kapitel er primært baseret på [Klein og Moeschberger, 2005].

2.1 Grundlæggende funktioner

Lad X være en ikke-negativ stokastisk variabel med tæthedsfunktionen f. Lad ydermere F være den kumulative fordelingsfunktion. Det er nu muligt at definere nogle af de grundlæggende funktioner i overlevelsesanalyse.

Definition 2.1.1 (Overlevelsesfunktionen)
Funktionen, $S\colon \mathbb{R}_+ \to [0,1]$, givet ved

$$S(x) = P(X > x)$$

kaldes *overlevelsesfunktionen.*

Det bemærkes, at $S(x) = 1 - F(x)$. Betragt et kontinuert X, da gælder

$$S(x) = 1 - F(x) = \int_x^\infty f(u)\,du. \tag{2.1}$$

Af 2.1 fremgår det, at

$$\frac{d}{dx}S(x) = -f(x). \tag{2.2}$$

Definition 2.1.2 (Hazard-funktionen)
Hazard-funktionen, $h(x)$, er defineret ved

$$h(x) = \lim_{dx\to 0} \frac{P\left[x \le X < x + dx \,|\, X \ge x\right]}{dx}. \tag{2.3}$$

Bemærk, at X kan anses som værende tiden, hvor en begivenhed af interesse indtræffer. Dermed kan $h(x)dx$ tolkes som sandsynligheden for, at en sådan begivenhed indtræffer i intervallet $[x, x + dx[$ givet, at $X \ge x$. Følgende lemma knytter hazard-, tætheds- og overlevelsesfunktionen.

Lemma 2.1.3
Lad X være kontinuert. Da gælder

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx}\ln S(x). \tag{2.4}$$

Bevis
Første lighed i (2.4) vises ved at anvende Bayes formel på (2.3). Følgende omskrivning fås:

$$h(x) = \lim_{dx\to 0} \frac{P\left[x \leq X < x + dx\right]}{dx\left(P(X > x) + P(X = x)\right)} = \frac{1}{S(x)} \lim_{dx\to 0} \frac{F(x + dx) - F(x)}{dx} = \frac{f(x)}{S(x)}.$$

Dermed kan anden lighed bevises ved at vise, at $\frac{d}{dx} \ln S(x) = -f(x)/S(x)$. Dette fremgår af

$$\frac{d}{dx} \ln S(x) = \frac{1}{S(x)} \frac{d}{dx} S(x) = -\frac{f(x)}{S(x)},$$

hvor (2.1) er anvendt. ■

Definition 2.1.4 (Den kumulative hazard-funktion)
Den kumulative hazard-funktion, $H(x)$, er defineret ved

$$H(x) = \int_0^x h(u)\, du.$$

Det følger af Lemma 2.1.3, at

$$S(x) = \exp\left(-\int_0^x h(u)\, du\right) = \exp(-H(x)). \tag{2.5}$$

Overlevelsesanalyse adskiller sig fra andre grene af statistik ved at tillade brugen af data, hvor en begivenhed af interesse ikke er indtruffet. Teorien bag dette introduceres i følgende afsnit.

2.2 Censorering

Lad $X_1, X_2, \ldots, X_n$ være ikke-negative i.i.d. stokastiske variable. Disse benævnes fremadrettet som *event-tider*. Lad ligeledes $C_1, C_2, \ldots C_n$ være ikke-negative uafhængige og identisk fordelte stokastiske variable. Fremadrettet kaldes disse *censoreringstider*. Event-tid indikerer tiden, hvor en begivenhed af interesse indtræffer. Censoreringstider indikerer derimod tiden, hvor overvågningen af begivenheden af interesse indstilles.

Lad $T_i = \min(X_i, C_i)$ for $i = 1, \ldots, n$. Det er kun parret (T_i, Δ_i), der observeres, hvor

$$\Delta_i = \begin{cases} 1, & \text{hvis } X_i \leq C_i \\ 0, & \text{hvis } X_i > C_i. \end{cases}$$

Hvis $\Delta_i = 0$, siges T_i at være censoreret.

Lad F_X og F_C være kumulative fordelingsfunktioner for henholdsvis event- og censoreringstider. Det antages gennem projektet, at disse er absolut kontinuerte. Ydermere betegnes de tilhørende tæthedsfunktioner for F_X og F_C ved f_X og f_C respektivt.

Sætning 2.2.1
Lad X og C være henholdsvis en event- og en censoreringstid. Lad ydermere X og C være uafhængige. Da er den simultane tæthed for (T, Δ) givet ved

$$f(t, \delta) = (f_X(t)(1 - F_c(t)))^{\delta} (f_C(t)(1 - F_X(t)))^{1-\delta}. \tag{2.6}$$

Bevis
Lad

$$F(t,\delta) = P(T \leq t, \Delta = \delta) = \int_{-\infty}^{t} f(u,\delta)\, du.$$

Da $\Delta \in \{0,1\}$ kan den simultane tæthed skrives på følgende måde:

$$\begin{aligned} f(t,\delta) &= \frac{d}{dt}F(t,\delta) \\ &= \frac{d}{dt}\left(\delta P(T \leq t, \Delta = 1) + (1-\delta)P(T \leq t, \Delta = 0)\right) \\ &= \delta\frac{d}{dt}P(T \leq t, \Delta = 1) + (1-\delta)\frac{d}{dt}P(T \leq t, \Delta = 0) \\ &= \left(\frac{d}{dt}P(T \leq t, \Delta = 1)\right)^{\delta}\left(\frac{d}{dt}P(T \leq t, \Delta = 0)\right)^{1-\delta}. \end{aligned}$$

Betragt nu $P(T \leq t, \Delta = 1)$. Denne kan omskrives på følgende vis

$$\begin{aligned} P(T \leq t, \Delta = 1) &= P(X \leq t, X \leq C) \\ &= \int_{x\leq t}\int_{x\leq c} f_X(x) f_C(c)\, dc\, dx \\ &= \int_{x\leq t} f_X(x) \int_{x\leq c} f_C(c)\, dc\, dx \\ &= \int_{x\leq t} f_X(x)(1 - F_C(x))\, dx. \end{aligned}$$

Dermed er $\frac{d}{dt}P(T \leq t, \Delta = 1) = f_X(t)(1 - F_C(t))$. På tilsvarende vis kan det vises, at $\frac{d}{dt}P(T \leq t, \Delta = 0) = f_C(t)(1 - F_X(t))$, hvilket beviser sætningen. ■

Betragt en endelig-dimensionel parametervektor θ og antag, at fordelingsfunktionen for event-tiderne afhænger af denne. Hvis det antages, at fordelingen af censoreringstider ikke afhænger af θ, kan $f_C(t)$ og $F_C(t)$ i (2.6) betragtes som værende konstanter i likelihood-funktionen. Af antagelsen og (2.6) fås den i'te faktor i likelihood-funktionen

$$L_i(\theta) = f_X(t_i;\theta)^{\delta_i}(1 - F_X(t_i;\theta))^{1-\delta_i} = f_X(t_i;\theta)^{\delta_i} S_X(t_i;\theta)^{1-\delta_i}. \tag{2.7}$$

Ved at anvende Lemma 2.1.3 kan (2.7) omskrives til

$$L_i(\theta) = h_X(t_i;\theta)^{\delta_i} S_X(t_i;\theta).$$

Likelihood-funktionen bliver da

$$L(\theta) = \prod_{i=1}^{n} h_X(t_i;\theta)^{\delta_i} S_X(t_i;\theta) = \prod_{i=1}^{n} h_X(t_i;\theta)^{\delta_i} \exp\left(-\int_0^{t_i} h_X(u;\theta)\, du\right). \tag{2.8}$$

2.3 Proportional hazard-modellen

En nyttig model inden for overlevelsesanalyse er *proportional hazard-modellen*, hvilken tillader indførelsen af kovariater. Lad p være antallet af kovariater. Da benævnes realiseringen af disse ved vektoren $\mathbf{z} \in \mathbb{R}^p$.

Definition 2.3.1 (Proportional hazard-modellen)
Lad $\boldsymbol{\beta} \in \mathbb{R}^p$ være en parametervektor. Da er proportional hazard-modellen givet ved

$$h(t \mid \mathbf{z}) = h_0(t) \exp\left(\mathbf{z}^\top \boldsymbol{\beta}\right),$$

hvor *baseline hazard-funktionen*, $h_0(t)$, er en vilkårlig ikke-negativ funktion.

Betragtes overlevelsesfunktionen for proportional hazard-modellen giver (2.5)

$$S(t \mid \mathbf{z}) = \exp\left(-H_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{z})\right), \tag{2.9}$$

hvor $H_0(t) = \int_0^t h_0(u)\, du$.

Proportional hazard-modellen, der ikke laver antagelser om den eksplicitte form for $h_0(\cdot)$, kaldes *Cox proportional hazard-modellen*. I praksis kendes baseline hazard-funktionen sjældent, hvorfor Cox proportional hazard-modellen har en fordel. Dog kræver likelihood-funktionen normalvis en antagelse om den eksplicitte form af $h_0(\cdot)$ ved anvendelse i parameterestimation. I det følgende introduceres intuitionen bag *den partielle likelihood-funktion*, hvilken benyttes til at estimere parametre i Cox proportional hazard-modellen.

Lad $(t_i, \delta_i, \mathbf{z}_i)$ for $i = 1, \ldots, n$ være realiseringer. Antag, at $t_i \neq t_j$ for $i \neq j$. Benyttes en profil likelihood tilgang kan den partielle likelihood udledes. Dette gøres ved at betragte hazard-funktionen som en uendelig dimensional 'nuisance' parameter og udtrykke denne som en funktion af parameteren af interesse, $\boldsymbol{\beta}$. Anvendes proportional hazard-modellen til data kan likelihood-funktionen i (2.8) omskrives til

$$L(\boldsymbol{\beta}, h_0(\cdot)) = \prod_{i=1}^{n} \left(h_0(t_i)\, dt_i \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)\right)^{\delta_i} \exp\left(-H_0(t_i) \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)\right). \tag{2.10}$$

Lad $D = \{i \mid \delta_i = 1\}$ være indeksmængden af event-tider. Lad ydermere $h_0(t_i)dt_i = a_i > 0$ i et lille interval $[t_i, t_i + dt_i[$ for $i \in D$, samt $h_0(t)dt = 0$ andetsteds. Dermed bidrager kun event-tider til en øget kumuleret hazard-funktion. For at lette notationen lad $a_i = 0$ for $i \notin D$. Den kumulative baseline hazard-funktion er integralet af hazard-funktionen, hvilken kun er ikke-nul på meget små intervaller. Dermed kan integralet approksimeres ved summen

$$H_0^*(t) = \sum_{t_i \leq t} a_i.$$

Denne approksimation kan indsættes i (2.10), hvorved der fås

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(a_i \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)\right)^{\delta_i} \exp\left(-\sum_{t_j \leq t_i} a_j \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)\right) \\ &= \prod_{i=1}^{n} \left(a_i \exp(\boldsymbol{\beta}^\top \mathbf{z}_i) \exp\left(-a_i \sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_j)\right)\right)^{\delta_i}, \end{aligned} \tag{2.11}$$

hvor $R(t) = \{t_i \mid t \leq t_i\}$. Det bemærkes, at likelihood-funktionen ikke direkte afhænger af værdierne af de realiserede tider, men derimod ordenen af disse. Tages logaritmen af (2.11) fås log-likelihood-funktionen

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\log(a_i) + \boldsymbol{\beta}^\top \mathbf{z}_i - a_i \sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_i) \right]. \tag{2.12}$$

Differentieres (2.12) i forhold til a_i fås udtrykket

$$\frac{\delta_i}{a_i} - \delta_i \sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_i).$$

Sættes dette lig med nul, kan a_i isoleres

$$\hat{a}_i = \frac{1}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)}. \tag{2.13}$$

Dette giver et estimat for den kumulative hazard-funktion. Indsættes $\hat{a}_i$ i (2.11), opnås den partielle likelihood-funktion

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\top \mathbf{z}_i)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^\top \mathbf{z}_i)} \right)^{\delta_i}. \tag{2.14}$$

Den partielle likelihood-funktion kan altså ses som en profil likelihood-funktion, hvilken gør det muligt at lave parameterestimationer uden at antage en eksplicit form af $h_0(\cdot)$.

2.4 Fordelinger

I dette afsnit betragtes fordelinger, der gennem projektet anvendes i eksempler, for at give et anvendt perspektiv.

2.4.1 Weibull-fordelingen

I overlevelsesanalyse bliver Weilbull-fordelingen ofte betragtet, da denne er mere fleksibel end fordelinger som eksponential-fordelingen. En stokastisk variabel X følger Weibull-fordelingen, hvis den tilhørende tæthedsfunktion, f_X, er givet ved

$$f_X(x) = \alpha\lambda x^{\alpha-1} \exp\left(-\lambda x^\alpha\right), \quad x > 0, \tag{2.15}$$

hvor α og λ er positive konstanter. Den kumulative fordelingsfunktion er da

$$F_X(x) = \int_0^x \alpha\lambda u^{\alpha-1} \exp\left(-\lambda u^\alpha\right) \, du = 1 - \exp\left(-\lambda x^\alpha\right). \tag{2.16}$$

Dermed giver 2.16 og 2.1 den tilhørende overlevelsesfunktion,

$$S_X(x) = 1 - F_X(x) = \exp\left(-\lambda x^\alpha\right). \tag{2.17}$$

Af (2.5) fås den kumulative hazard-funktion

$$H_X(x) = \lambda x^\alpha. \tag{2.18}$$

Dermed er hazard-funktionen givet ved

$$h_X(x) = \alpha\lambda x^{\alpha-1}. \tag{2.19}$$

Weibull-fordelingen er fleksibel grundet α-parameteren. Hazard-funktionen er aftagende, når $0 < \alpha < 1$, konstant, når $\alpha = 1$, samt voksende, når $\alpha > 1$. Denne fleksibilitet kombineret med de pæne lukkede udtryk for Weibull-funktionerne gør, at denne fordeling er meget anvendt inden for overlevelsesanalyse.

Eksempel 2.4.1
Lad $(t_i, \delta_i, \mathbf{z}_i)$ være realiseringer for $i = 1, \ldots, n$. Betragt proportional hazard-modellen,

$$h(t \mid \mathbf{z}) = h_0(t)\exp(\mathbf{z}^\top \boldsymbol{\beta}),$$

hvor $h_0(t)$ er givet som i (2.19) og $\mathbf{z}, \boldsymbol{\beta} \in \mathbb{R}^p$. Af (2.8) følger den tilhørende likelihood-funktion

$$\begin{aligned} L(\alpha, \lambda, \boldsymbol{\beta}) &= \prod_{i=1}^{n} \left(h_0(t_i)\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right)^{\delta_i} \exp\left(-\int_0^{t_i} h_0(u)\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\, du\right) \\ &= \prod_{i=1}^{n} \left(h_0(t_i)\right)^{\delta_i} \left(\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right)^{\delta_i} \exp\left(-H_0(t_i)\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right) \\ &= \prod_{i=1}^{n} \left(\alpha\lambda t_i^{\alpha-1}\right)^{\delta_i} \left(\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right)^{\delta_i} \exp\left(-\lambda t_i^{\alpha}\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right). \end{aligned}$$

Log-likelihood-funktionen kan da skrives som

$$\ell(\alpha, \lambda, \boldsymbol{\beta}) = D(\log\lambda + \log\alpha) + \sum_{i=1}^{n}\left[\delta_i(\alpha-1)\log(t_i) - \lambda t_i^{\alpha}\exp\left(\mathbf{z}_i^\top \boldsymbol{\beta}\right) + \delta_i \mathbf{z}_i^\top \boldsymbol{\beta}\right],$$

hvor $D = \sum_{i=1}^{n}\delta_i$. Differentieres denne i forhold til parametrene α, λ og β_j for $j = 1, \ldots, p$ fås

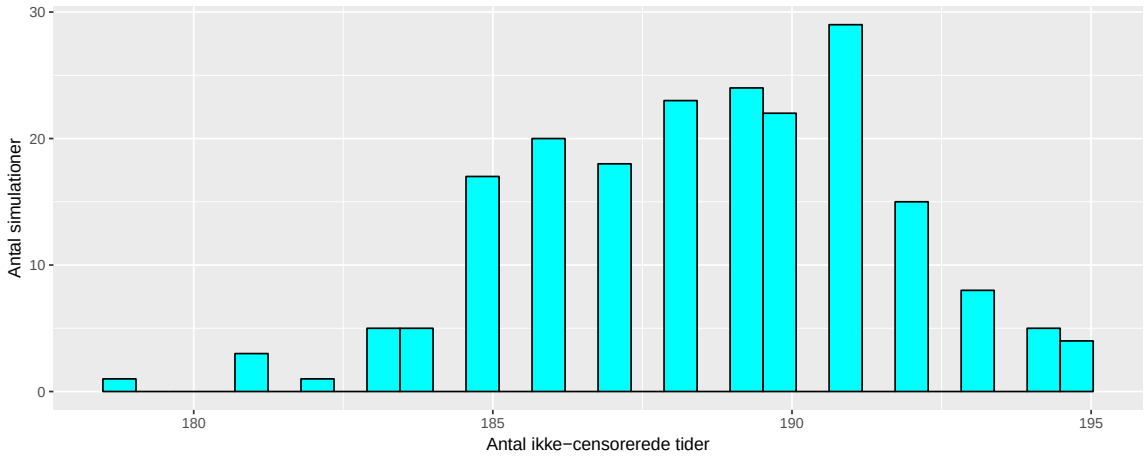
$$\begin{aligned} \frac{\partial}{\partial\alpha}\ell(\alpha, \lambda, \boldsymbol{\beta}) &= \frac{D}{\alpha} + \sum_{i=1}^{n}\left[\delta_i\log(t_i) - \lambda\log(t_i)t_i^{\alpha}\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right], \\ \frac{\partial}{\partial\lambda}\ell(\alpha, \lambda, \boldsymbol{\beta}) &= \frac{D}{\lambda} - \sum_{i=1}^{n} t_i^{\alpha}\exp(\mathbf{z}_i^\top \boldsymbol{\beta}), \text{ og} \\ \frac{\partial}{\partial\beta_j}\ell(\alpha, \lambda, \boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[\delta_i z_{ij} - \lambda z_{ij} t_i^{\alpha}\exp(\mathbf{z}_i^\top \boldsymbol{\beta})\right], \end{aligned}$$

hvor z_{ij} indikerer den j'te indgang i $\mathbf{z}_i$. Ovenstående funktioner sættes lig nul for at estimere parametrene. Løsningen herfor kræver numeriske metoder som Newton-Raphson. De dobbelt afledede til, der anvendes til denne metode, kan findes i Appendiks A.1.

I Appendiks C.1 defineres funktionen `WeibDataSim`, hvilken simulerer binære kovariater og dertilhørende event-tider, som følger proportional hazard-modellen med en specificeret Weibull-hazard-funktion som $h_0(t)$. Teorien bag simuleringen af event-tider, kan findes i Appendiks B. Endvidere kan funktionerne `survreg` og `coxph` i `R`-pakken `Survival` anvendes til at tilpasse henholdsvis en proportional hazard-model og en Cox proportional hazard-model på data.

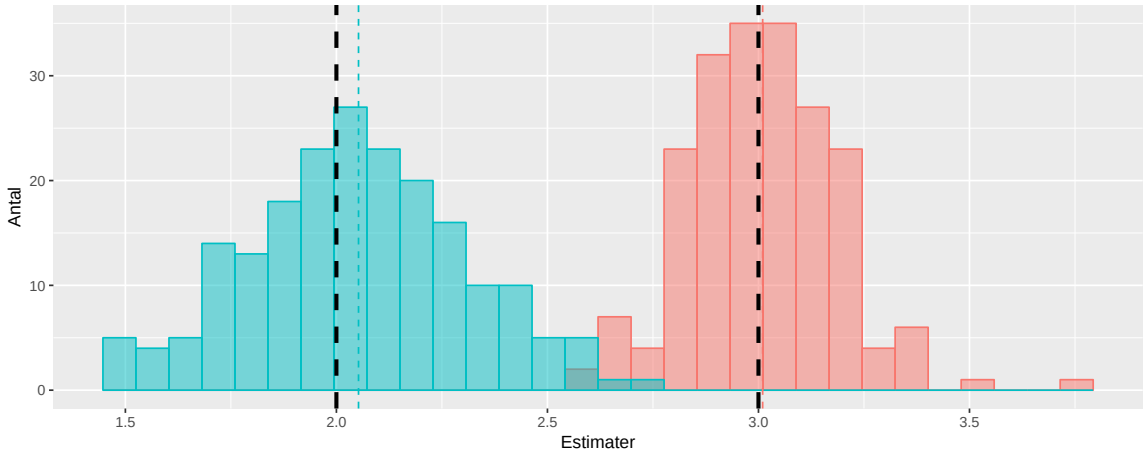
I Appendiks C.2 ses en `R`-kode, der anvender `WeibDataSim` til at simulerere data. Dette sker 200 gange med 200 observationer hver. Data simuleres med Weibull-parametrene `alpha =`

`3` og `lambda = 2`. Endvidere benyttes parameterværdierne `beta1 = 2` og `beta2 = -0.6` som koefficienterne til kovariaterne. Fordelingen, hvorfra censoreringstiderne genereres, er en eksponentialfordeling med 'censoreringsraten', `rateC`, som sættes til 0.1.[1]



Figur 2.4.1: Antallet af simulationer, der har et givent antal event-tider.

I Figur 2.4.1 ses antallet simulationer, der har et givent antal observerede event-tider. Fordelingen af event-tider (ligeledes fordelingen af censoreringstider) betragtes som værende plausible scenarier for virkeligt data. Derfor vil fremtidige simulationer også benytte disse fordelinger.

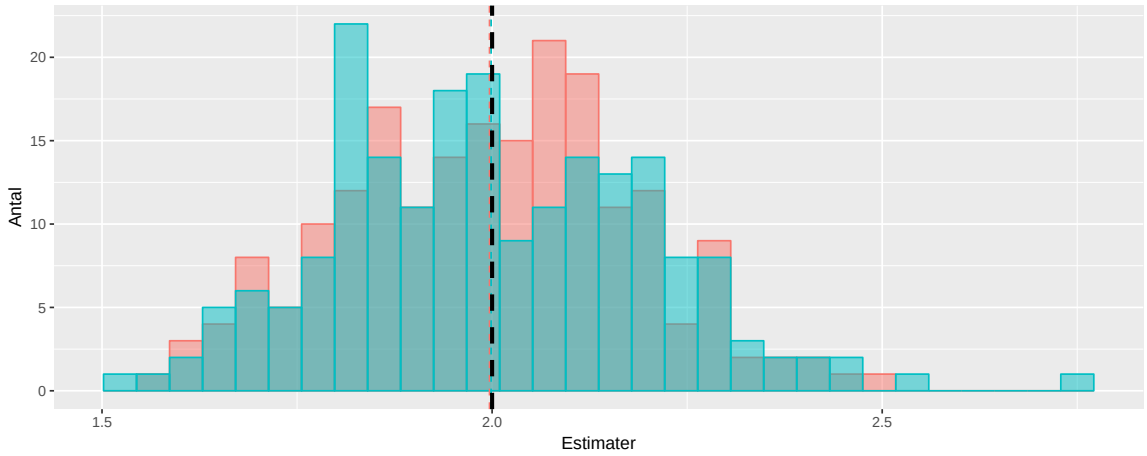


Figur 2.4.2: Estimaterne er for α (rød) og λ (blå). Stiplede linjer indikerer gennemsnitlige estimater. De sorte stiplede linjer indikerer de sande parametre, `alpha=3` og `lambda=2`.

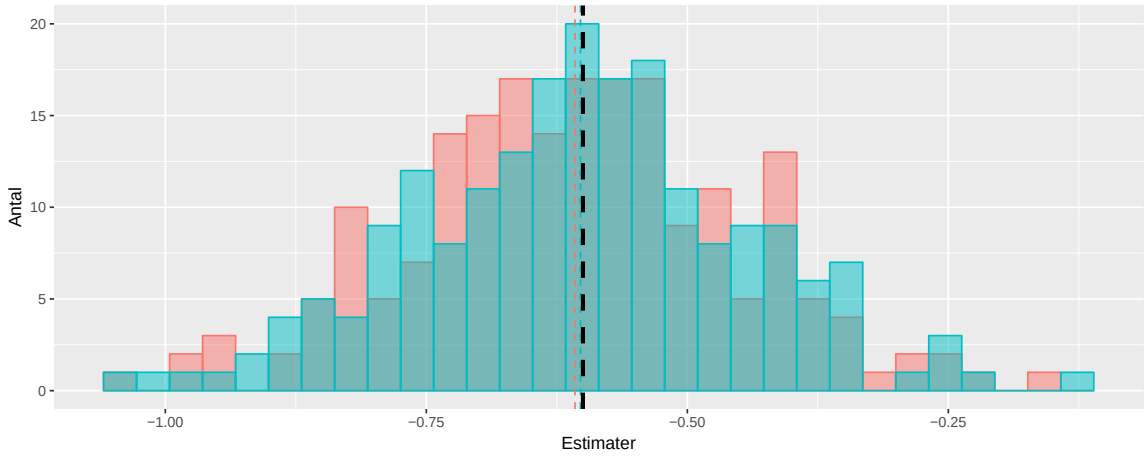
Figur 2.4.2 viser, at en parametrisk likelihood-funktion[2] er forholdsvist god til at estimere α-parameteren i tilfældet med en Weibull-hazard-funktion som baseline hazard-funktion. Variansen for λ lader til at være en smule større, men det lader til, at det er fornuftige estimater, der opnås. De gennemsnitlige værdier for α og λ er begge forholdsvist tæt på deres respektive sande værdi.

[1] Dette svarer til en Weibull-fordeling med $\alpha = 1$ og $\lambda = 0.1$.

[2] 'Parametrisk' anvendes om likelihood-funktionen, når der er tale om en specificeret baseline hazard-funktion.



Figur 2.4.3: Estimaterne er for β_1 fundet ved parametrisk likelihood-funktion (rød) og partiel likelihood-funktion (blå). Stiplede linjer indikerer gennemsnitlige estimater. Den sorte stiplede linje indikerer den sande parameter, `Beta1=2`.



Figur 2.4.4: Estimaterne er for β_2 fundet ved parametrisk likelihood-funktion (rød) og partiel likelihood-funktion (blå). Stiplede linjer indikerer gennemsnitlige estimater. Den sorte stiplede linje indikerer den sande parameter, `Beta2=-0.6`.

Det fremgår af Figur 2.4.3 og Figur 2.4.4, at estimaterne fra den parametriske og den partielle likelihood-funktion er nogenlunde tilsvarende. Estimaterne fra den partielle likelihood-funktion giver gennemsnitligt et estimat, der er tættere på den sande paramter.

MSE	β_1	β_2
Parametrisk	0.03	0.02
Partiel	0.04	0.03

I Tabel 2.4.1 ses 'Mean Square Error' (MSE) for hver af estimaterne fra de to likelihood-funktioner. Det bemærkes, at vurderingen på baggrund af MSE er, at estimaterne fra den parametriske likelihood-funktion passer (lidt) bedre end dem fra den partielle likelihood-funktion. Fordelingen af estimaterne samt den lille forskel i MSE'en tyder dog på, at det principielt er 'ligegyldigt', hvilken af likelihood-funktionerne, der vælges til at estimere. Dette giver et større incitament til at anvende den partielle likelihood-funktion, da denne ingen kendskab kræver til den eksplicitte form af baseline hazard-funktionen. □

2.4.2 Log-normal-fordelingen

En anden fordeling, der er værd at betragte, er log-normal-fordelingen. Lad $X \sim \log N(\mu, \sigma^2)$. Da er tæthedsfunktionen, f_X, givet ved

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

Lad $\Phi(\cdot)$ være den kumulative fordelingsfunktion for standard normalfordelingen. Hermed er overlevelsesfunktionen for X givet ved

$$S_X(x) = 1 - \Phi\left(\frac{\log(x)-\mu}{\sigma}\right)$$

Af (2.4) fås hazard-funktionen

$$h_X(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right)}{1-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)} = \frac{\frac{1}{\sigma x}\phi(\frac{\log(x)-\mu}{\sigma})}{1-\Phi(\frac{\log(x)-\mu}{\sigma})},$$

hvor $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ er tæthedsfunktionen for standard normalfordelingen. Ved brug af L'Hôpitals regel kan grænseværdien for hazard-funktionen, når x går med uendelig, udregnes.

$$\begin{aligned}\lim_{x\to\infty} h_X(x) &= \lim_{x\to\infty} \frac{\frac{1}{\sigma x}\phi(\frac{\log(x)-\mu}{\sigma})}{1-\Phi(\frac{\log(x)-\mu}{\sigma})} = \lim_{x\to\infty} \frac{-\frac{1}{\sigma x^2}\phi(\frac{\log(x)-\mu}{\sigma}) + \frac{1}{\sigma x}\phi(\frac{\log(x)-\mu}{\sigma})\left(-\frac{\log(x)-\mu}{\sigma}\right)\frac{1}{x}}{-\phi(\frac{\log(x)-\mu}{\sigma})} \\ &= \lim_{x\to\infty} \frac{1}{\sigma x^2} + \frac{1}{\sigma x}\cdot\frac{\log(x)-\mu}{\sigma}\cdot\frac{1}{x} = 0.\end{aligned}$$

Hazard-funktionen er gående mod 0, hvorfor denne er uhensigtsmæssig at betragte i forhold til menneskers overlevelse. Dette ville betyde, at stigningen i sandsynligheden for at dø bliver mindre des ældre, en person bliver. Log-normal-fordelingen vil senere blive betragtet i en anden sammenhæng.

2.4.3 Gamma-fordelingen

Lad X følge en gamma-fordeling. Dette noteres $X \sim \Gamma(k, \lambda)$ for $k, \lambda > 0$. Tæthedsfunktionen, f_X, er givet ved

$$f_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}.$$

Overlevelsesfunktionen bliver da

$$S_X(x) = 1 - \frac{\int_0^x \lambda^k u^{k-1} e^{-\lambda u}\, du}{\Gamma(k)}.$$

Dermed følger hazard-funktionen

$$h_X(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k) - \int_0^x \lambda^k u^{k-1} e^{-\lambda u}\, du}.$$

Det bemærkes, at der ikke er en lukket form for integralerne i funktionerne, hvorfor likelihood-funktionen for gamma-fordelte event-tider ikke har en simpel form. Lige som log-normal-fordelingen betragtes denne senere.

3. Frailty-modeller

I Kapitel 2 blev proportional hazard-modellen introduceret. Denne model er baseret på en antagelse om homogenitet op til nogle kendte kovariater i populationen. I nærværende kapitel introduceres udvidelser af proportional hazard-modellen, som indeholder en eller flere uobserverede variable, der redegør for en inhomogenitet i en population eller grupper i populationen. Dette kapitel er primært baseret på Wienke [2011].

3.1 Den en-dimensionelle frailty-model

Den første udvidelse af proportional hazard-modellen, der introduceres, er *den univariate frailty-model*.

Definition 3.1.1 (Den en-dimensionelle frailty-model)
Lad $h^*(t \mid \mathbf{z})$ være en hazard-funktion, der tillader kovariater. Da er den en-dimensionelle frailty-model givet ved

$$h(t \mid \mathbf{z}, W) = W h^*(t \mid \mathbf{z}), \tag{3.1}$$

hvor W er en ikke-negativ stokastisk variabel.

Variablen W kaldes for *frailty-variablen* eller blot *frailty*. Bemærk, at W kan skaleres således, at $\mathbb{E}[W] = 1$, da en eventuel skalar kan absorberes i baseline hazard-funktionen.[1] Af (3.1) samt (2.5) er

$$S(t \mid \mathbf{z}, W) = \exp\left(-W H^*(t \mid \mathbf{z})\right), \tag{3.2}$$

hvor $H^*(t \mid \mathbf{z}) = \int_0^\infty h^*(u \mid \mathbf{z})\, du$. Populationens overlevelsesfunktion $S(t \mid \mathbf{z})$ kan opnås ved at integrere frailty-variablen ud i (3.2). Dermed er

$$S(t \mid \mathbf{z}) = \int_0^\infty S(t \mid \mathbf{z}, w) f_W(w)\, dw = \mathbb{E}_W[S(t \mid \mathbf{z}, W)] = \mathcal{L}_W[H^*(t \mid \mathbf{z})], \tag{3.3}$$

hvor $f_W(\cdot)$ er tæthedsfunktionen for W og $\mathcal{L}[\cdot]$ er Laplacetransformationen givet ved

$$\mathcal{L}_W[u] = \int_0^\infty e^{-uw} f_W(w)\, dw.$$

Laplacetransformen nævnes af praktiske årsager, denne er kendt for visse sandsynlighedsfordelinger. Laplacetransformen er derfor bekvem ved udregning af middelværdier, da der under passende regularitetsbetingelser gælder, at

$$\mathcal{L}'_W[u] = \frac{d}{du}\mathcal{L}_W[u] = \int_0^\infty \frac{d}{du} e^{-uw} f_W(w)\, dw. = -\int_0^\infty w e^{-uw} f_W(w)\, dw.$$

Sættes $u = 0$ opnås følgende relation

$$\mathcal{L}'_W[0] = -\mathbb{E}_W[W]. \tag{3.4}$$

[1] Dette er under antagelsen, at middelværdien af W er endelig.

Eksempel 3.1.2
Betragt den stokastiske variabel $W \sim \Gamma(k, \lambda)$. Den tilhørende tæthedsfunktion er givet i Afsnit 2.4.3. Laplace-transformationen bliver da

$$\mathcal{L}_W[u] = \int_0^\infty \frac{1}{\Gamma(k)} \lambda^k w^{k-1} e^{-\lambda w} e^{-uw}\, dw$$
$$\frac{\lambda^k}{(\lambda+u)^k} \int_0^\infty (\lambda+u)^k \frac{1}{\Gamma(k)} w^{k-1} e^{-(\lambda+u)w}\, dw.$$

Funktionen, der integreres over, svarer til en gamma-fordeling med parametrene k og $\lambda+u$, hvorfor integralet bliver 1. Dermed er

$$\mathcal{L}_W[u] = \frac{\lambda^k}{(\lambda+u)^k} = \left(\frac{\lambda+u}{\lambda}\right)^{-k} = \left(1+\frac{u}{\lambda}\right)^{-k}. \tag{3.5}$$

Dette betyder, at der for populationens overlevelsesfunktion for den en-dimensionelle frailty-model med frailty $W \sim \Gamma(1,1)$ gælder

$$S(t \mid \mathbf{z}) = \mathcal{L}_W[H^*(t \mid \mathbf{z})] = (1 + H^*(t \mid \mathbf{z}))^{-1}. \tag{3.6}$$

□

I følgende sætning kobles hazard-funktionen for populationen, $h(t \mid \mathbf{z})$, sammen med det gennemsnitlige individ i populationen til tid $T > t$.

Sætning 3.1.3
Betragt den univariate frailty-model som i (3.1). Da er

$$h(t \mid \mathbf{z}) = h^*(t \mid \mathbf{z}) \mathbb{E}_W[W \mid T > t].$$

Bevis
Af (2.4) og (3.1) fås

$$h(t \mid \mathbf{z}, W = w) = \frac{f(t \mid \mathbf{z}, W = w)}{S(t \mid \mathbf{z}, W = w)} = w h^*(t \mid \mathbf{z}).$$

Dette kan omskrives til

$$f(t \mid \mathbf{z}, W = w) = w h^*(t \mid \mathbf{z}) S(t \mid \mathbf{z}, W = w).$$

Den simultane tæthed for T og W (betinget med $\mathbf{z}$) kan da udtrykkes ved

$$f(t, w \mid \mathbf{z}) = w h^*(t \mid \mathbf{z}) S(t \mid \mathbf{z}, W = w) f_W(w) \tag{3.7}$$

hvormed frailty-variablen kan integreres ud,

$$f(t \mid \mathbf{z}) = h^*(t \mid \mathbf{z}) \int_0^\infty w S(t \mid \mathbf{z}, w) f_W(w)\, dw. \tag{3.8}$$

Dette samt (2.4) giver, at

$$h(t \mid \mathbf{z}) = \frac{h^*(t \mid \mathbf{z}) \int_0^\infty w S(t \mid \mathbf{z}, W = w) f_W(w)\, dw}{S(t \mid \mathbf{z})}. \tag{3.9}$$

Integralet i (3.9) forsøges nu omskrevet. Betragtes

$$\mathbb{E}_W[W \mid T > t] = \int_0^\infty w f_W(w \mid T > t)\, dw,$$

søges et udtryk for $f_W(w \mid T > t)$. Denne tæthed er defineret ud fra relationen

$$P(W \leq w, T > t \mid \mathbf{z}) = \int_0^w f_W(v \mid T > t) P(T > t \mid \mathbf{z})\, dv = \int_0^w f_W(v \mid T > t) S(t \mid \mathbf{z})\, dv. \tag{3.10}$$

Ydermere gælder for simultane tætheder, at

$$\begin{aligned} P(W \leq w, T > t \mid \mathbf{z}) &= \int_0^w \int_t^\infty f_{(T,W)}(u, v \mid \mathbf{z})\, du\, dv \\ &= \int_0^w \int_t^\infty f_{T|W}(u \mid \mathbf{z}, W = v) f_W(v)\, du\, dv \\ &= \int_0^w S(t \mid \mathbf{z}, W = v) f_W(v)\, dv \\ &= \int_0^w \frac{S(t \mid \mathbf{z}, W = v) f_W(v)}{S(t \mid \mathbf{z})} S(t \mid \mathbf{z})\, dv. \end{aligned} \tag{3.11}$$

Af (3.10) og (3.11) følger

$$f_W(w \mid T > t) = \frac{S(t \mid \mathbf{z}, W = w) f_W(w)}{S(t \mid \mathbf{z})}.$$

Hermed bliver

$$h(t \mid \mathbf{z}) = h^*(t \mid \mathbf{z}) \int_0^\infty w f_W(w \mid T > t)\, dw = h^*(t \mid \mathbf{z}) \mathbb{E}_W[W \mid T > t],$$

hvilket var, hvad der skulle vises. ■

Sætning 3.1.3 giver anledning til at betragte monotoniforholdene for $\mathbb{E}_W[W \mid T > t]$.

Proposition 3.1.4
Under passende regularitetsbetingelser er

$$\frac{d}{dt} \mathbb{E}_W[W \mid T > t] \leq 0.$$

Bevis
Beviset tager udgangspunkt i, at

$$\mathbb{E}_W[W \mid T > t] = \frac{\int_0^\infty w S(t \mid \mathbf{z}, w) f_W(w)\, dw}{S(t \mid \mathbf{z})} = \frac{E_W[W S(t \mid \mathbf{z}, W)]}{S(t \mid \mathbf{z})}.$$

Under passende regularitetsbetingelser er

$$\frac{d}{dt} \mathbb{E}_W[W S(t \mid \mathbf{z}, W)] = \mathbb{E}_W \left[W \frac{d}{dt} S(t \mid \mathbf{z}, W) \right] = -h^*(t \mid \mathbf{z}) \mathbb{E}_W[W^2 S(t \mid \mathbf{z}, W)]. \tag{3.12}$$

Det ses desuden ved brug af (3.3), at

$$\begin{aligned}\frac{d}{dt}S(t \mid \mathbf{z}) &= \frac{d}{dt}\int_0^\infty S(t \mid \mathbf{z}, w) f_W(w)\, dw \\ &= \int_0^\infty -w h^*(t \mid \mathbf{z}) S(t \mid \mathbf{z}, w) f_W(w)\, dw \\ &= -h^*(t \mid \mathbf{z}) \mathbb{E}_W[W S(t \mid \mathbf{z}, W)].\end{aligned}$$

Ved differentiering af $\mathbb{E}_W[W \mid T > t]$ giver kvotientreglen

$$\begin{aligned}\frac{d}{dt}\mathbb{E}_W[W \mid T > t] &= -h^*(t \mid \mathbf{z}) \frac{\mathbb{E}_W[W^2 S(t \mid \mathbf{z}, W)] S(t \mid \mathbf{z}) - \mathbb{E}_W[W S(t \mid \mathbf{z}, W)]^2}{S(t \mid \mathbf{z})^2} \\ &= -h^*(t \mid \mathbf{z}) \left(\mathbb{E}_W \left[W^2 \frac{S(t \mid \mathbf{z}, W)}{S(t \mid \mathbf{z})} \right] - \mathbb{E}_W \left[W \frac{S(t \mid \mathbf{z}, W)}{S(t \mid \mathbf{z})} \right]^2 \right).\end{aligned}$$

Fra beviset af Sætning 3.1.3 er

$$f_W(w \mid T > t) = \frac{S(t \mid \mathbf{z}, W = w) f_W(w)}{S(t \mid \mathbf{z})}.$$

Dermed bliver

$$\begin{aligned}\frac{d}{dt}\mathbb{E}_W[W \mid T > t] &= -h^*(t \mid \mathbf{z}) \left(\mathbb{E}_W \left[W^2 \mid T > t \right] - \mathbb{E}_W \left[W \mid T > t \right]^2 \right) \\ &= -h^*(t \mid \mathbf{z}) \mathrm{Var}[W \mid T > t] \leq 0,\end{aligned}$$

da både $h^*(t \mid \mathbf{z})$ og variansen er ikke-negative. ■

Proposition 3.1.4 betyder, at des større t bliver, jo 'stærkere' bliver populationen med $T > t$. Betragt forholdet mellem hazard-funktionerne for populationen, $h(t \mid \mathbf{z})$ og et individ, $h(t \mid \mathbf{z}, w)$, hvor $\mathbb{E}[W] = w$. Da gælder

$$\frac{h(t \mid \mathbf{z})}{h(t \mid \mathbf{z}, w)} = \frac{\mathbb{E}[W \mid T > t]}{\mathbb{E}[W]} = \frac{\mathbb{E}[W \mid T > t]}{\mathbb{E}[W \mid T > 0]} < 1.$$

Altså er det gennemsnitlige individ 'svagere' end det gennemsnitlige individ i populationen til tid t, hvilket bør medtages i overvejelser, inden der drages konklusioner for individer på baggrund af populationens hazard- eller overlevelsesfunktion. I det følgende eksempel undersøges forskellen mellem hazard-funktionen for populationen og for individerne yderligere. Eksemplet er baseret på Martinussen [2017].

Eksempel 3.1.5
Betragt et studie med syge patienter. Lad z være en observeret kovariat, der er 1, hvis en patient har modtaget behandling, og 0, hvis ingen behandling er modtaget. Lad populationens hazard-funktion, $h(t \mid z)$, være givet som en proportional hazard-model med en tidsvarierende effekt for z. Mere præcist er

$$h(t \mid z) = \exp(\beta z \mathbb{1}[t \leq v]),$$

hvor v er et positivt tal. Hvis $\beta < 0$ kan tiden v tolkes som grænsen mellem, at behandlingen har en effekt og ikke har en effekt. Her er baseline-hazard-funktionen konstant 1. Betragt endvidere frailty-modellen

$$h(t \mid z, W) = W h^*(t \mid z),$$

hvor $W \sim \Gamma(1,1)$. Hazard-funktionen, $h^*(t \mid z)$, kobler hazard-funktionerne for populationen, $h(t \mid z)$, og individet $h(t \mid z, W)$. Målet er at finde et udtryk for individets hazard-funktion ud fra populationens, hvilket gør det muligt at sammenligne disse.

Af Eksempel 3.1.2 er

$$S(t \mid z) = \mathcal{L}_W[H^*(t \mid z)] = (1 + H^*(t \mid z))^{-1}.$$

Heri kan $H^*(t \mid z)$ isoleres. Dette giver

$$H^*(t \mid z) = S(t \mid z)^{-1} - 1 = \exp(H(t \mid z)) - 1.$$

Definitionen af den kumulative hazard-funktion medfører da

$$h^*(t \mid z) = \frac{d}{dt} H^*(t \mid z) = h(t \mid z) \exp(H(t \mid z)).$$

Udtrykket $H(t \mid z)$ afhænger af, om $t \leq v$ eller $t > v$. For $t \leq v$ er

$$h^*(t \mid z) = \exp(\beta z \cdot 1) \exp\left(\int_0^t \exp(\beta z \mathbb{1}\left[u \leq v\right]) \, du \right) = \exp(\beta z) \exp\left(\exp(\beta z) t \right).$$

Endvidere gælder for $t > v$, at

$$\begin{aligned} h^*(t \mid z) &= \exp(\beta z \cdot 0) \exp\left(\int_0^t \exp(\beta z \mathbb{1}\left[u \leq v\right]) \, du \right) \\ &= \exp\left(\int_0^v \exp(\beta z) \, du + \int_v^t \exp(\beta z \cdot 0) \, du \right) \\ &= \exp\left(\exp(\beta z) t + t - v \right). \end{aligned}$$

Dermed er den hazard-funktionen for individerne i en gruppe kendt. Antag, at $\beta < 0$ og betragt forholdet mellem de betingede hazard-funktioner for et individ,

$$\frac{h(t \mid z = 1, W = w)}{h(t \mid z = 0, W = w)} = \frac{h^*(t \mid z = 1)}{h^*(t \mid z = 0)} < 1.$$

Dette gælder for alle (tilladte) værdier af w og t. Betragtes derimod det tilsvarende forhold for populationen, er

$$\frac{h(t \mid z = 1)}{h(t \mid z = 0)} = \exp(\beta \mathbb{1}\left[t \leq v\right]).$$

Det ses, at forholdet er mindre end 1, når $t \leq v$. Derimod er værdien lig med 1, hvis $t > v$. Det betyder altså, at konklusioner taget på baggrund af populationen, muligvis ikke holder i det individuelle tilfælde. I dette eksempel ville det være muligt at konkludere, at en behandling ingen effekt har efter et tidspunkt v, hvis populationen betragtes. Konklusionen for individet er dog, at behandling altid har en effekt. □

Eksempel 3.1.5 viser, at der er forskel på at konklusionerne kan være forskellige for populationen kontra individet. Ingen af konklusionerne er 'forkerte', de kan blot varierere fra marginale til betingede fordelinger.

3.2 Den delte frailty-model

I forrige afsnit blev den en-dimensionelle frailty-model præsenteret. Denne type model bygger på, at individerne i et studie er uafhængige. Dette vil ikke nødvendigvis være tilfældet. Eksempelvis kunne et studie betragte individer med en dødelig sygdom. Endvidere kan nogle individer være fra samme familie. Nogle af disse familier da have større risici for at dø end andre, da individerne i en familie har samme 'genpulje', hvilken måske ikke er inddraget i kovariaterne. I nærværende afsnit introduceres en multi-dimensionel frailty-model, som betragter grupper i data og knytter en frailty til hver af disse.

Lad i dette afsnit n være antallet af grupper, og lad den i'te gruppe have n_i observationer og en tilknyttet frailty W_i for $i = 1, \ldots, n$. Lad ydermere T_{ij}, δ_{ij} og $\mathbf{z}_{ij}$ være henholdsvis tid, status samt kovariaterne for den j'te observation i den i'te gruppe for $i = 1, \ldots, n$ og $j = 1, \ldots, n_i$. Endvidere antages der fremadrettet for event-tiderne X_{ij} for grupperne $i = 1, \ldots n$ og observationerne $j = 1, \ldots, n_i$, at X_{ij}'erne er uafhængige givet $W_1, \ldots W_n$.

Definition 3.2.1 (Den delte frailty-model)
Lad $h_0(\cdot) > 0$ være en vilkårlig hazard-funktion. Da er den delte frailty-model givet ved

$$h(t \mid \mathbf{z}_{ij}, W_i) = W_i h(t \mid \mathbf{z}_{ij}).$$

I nærværende projekt betragtes $h(t \mid \mathbf{z}_{ij})$ som en proportional hazard-model, hvorfor den delte frailty-model kan skrives som

$$h(t \mid \mathbf{z}_{ij}, W_i) = W_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}), \tag{3.13}$$

hvor $h_0(\cdot) = h_0(\cdot, \xi)$ i dette afsnit betragtes som værende specificeret med parametervektoren ξ.

Bemærkning 3.2.2
Grundet den betingede uafhængighed i en gruppe kan den simultane betingede overlevelsesfunktion for den i'te gruppe skrives

$$S(t_{i1}, \ldots, t_{in_i} \mid \mathbf{z}_i, W_i = w_i) = \prod_{j=1}^{n_i} S(t_{ij} \mid \mathbf{z}_{ij}, W_i = w_i) = \exp\left(-W_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right).$$

Her er $\mathbf{z}_i = \{\mathbf{z}_{i1}, \ldots, \mathbf{z}_{in_i}\}$. Som ved den en-dimensionelle frailty-model kan den marginale overlevelsesfunktion for den i'te gruppe opnås ved at midle over frailty-variablen.

$$S(t_{i1}, \ldots, t_{in_i} \mid \mathbf{z}_i) = \mathbb{E}_{W_i}\left[S(t_{i1}, \ldots, t_{in_i} \mid \mathbf{z}_i, W_i)\right] = E_{W_i}\left[\exp\left(-W_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right)\right].$$

Dette svarer endvidere til Laplace-transformationen

$$S(t_{i1}, \ldots, t_{in_i} \mid \mathbf{z}_i) = \mathcal{L}_{W_i}\left[\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right].$$

Betragtes tilfældet, hvor $W_i \sim \Gamma(1/\theta, 1/\theta)$, opnås

$$S(t_{i1}, \ldots, t_{in_i} \mid \mathbf{z}_i) = \left(1 + \theta \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right)^{-\frac{1}{\theta}}.$$

Denne kan ikke faktoriseres. Dette betyder, at gruppens frailty medfører, at observationerne ikke er uafhængige i den marginale simultane fordeling. □

Betragt de observerede realiseringer $(t_{ij}, \delta_{ij}, \mathbf{z}_{ij})$ for $i = 1, \ldots, n$ og $j = 1, \ldots, n_i$ samt de uobserverede frailty-variable $\mathbf{w} = \{w_1, \ldots, w_n\}$. Den frailty-betingede likelihood-funktion for den delte frailty-model på formen (3.13) kan opskrives ved brug af (2.8). For den i'te gruppe er den betingede likelihood-funktion givet ved

$$L_i(\xi, \boldsymbol{\beta} \mid W_i = w_i) = \prod_{j=1}^{n_i} \left(w_i h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}} \exp\left(-w_i H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right) \tag{3.14}$$

$$= w_i^{d_i} \exp\left(-w_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right) \prod_{j=1}^{n_i} \left(h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}}, \tag{3.15}$$

hvor $d_i = \sum_{j=1}^{n_i} \delta_{ij}$, og ξ blot er en pladsholder for eventuelle parametre i eksempelvis baseline hazard-funktionen, $h_0(\cdot)$. I det følgende eksempel betragtes den marginale likelihood-funktion, hvor frailty-variablen integreres ud.

Eksempel 3.2.3
Betragt den betingede likelihood-funktion givet i (3.15). Lad $W_i \sim \Gamma(1/\theta, 1/\theta)$. Den marginale likelihood-funktion kan opstilles ud fra følgende midling af den betingede.

$$L_i(\boldsymbol{\beta}, \xi) = \mathbb{E}_{W_i}[L_i(\theta, \boldsymbol{\beta} \mid W_i = w_i)] = \int_0^\infty L_i(\theta, \boldsymbol{\beta} \mid W_i = w_i) f_{W_i}(w_i)\, dw_i.$$

De faktorer, hvor w_i ikke indgår i (3.15), kan betragtes som værende konstanter. Dermed fås

$$\begin{aligned} L_i(\theta, \boldsymbol{\beta}) &= \int_0^\infty w_i^{d_i} \exp\left(-w_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right) \frac{w_i^{1/\theta - 1} \exp(-\frac{w_i}{\theta})}{\theta^{1/\theta} \Gamma(1/\theta)}\, dw_i \\ &\quad \times \prod_{j=1}^{n_i} \left(h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}} \\ &= \int_0^\infty w_i^{d_i + 1/\theta - 1} \exp\left(-w_i \left[\theta^{-1} + \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right] \right) dw_i \\ &\quad \times \frac{1}{\theta^{1/\theta} \Gamma(1/\theta)} \prod_{j=1}^{n_i} \left(h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}}. \end{aligned}$$

Sættes $k = d_i + 1/\theta$ og $\lambda = \theta^{-1} + \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})$ samt forlænges brøken med $\Gamma(k)\lambda^k$ opnås

$$\begin{aligned} L_i(\theta, \boldsymbol{\beta}) &= \int_0^\infty \frac{\lambda^k w_i^{k-1} \exp(-\lambda w_i)}{\Gamma(k)}\, dw_i \\ &\quad \times \frac{\Gamma(k)}{\lambda^k \theta^{1/\theta} \Gamma(1/\theta)} \prod_{j=1}^{n_i} \left(h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}} \\ &= \frac{\Gamma(d_i + 1/\theta) \prod_{j=1}^{n_i} \left(h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}}}{\left(\theta^{-1} + \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{d_i + 1/\theta} \theta^{1/\theta} \Gamma(1/\theta)}. \end{aligned} \tag{3.16}$$

Dermed er det vist, at en $\Gamma(1/\theta, 1/\theta)$-fordelt frailty kan integreres ud i den betingede likelihood-funktion. Hvis baseline hazard-funktionen er kendt, det vil sige, at der arbejdes med en parametrisk model, kan parametrene i denne samt $\boldsymbol{\beta}$ og θ estimeres. □

Eksempel 3.2.3 kræver kendskab til formen af $h_0(\cdot)$. Dette er imidlertid en stor antagelse at lave, da denne sjældent er kendt. Dette blev også berørt ved Cox proportional hazard-modellen, hvor problemet blev løst med den partielle likelihood-funktion. Tilsvarende metoder vil blive undersøgt i det følgende kapitel.

4. Estimationsmetoder

I nærværende kapitel præsenteres to metoder til at estimere parametre i en delt frailty-model med uspecificeret baseline hazard-funktion. Den ene metode er en tilpasset version af *EM-algoritmen*. Den generelle teori bag EM-algoritmen kan findes i Appendiks D.1. Den anden metode er kendt som *penalised partial likelihood-metoden* og forkortes PPL. Disse metoder tages i brug, når baseline hazard-funktionen er uspecificeret, hvorfor likelihood-funktionen ikke kan anvendes. Dette kapitel er primært baseret på Duchateau og Janssen [2008].

4.1 EM-algoritmen

I dette afsnit opstilles teorien for at anvende EM-algoritmen på en delt frailty-model med gamma-fordelte frailty-variable. Betragt derfor den delte frailty-model

$$h(t \mid \mathbf{z}_{ij}, W_i) = W_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}),$$

hvor $h_0(\cdot)$ er uspecificeret, og $W_i \sim \Gamma(1/\theta, 1/\theta)$ for $i = 1, \ldots, n$. I praksis er W_i latente variable, hvilket vil sige, at disse ikke observeres. Dette kan dog omgås ved at bruge EM-algoritmen. I denne vælges begyndelsesværdier for parametrene, der ønskes estimeret, ud fra hvilke de latente variable kan estimeres ved den forventede værdi betinget med parametrene. Antag derfor, at W_i er observeret med værdien w_i for $i = 1, \ldots, n$. Dette giver

$$h(t \mid \mathbf{z}_{ij}, W_i = w_i) = w_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) = h_0(t) \exp\left(\mathbf{z}_{ij}^\top \boldsymbol{\beta} + \log(w_i)\right) \tag{4.1}$$

Dette er en Cox proportional hazard-model med et offset-led, da $h_0(\cdot)$ er uspecificeret. For fuldstændighedens skyld udledes i Appendiks D.2 den partielle likelihood for Cox proportional hazard-modellen med et offset-led.

Lad $\mathscr{T} = (\mathbf{T}, \boldsymbol{\Delta})$, hvor $\mathbf{T} = (T_{1\,1}, \ldots, T_{n\,n_i})$, og $\boldsymbol{\Delta} = (\Delta_{1\,1}, \ldots, \Delta_{n\,n_i})$. Lad ydermere $\mathbf{w}$ være realiseringen af $\mathbf{W} = (W_1, \ldots, W_n)$. Genkald, at $\mathbf{w}$ antages observeret. Betragt da relationen

$$f_{\mathscr{T},\mathbf{W}}(\mathbf{t}, \mathbf{w} \mid h_0(\cdot), \boldsymbol{\beta}, \theta) = f_{\mathscr{T}\mid\mathbf{W}}(\mathbf{t} \mid h_0(\cdot), \boldsymbol{\beta}) f_{\mathbf{W}}(\mathbf{w} \mid \theta), \tag{4.2}$$

hvor $\mathbf{t}$ er en realisering af $\mathscr{T}$. Kovariaterne $\mathbf{z}_{ij}$ er udeladt for at lette notationen. Den fulde log-likelihood-funktion, $\ell_f(h_0(\cdot), \boldsymbol{\beta}, \theta) = \log(f_{\mathscr{T},\mathbf{W}}(\mathbf{t}, \mathbf{w} \mid h_0(\cdot), \boldsymbol{\beta}, \theta))$, kan da skrives som en sum af to log-likelihood-funktioner, $\ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) = \log(f_{\mathscr{T}\mid\mathbf{W}}(\mathbf{t} \mid h_0(\cdot), \boldsymbol{\beta}))$ og $\ell_{f2}(\theta) = \log(f_{\mathbf{W}}(\mathbf{w} \mid \theta))$. Dette giver

$$\ell_f(h_0(\cdot), \boldsymbol{\beta}, \theta) = \ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) + \ell_{f2}(\theta).$$

Af (3.14) gælder

$$\ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[\delta_{ij} \log\left(w_i h_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right) - w_i H_0(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right]. \tag{4.3}$$

Endvidere er

$$\ell_{f2}(\theta) = \sum_{i=1}^{n} \log f_W(w_i \mid \theta) = \sum_{i=1}^{n} -\log\Gamma\left(\frac{1}{\theta}\right) - \log(\theta^{1/\theta}) + \left(\frac{1}{\theta} - 1\right)\log(w_i) - \frac{1}{\theta} w_i. \tag{4.4}$$

Lad estimaterne for parametrene til det k'te trin i EM-algoritmen, $\xi^{(k)} = (h_0^{(k)}(\cdot), \boldsymbol{\beta}^{(k)}, \theta^{(k)})$, være givet. I de følgende afsnit betragtes henholdsvis E- og M-trinet i EM-algoritmen.

4.1.1 E-trinet

I dette underafsnit betragtes det $k+1$'te trin. Det vil sige, at parametrene er blevet maksimeret for det k'te trin. Definitionen af EM-algoritmen i Appendiks D.1 kræver, at middelværdien af den fulde log-likelihood-funktion[1]betinget med det observerede data og parameterestimaterne til det k'te trin for $k = 0, 1, \ldots$ skal udregnes. Dette betyder helt konkret, at

$$\mathbb{E}\left[\ell_f(h_0(\cdot), \boldsymbol{\beta}, \theta; \mathbf{W}) \mid \mathbf{t}, \xi^{(k)}\right]$$

skal bestemmes. Relationen mellem ℓ_f, ℓ_{f1} og ℓ_{f2} gør, at den betingede middelværdi af ℓ_{f1} og ℓ_{f2} kan betragtes. Det ses, at

$$\mathbb{E}\left[\ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) \mid \mathbf{t}, \xi^{(k)}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\Bigg[\delta_{ij}\left(\mathbb{E}\left[\log(W_i) \mid \mathbf{t}, \xi^{(k)}\right] + \log h_0(t_{ij})\right)\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta}) \\ -\mathbb{E}\left[W_i \mid \mathbf{t}, \xi^{(k)}\right] H_0(t_{ij})\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta})\Bigg]$$

samt, at

$$\mathbb{E}\left[\ell_{f2}(\theta) \mid \mathbf{t}, \xi^{(k)}\right] = \sum_{i=1}^{n} -\log\left(\Gamma\left(\frac{1}{\theta}\right)\theta^{1/\theta}\right) + \left(\frac{1}{\theta} - 1\right)\mathbb{E}\left[\log(W_i) \mid \mathbf{t}, \xi^{(k)}\right] - \frac{1}{\theta}\mathbb{E}\left[W_i \mid \mathbf{t}, \xi^{(k)}\right].$$

Det ses, at de eneste udtryk, der mangler at bestemmes i forhold til den betingede middelværdi, er $\mathbb{E}\left[W_i \mid \mathbf{t}, \xi^{(k)}\right]$ og $\mathbb{E}\left[\log(W_i) \mid \mathbf{t}, \xi^{(k)}\right]$. Disse værdier er baseret på de k'te estimater, hvilke er kendte jævnfør a. Derfor giver det $k+1$'te trin, at

$$\mathbb{E}^{(k+1)}[W_i] = \mathbb{E}\left[W_i \mid \mathbf{t}, \xi^{(k)}\right] = \int_0^\infty w_i f_{W|\mathscr{T}}(w_i \mid \mathbf{t}, \xi^{(k)})\, dw_i.$$

Anvendes Bayes Sætning kan den betingede tæthedsfunktion omskrives til

$$\begin{aligned} f_{W|\mathscr{T}}(w_i \mid \mathbf{t}, \xi^{(k)}) &= \frac{f_{\mathscr{T}|W}(\mathbf{t} \mid W_i = w_i, h_0^{(k)}(\cdot), \boldsymbol{\beta}^{(k)}) f_W(w_i \mid \theta^{(k)})}{f_{\mathscr{T}}(\mathbf{t} \mid \xi^{(k)})} \\ &= \frac{L_i(h_0^{(k)}(\cdot), \boldsymbol{\beta}^{(k)} \mid W_i = w_i) f_W(w_i \mid \theta^{(k)})}{L_i(\xi^{(k)})}. \end{aligned} \tag{4.5}$$

Af (3.15), (3.16) samt $f_W(w_i) = \frac{1}{\Gamma(1/\theta)\theta^{1/\theta}} w_i^{1/\theta-1}\exp(-w_i/\theta)$ kan (4.5) skrives

$$\begin{aligned} f_{W|\mathscr{T}}(w_i \mid \mathbf{t}, \xi^{(k)}) =& \frac{1}{\Gamma\left(d_i + \frac{1}{\theta^{(k)}}\right)} w_i^{d_i + 1/\theta^{(k)} - 1}\left(\frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij})\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta}^{(k)}\right) \\ &\times \exp\left[-w_i\left(\frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij})\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta}^{(k)}\right)\right]. \end{aligned} \tag{4.6}$$

Sættes $\alpha = d_i + \frac{1}{\theta^{(k)}}$ og $\lambda = \frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij})\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta}^{(k)})$ er (4.6) tæthedsfunktionen for $\Gamma(\alpha, \lambda)$. Middelværdien for en $\Gamma(\alpha, \lambda)$-fordelt variabel er $\frac{\alpha}{\lambda}$.[2] Dette giver

$$\mathbb{E}^{(k+1)}[W_i] = \frac{d_i + \frac{1}{\theta^{(k)}}}{\frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij})\exp(\mathbf{z}_{ij}^\top\boldsymbol{\beta}^{(k)})}. \tag{4.7}$$

[1]Dette er i forhold til de latente variable, hvilket vil sige $\mathbf{W}$ i dette afsnit.
[2]Dette ses også af (3.4) og (3.5)

I Appendiks D.3 betragtes $\mathbb{E}[\log(W)]$, hvor W er en gamma-fordelt variabel. Da $f_{W|\mathscr{T}}(w_i \mid \mathbf{t}, \xi^{(k)})$ er tætheden for gamma-fordelingen $\Gamma(\alpha, \lambda)$, hvor $\alpha = d_i + \frac{1}{\theta^{(k)}}$ og $\lambda = \frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}^{(k)}$, gælder resultatet i Appendiks D.3. Dermed fås

$$\mathbb{E}^{(k+1)}[\log W_i] = \psi\left(d_i + \frac{1}{\theta^{(k)}}\right) - \log\left(\frac{1}{\theta^{(k)}} + \sum_{j=1}^{n_i} H_0^{(k)}(t_{ij}) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right). \tag{4.8}$$

Dermed anses disse forventede værdier som de observerede værdier for frailty-variablene i det $k+1$'te trin. E-trinet er derfor afsluttet, hvorfor M-trinet nu kan betragtes.

M-trinet

Da E-trinet er fundet for det $k+1$'te trin, kan M-trinet nu udledes. I dette trin skal den fulde likelihood-funktion maksimeres med hensyn til $\boldsymbol{\beta}$ og θ betinget med de observerede værdier samt de forventede værdier, der blev fundet i E-trinet. Ligeledes skal $h_0(\cdot)$ maksimeres, da denne er krævet i E-trinet. Eftersom baseline hazard-funktionen er uspecificeret, betragtes denne dog som en 'nuisance' parameter. Fra tidligere blev det vist, at

$$\ell_f(h_0(\cdot), \boldsymbol{\beta}, \theta) = \ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) + \ell_{f2}(\theta).$$

Dette betyder, at $\boldsymbol{\beta}$ kan estimeres ved at udelukkende at maksimere $\ell_{f1}(h_0(\cdot), \boldsymbol{\beta})$. I Appendiks D.2 anses $h_0(\cdot)$ også som værende en 'nuisance' parameter. Dette anvendes i et profil likelihood-scenarie, hvilket giver den partielle likelihood-funktion med et offset-led. Denne ses i (D.10), hvilken anvendes til at estimere $\boldsymbol{\beta}$. Heri anses frailty-variablene som værende kendte. Dette overkommes ved at bruge resultaterne fra E-trinet. Fremadrettet benævnes denne log-likelihood-funktion ved ℓ_{f1p}. Endvidere kan $h_0(\cdot)$ og $H_0(\cdot)$ estimeres ved henholdsvis (D.9) og (D.6), da disse estimater skal bruges i E-trinet. Ligeledes kan θ estimeres ved at maksimere $\ell_{f2}(\theta)$. Differentieres (D.10) og (4.4) i forhold til henholdsvis den o'te komponent i $\boldsymbol{\beta}$, $\boldsymbol{\beta}_o$ og θ fås:

$$\frac{\partial}{\partial \beta_o} \ell_{f1p}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \left[-\frac{\sum_{kl \in R(t_{ij})} \mathbf{z}_{kl,o} w_k \exp(\mathbf{z}_{kl}^\top \boldsymbol{\beta})}{\sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^\top \boldsymbol{\beta})} + \mathbf{z}_{ij,o} \right],$$

samt

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell_{f2}(\theta) &= \sum_{i=1}^{n} \frac{\Gamma'\left(\frac{1}{\theta}\right)}{\theta^2 \Gamma\left(\frac{1}{\theta}\right)} + \frac{\log(\theta)}{\theta^2} - \frac{1}{\theta^2} - \frac{\log(w_i)}{\theta^2} + \frac{w_i}{\theta^2} \\ &= \frac{1}{\theta^2}\left(\psi\left(\frac{1}{\theta}\right) + \log(\theta) - 1 - \log(w_i) + w_i\right). \end{aligned}$$

For at løse $\frac{\partial}{\partial \beta_o} \ell_{f1p}(\boldsymbol{\beta}) = 0$ og $\frac{\partial}{\partial \theta} \ell_{f2} = 0$ kræves numeriske metoder. Newton-Raphson-metoden kan eventuelt anvendes til førstnævnte ligning, og bisektionsmetoden kan anvendes til anden ligning på et interval $]0, a]$, da θ skal være positiv.

Startværdier for parametrene, $\boldsymbol{\beta}$ og $h_0(\cdot)$, i EM-algoritmen kan findes ved at bruge den partielle likelihood-funktion. Det vil sige lade være med at tage højde for frailty-variablene. Endvidere foreslår Duchateau og Janssen [2008], at startværdien for θ skal være 1. Det medfører, at start-fordelingen for frailty-variablene er en gamma-fordeling med middelværdi og varians 1. Ud fra disse kan E-trinet da udregnes, hvorefter dette kan bruges i

M-trinet. Det er ikke garanteret, at parametrene konvergerer, da der principielt kan være flere stationære punkter. Derimod vides det, at likelihood-funktionen stiger jævnfør Appendiks D.1. Dermed kan (3.16) udregnes for hvert trin. Endvidere kan forskellen mellem likelihood-funktionen med de nuværende parametre og de forrige parametre udregnes. Hvis denne forskel er mindre end en given tolerance, kan EM-algoritmen afsluttes.

4.2 Penalised partial likelihood

I nævværende afsnit bliver *penalised partial likelihood-metoden* (PPL-metoden) præsenteret. Denne metode betrager frailty-variablene som parametre (noteres W_i), der skal estimeres. PPL-metoden minder om 'coordinate ascend'. Først betragtes variansen, $1/\theta$, som værende kendt, hvormed $\boldsymbol{\beta}$ og W_i kan estimeres. Anden del søger at estimere θ ud fra estimaterne for $\boldsymbol{\beta}$ og W_i. PPL-metoden stopper, når estimatet for θ ændrer sig mindre end en givet tolerance. I det følgende opstilles funktionerne, der bruges i denne metode, i et tilfælde med gamma-fordelte frailty-variable.

Betragt først den delte frailty-model med opskrivningen

$$h(t \mid \mathbf{z}_{ij}, W_i) = W_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta} + \log(W_i) = h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta} + V_i),$$

hvor W_i er en gamma-fordelt frailty med $\Gamma(1/\theta, 1/\theta)$, og $V_i = \log(W_i)$ kaldes en *tilfældig effekt* for $i = 1, \ldots, n$. Likelihood-funktionen, der anvendes under M-trinet i EM-algoritmen, kunne udtrykkes ved følgende sum:

$$\ell_f(h_0(\cdot), \boldsymbol{\beta}, \theta) = \ell_{f1}(h_0(\cdot), \boldsymbol{\beta}) + \ell_{f2}(\theta).$$

Hvis $h_0(\cdot)$ 'udprofileres' i en profil-likelihood, kan denne skrives

$$\ell_{fp}(\boldsymbol{\beta}, \theta) = \ell_{f1p}(\boldsymbol{\beta}) + \ell_{f2}(\theta),$$

hvor ℓ_{f1p} er givet som i (D.10). I PPL-metoden betragtes de tilfældige effekter, $\mathbf{V} = (V_1, \ldots, V_n)$, som værende ukendte parametre, der skal estimeres. Dette leder til den penalised partielle likelihood-funktion, forkortes blot PPL:

$$\ell_{PPL}(\boldsymbol{\beta}, \mathbf{V} \mid \theta) = \ell_{f1p}(\boldsymbol{\beta}, \mathbf{V}) + \ell_{f2}(\mathbf{V} \mid \theta). \tag{4.9}$$

Heri kan $\ell_{f2}(\mathbf{V} \mid \theta) = \sum_{i=1}^n \log f_V(V_i, \theta)$ betragtes som et straf-led, da der (ofte) gælder, at des længere tilfældige effekter er fra deres middelværdi, jo mindre værdi har den resulterende tæthedsfunktion for denne værdi. Eftersom W_i er gamma-fordelt kan tætheden for V_i udregnes til

$$f_V(v) = \frac{(\exp(v))^{1/\theta} \exp\left[-\exp(v)/\theta\right]}{\theta^{1/\theta} \Gamma\left(\frac{1}{\theta}\right)},$$

Da θ er konstant, kan $\ell_{f2}(\mathbf{V} \mid \theta)$ reduceres til en ækvivalent likelihood-funktion.

$$\begin{aligned}
\ell_{f2}(\mathbf{V} \mid \theta) &= \sum_{i=1}^n -\log \Gamma\left(\frac{1}{\theta}\right) - \log(\theta^{1/\theta}) + \frac{1}{\theta} V_i - \frac{1}{\theta} \exp(V_i) \\
&\equiv \sum_{i=1}^n \frac{1}{\theta} V_i - \frac{1}{\theta} \exp(V_i) \\
&= \frac{1}{\theta} \sum_{i=1}^n V_i - \exp(V_i).
\end{aligned}$$

Parametrene $\boldsymbol{\beta}$ og $\mathbf{V}$ kan nu estimeres på normal vis. Som i EM-algoritmen kræver dette en iterativ metode som Newton-Raphson-metoden til at finde løsningerne. Hermed fås estimater for $\boldsymbol{\beta}$ og $\mathbf{V}$. Dernæst foreslår Duchateau og Janssen [2008], at anvende $\boldsymbol{\beta}$ og $\mathbf{V}$

til at opnå estimater for henholdsvis $H_0(\cdot)$ og $h_0(\cdot)$ ved at bruge (D.6) og en omskrivning af (D.9) respektivt. Omskrivningen af (D.9) er blot

$$\hat{a}_{ij} = \frac{1}{\sum_{kl \in R(t_{ij})} \exp(\mathbf{z}_{kl}^\top \boldsymbol{\beta} + V_k)}.$$

Næste trin består i at estimere θ ud fra den marginale likelihood-funktion, (3.16), hvor estimaterne for $\boldsymbol{\beta}$, $h_0(\cdot)$ og $H_0(\cdot)$ anvendes.

En fordel ved PPL-metoden er, at den også kan udvides til at finde estimater i tilfælde af en delt frailty-model med log-normal-fordelte frailty-variable.

4.2.1 Log-normal-fordelte frailty-variable

Gamma-fordelingen er den primært anvendte fordeling for frailty-variablene i nærværende projekt, da denne har forholdsvist pæne egenskaber. Dette betyder dog ikke, at andre fordelinger ikke er tilgængelige. Et alternativ til gamma-fordelingen er log-normal-fordelte frailty-variable. Lad $W_i \sim \text{logN}(\mu_W, \sigma_W^2)$ og betragt igen

$$h(t \mid \mathbf{z}_{ij}, W_i) = W_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta} + V_i),$$

hvor $V_i = \log(W_i)$. Da er $V_i \sim \text{N}(\mu_V, \sigma_V^2)$. I det følgende betragtes det simple tilfælde, hvor $\mu_V = 0$.

Ligesom for tilfældet med gamma-fordelte frailty-variable betragtes PPL fra (4.9). Her er ℓ_{f1p} uændret ved skift af fordelingen for frailty-variablene. Derimod er ℓ_{f2} forskellig fra fordeling til fordeling. Da

$$f_V(v \mid \sigma^2) = \frac{1}{\sqrt{2\pi\sigma_V^2}} \exp\left(-\frac{v^2}{2\sigma_V^2}\right),$$

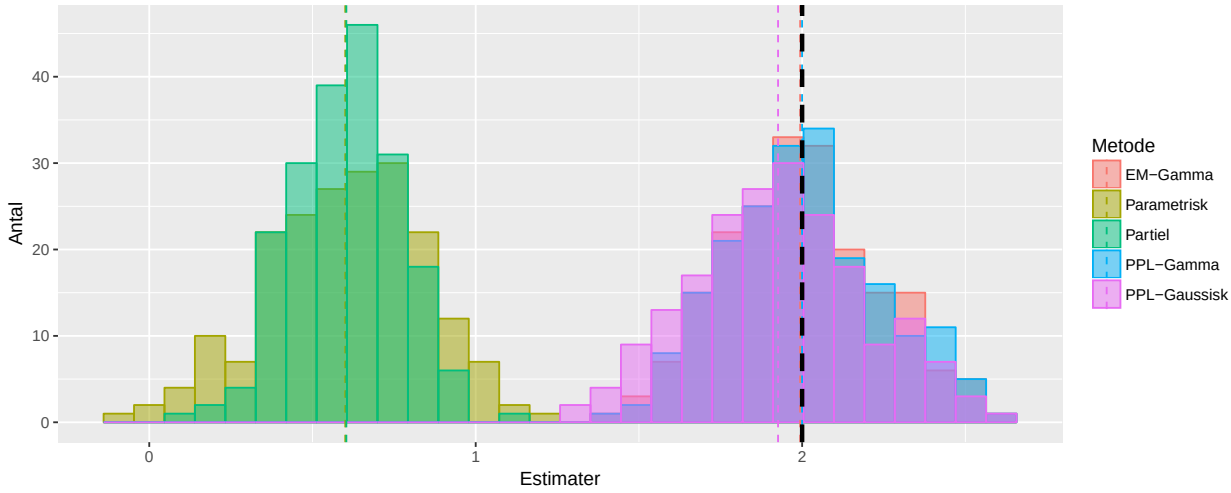
er ℓ_{f2} givet ved

$$\ell_{f2}(\mathbf{V} \mid \sigma_V^2) = \sum_{i=1}^{n} \left[-\frac{1}{2}\left(\log(2\pi) + \log(\sigma^2)\right) - \frac{V_i^2}{2\sigma_V^2}\right] \equiv -\frac{1}{2}\sum_{i=1}^{n}\left[\log(\sigma_V^2) + \frac{V_i^2}{\sigma_V^2}\right].$$

Estimater for $\boldsymbol{\beta}$ og $\mathbf{V}$ kan som ved gamma-fordelte frailty-variable opnås ved Newton-Raphson metoden. Til at estimere σ^2 foreslår Duchateau og Janssen [2008] et REML-estimat.

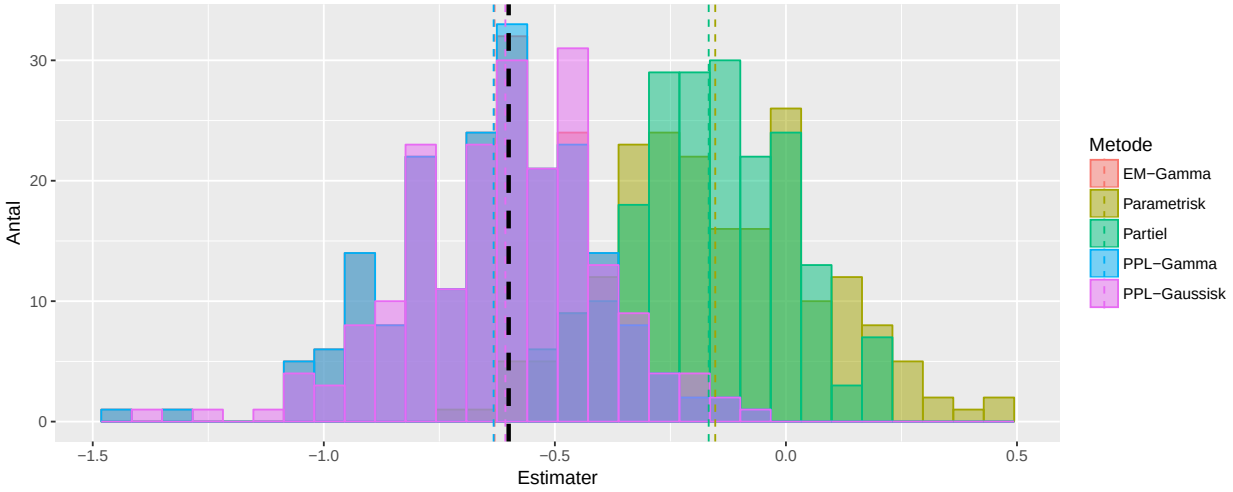
4.3 Sammenligning af metoderne

I dette afsnit simuleres data med `WeibDataSim` fra Appendiks C.1. Data sættes til at have Weibull-fordelingen med $\alpha = 2$ og $\lambda = 3$ som baseline hazard-funktion. Endvidere betragtes to binære kovariater med henholdsvis 2 og -0.6 som sande parametre. Ydermere tilføjes en frailty-variabel til den sande model, hvor frailty-variablen følger gamma-fordelingen $\Gamma(1/\theta, 1/\theta)$, hvor $\theta = 3$. Med andre ord følger den en gamma-fordeling med middelværdi 1 og varians 3. Slutteligt sættes antallet af grupper til 50 samt antallet af individer pr. gruppe til 4, således der simuleres 200 observationer. Censoreringsraten, `rateC`, sættes til 0.1. Sådanne datasæt simuleres 200 gange. På hvert datasæt tilpasses en proportional hazard-model med Weibull baseline hazard-funktion, en Cox proportional hazard-model og en delt frailty-model med gamma-fordelte frailty-variable tilpasset med både EM-algoritmen og PPL-metoden. Derudover tilpasses også en delt frailty-model med log-normal-fordelte frailty-variable med PPL-metoden. Sidstnævnte model er taget med for at se, hvordan en misspecificeret frailty-model klarer sig. `R`-pakkerne `survival` og `frailtyEM` indeholder funktioner, der kan anvende disse metoder. I Appendiks C.3 ses `R`-koden til at generere outputtet i nærværende afsnit.



Figur 4.3.1: Fordeling af estimater for `Beta1`=2. De farvede stiplede linjer indikerer det gennemsnitlige estimat for hver metode. Den sorte stiplede linje indikerer den sande parameter.

I Figur 4.3.1 ses estimaterne for `Beta1` fra hver af de fem forskellige metoder. De stiplede linjer indikerer det gennemsnitlige estimat for hver metode, undtagen den sorte stiplede linje, som markerer den sande parameter. Det ses, at den parametriske model, som har en Weibull-hazard som baseline hazard-funktion, samt Cox proportional hazard-modellen (angivet med henholdsvis 'Parametrisk' og 'Partiel' i Figur 4.3.1) underestimerer `Beta1`. Ydermere bemærkes det, at de to metoder for den delte frailty-model med gamma-fordelte frailty-variable giver stort set identiske estimatfordelinger. EM-algoritmen og PPL-metoden for gamma-fordelte frailty-variable ligger gennemsnitligt oven i den sande parameter `Beta1=2`. Det noteres endvidere, at modellen med log-normal-fordelte frailty-variable gennemsnitligt præsterer estimater i nærheden af den sande parameter. Dette er på trods af, at datasættene er simuleret med gamma-fordelte frailty-variable.



Figur 4.3.2: Fordeling af estimater for `Beta2`=-0.6. De farvede stiplede linjer indikerer det gennemsnitlige estimat for hver metode. Den sorte stiplede linje indikerer den sande parameter.

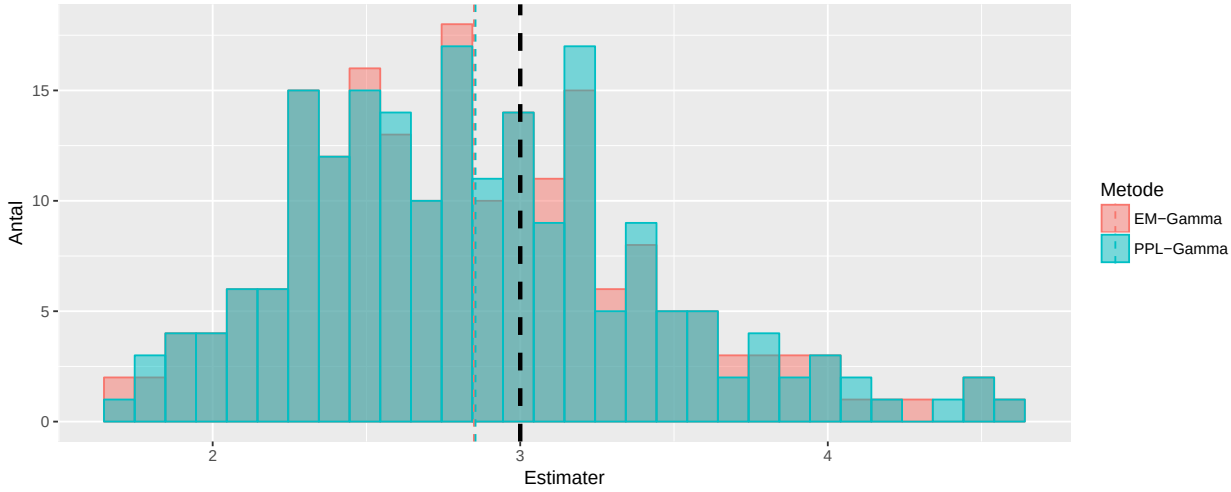
I Figur 4.3.2 ses ligeledes for `Beta2` en underestimation (numerisk set) i forhold til den sande parameter `Beta2=-0.6` ved antagelse af en proportional hazard-model med Weibull-hazard-funktionen som baseline hazard-funktion samt en cox proportional hazard-model. Derimod rammer den delte frailty-model, uafhængig af estimationsmetode, relativt tæt på den sande parameter gennemsnitligt set. For modellen med log-normal-fordelte frailty er det gennemsnitlige estimat for `Beta2` bedre.

Betragtes MSE (Mean Square Error) for de 3 metoder på hver af de to parametre, så giver `R`-koden i Appendiks C.3 følgende:

Tabel 4.3.1: MSE for de 5 forskellige metoder.

MSE	`Beta1`	`Beta2`
Parametrisk	2.02	0.25
Partiel	1.98	0.22
EM-Gamma	0.06	0.05
PPL-Gamma	0.06	0.05
PPL-Gaussisk	0.08	0.04

Af Tabel 4.3.1 fremgår det, at den kvadratiske fejl er mere end 30 gange så stor for estimaterne i de to proportional hazard-modeller for `Beta1` i forhold til den delte frailty-model med både EM-algoritmen og PPL-metoden for gamma-fordelte frailty-variable. Dette forhold er dog reduceret til lidt over 4 for `Beta2`, hvilket formentligt skyldes, at effekten af `Beta2` ikke er lige så stor som den for `Beta1`. Log-normal-fordelte frailty-variable lader til at give et bedre estimat for sande parametre med (relativt) numerisk lave værdier. Umiddelbart tyder disse få simuleringer på, at frailty-modellen ikke er irrelevant. Ydermere kunne det tyde på, at den eksakte fordeling for frailty-variablene er 'ligegyldig' i den forstand, at estimaterne for både gamma- og log-normal-fordelte frailty giver tilsvarende estimater.



Figur 4.3.3: Fordeling af estimater for `theta`=3. De farvede stiplede linjer indikerer det gennemsnitlige estimat for metoderne. Den sorte stiplede linje indikerer den sande parameter.

Figur 4.3.3 viser, hvorledes estimaterne for `theta=3` fordeler sig for henholdsvis EM-algoritmen og PPL-metoden ved en delt frailty-model med gamma-fordelte frailty-variable. Fordelingerne af disse estimater er forholdsvist ens. `R`-funktionen `emfrail` anvender en modificeret version af EM-algoritmen. Denne modificerede version giver estimater (teoretisk), der er lig dem for PPL-metoden. I det følgende afsnit bevises disse egenskaber.

4.4 Yderligere bemærkninger

Kapitlet afrundes med dette afsnit, som indeholder nogle bemærkninger til metoderne og resultaterne herfor.

I Afsnit 4.1 betragtedes EM-algoritmen i forhold til en delt frailty-model med gamma-fordelte frailty-variable. Likelihood-funktionen, der anvendes i den generelle teori for EM-algoritmen i Appendiks D.1, er en parametrisk likelihood-funktion. Dette er dog ikke tilfældet i Afsnit 4.1, hvor den partielle likelihood-funktion med offset anvendes. Gyldigheden af denne 'udskiftning' kan der blive sat spørgsmålstegn ved. I Nielsen et al. [1992] benyttes tælleprocesser til at udlede teorien bag frailty-modeller og estimation af parametre ved EM-algoritmen. Heri argumenteres der for, at den partielle likelihood-funktion kan anvendes i stedet for den parametriske. Argumentet går på, at EM-algoritmen virker ud fra strukturen i relationen mellem den fulde data-likelihood-funktion og den observerede data-likelihood-funktion. Denne relation, argumenterer Nielsen et al. [1992], er intakt, når den partielle likelihood-funktion bruges frem for den parametriske likelihood-funktion.

I Afsnit 4.3 blev forskellige estimationsmetoder og deres estimater sammenlignet. Herunder var EM-algoritmen og PPL-metoden, der begge antog en delt frailty-model med gamma-fordelte frailty-variable. I Figur 4.3.1, Figur 4.3.2 og Figur 4.3.3 var estimaterne for disse to metoder stort set identiske. Teoretisk kan det vises, at estimaterne bør være ens. Dette skyldes, at EM-algoritmen, der anvendes i `R`-pakken `frailtyEM`, er en modificeret version af den beskrevne EM-algoritme fra Afsnit 4.1. I den modificerede version betragtes to dele som ved PPL-metoden. I første del anses θ som værende én fast værdi. Ud fra denne værdi for θ køres EM-algoritmen som i Afsnit 4.1 blot uden at estimere θ. Når den marginale likelihood-funktion anses som værende konvergeret for de estimerede parametre begynder anden del. I denne del estimeres en ny θ ved brug af (3.16). Dette betyder, at anden del for både PPL-metoden og den modificerede EM-algoritme er den samme. Hvis det kan vises, at første del for henholdsvis den modificerede EM-algoritme og PPL-metoden producerer samme estimater, da er det vist, at metoderne giver samme estimater. Bemærk dog, at dette kun gælder for gamma-fordelte frailty-variable. Begge metoder bruger $\ell_{f1p}(\boldsymbol{\beta})$ med offset-ledene $\mathbb{E}[W_m]$ og V_m til at estimere $\boldsymbol{\beta}$. Da

$$\ell_{fp}(\boldsymbol{\beta}, \theta) = \ell_{f1p}(\boldsymbol{\beta}) + \ell_{f2}(\theta),$$

er

$$\frac{\partial}{\partial V_m}\ell_{fp}(\boldsymbol{\beta}, \theta) = \frac{\partial}{\partial V_m}\ell_{f1p}(\boldsymbol{\beta}) + \frac{\partial}{\partial V_m}\ell_{f2}(\theta).$$

For $\frac{\partial}{\partial V_m}\ell_{f1p}(\boldsymbol{\beta})$ gælder der, at

$$\frac{\partial}{\partial V_m}\ell_{f1p}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}\left[\mathbb{1}\left[i=m\right] - \frac{\sum_{ql\in R(t_{ij})}\mathbb{1}\left[q=m\right]\exp(\mathbf{z}_{ql}^\top\boldsymbol{\beta}+V_q)}{\sum_{ql\in R(t_{ij})}\exp(\mathbf{z}_{ql}^\top\boldsymbol{\beta}+V_q)}\right]. \tag{4.10}$$

Bemærk, at

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}\mathbb{1}\left[i=m\right] = \sum_{j=1}^{n_i}\delta_{mj} = d_m.$$

I Appendiks D.2 blev en ombytning af sum-grænser anvendt. Samme argumentation giver følgende brugbare relation

$$\sum_{q=1}^{n}\sum_{l=1}^{n_i}\sum_{ij:t_{ij}\leq t_{ql}} a_{ij}\exp(\mathbf{z}_{ql}^\top\boldsymbol{\beta}+V_q) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} a_{ij}\sum_{ql\in R(t_{ij})}\exp(\mathbf{z}_{ql}^\top\boldsymbol{\beta}+V_q),$$

hvor a_{ij} er givet som i (D.9). Dette medfører

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}\sum_{ql\in R(t_{ij})}\mathbb{1}\left[q=m\right]\exp(\mathbf{z}_{ql}^{\top}\boldsymbol{\beta}+V_q)\frac{1}{\sum_{ql\in R(t_{ij})}\exp(\mathbf{z}_{ql}^{\top}\boldsymbol{\beta}+V_q)}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\delta_{ij}\sum_{ql\in R(t_{ij})}\mathbb{1}\left[q=m\right]\exp(\mathbf{z}_{ql}^{\top}\boldsymbol{\beta}+V_q)a_{ij}\\
&=\sum_{i=1}^{n}\sum_{j=1}^{n_i}a_{ij}\sum_{ql\in R(t_{ij})}\mathbb{1}\left[q=m\right]\exp(\mathbf{z}_{ql}^{\top}\boldsymbol{\beta}+V_q)\\
&=\sum_{q=1}^{n}\sum_{l=1}^{n_i}\mathbb{1}\left[q=m\right]\exp(\mathbf{z}_{ql}^{\top}\boldsymbol{\beta}+V_q)\sum_{ij\colon t_{ij}\leq t_{ql}}a_{ij}\\
&=\sum_{l=1}^{n_m}\exp(\mathbf{z}_{ml}^{\top}\boldsymbol{\beta}+V_m)H_0(t_{ml}),
\end{aligned}$$

hvor $H_0(\cdot)$ er givet som i D.6. Bemærk, at δ_{ij} kan absorberes i a_{ij}, da denne er 0 for $\delta_{ij}=0$ og forskellig fra 0, hvis $\delta_{ij}=1$. Disse udregninger giver samlet, at

$$\frac{\partial}{\partial V_m}\ell_{f1p}(\boldsymbol{\beta})=d_m-\sum_{l=1}^{n_m}\exp(\mathbf{z}_{ml}^{\top}\boldsymbol{\beta}+V_m l)H_0(t_{ml}).$$

Endvidere er

$$\frac{\partial}{\partial V_m}\ell_{f2}=\frac{1-\exp(V_m)}{\theta}.$$

Dermed opnås

$$\frac{\partial}{\partial V_m}\ell_{fp}(\boldsymbol{\beta},\theta)=d_m-\sum_{l=1}^{n_m}\exp(\mathbf{z}_{ml}^{\top}\boldsymbol{\beta}+V_m)H_0(t_{ml})+\frac{1-\exp(V_m)}{\theta}\tag{4.11}$$

Betragt den modificerede EM-algoritme for et givent θ_k og antag, at algoritmen er konvergeret i M-trinet for denne værdi af θ. Da kan estimatet for w_a, $\hat{w}_a$, beregnes ud fra (4.7), hvor θ_k kan indsættes på pladsen for $\theta^{(k)}$. Ligeledes anses også $H_0(\cdot)$ som værende konvergeret med estimatet $\hat{H}_0(\cdot)$. For at sammenligne den modificerede EM-algoritme og PPL-metoden skal w_a transformeres, således at $\hat{w}_a=\log(\hat{V}_a)$. En omskrivning af (4.7) samt transformationen af estimaterne giver

$$\sum_{j=1}^{n_m}H_0(t_{aj})\exp(\mathbf{z}_{ml}^{\top}\boldsymbol{\beta})=\exp(-\hat{V}_m)(d_m+\frac{1}{\theta_k})-\frac{1}{\theta_k}.$$

Dette kan indsættes i (4.11),

$$\frac{\partial}{\partial V_m}\ell_{fp}(\boldsymbol{\beta},\theta_k)=d_m-\left(\exp(-\hat{V}_m)(d_m+\frac{1}{\theta_k})-\frac{1}{\theta_k}\right)\exp(\hat{V}_m)+\frac{1-\exp(\hat{V}_m)}{\theta_k},$$

hvilket giver 0. Derfor giver metoderne samme estiamter for den delte frailty-model med gamma-fordelte frailty-variable, da $\hat{V}_m$ fra EM-algoritmen løser scorefunktionen af PPL.

4.5 Anvendelse af frailty-modeller

I nærværende kapitel undersøges et datasæt ved at tilpasse frailty-modeller. Datasættet er lavet ud fra et forsøg på rotter. I alt er der 100 kuld, hvor der fra hvert kuld er udvalgt 3 rotter. Den ene af rotterne gav man et stof, og de resterende to fungerede som kontrolgruppe. Man ønskede at teste, hvorvidt stoffet havde en indflydelse på udviklingen af tumor. I datasættet findes hver rottes kuld, `litter`, nummereret fra 1 til 100. Derudover findes tiden, `time`, hvor en tumor blev opdaget, eller rotten blev taget ud af studiet. Derfor er også `status` inkluderet, som indikerer, hvorvidt der er tale om en tumor eller en censorering. Endvidere betragtes `rx`, som er 1, hvis rotten modtog stoffet, og 0, hvis denne var i kontrolgruppen. Slutteligt findes også `sex`, som er kønnet på rotten. Alle ulige nummererede kuld indeholder kun hunkøns-rotter. Ligeledes findes der kun hankøns-rotter i kuldene med lige numre.

I Appendiks C.4 kan `R`-koden, der er blevet brugt til at tilpasse modellerne og lave tabellerne i dette afsnit, findes.

Tabel 4.5.1: Summary af tilpassede modeller.

	coef	exp(coef)	se(coef)	z	p
Cox					
rx	0.79	2.21	0.31	2.56	0.01
sexm	-3.07	0.05	0.72	-4.23	0.00
Gamma					
rx	0.79	2.21	0.31	2.53	0.01
sexm	-3.14	0.04	0.74	-4.26	0.00
Log-normal					
rx	0.79	2.21	0.31	2.53	0.01
sexm	-3.10	0.05	0.74	-4.21	0.00

I Tabel 4.5.1 er der blevet tilpasset en Cox proportional hazard-model og en delt frailty-model med henholdsvis gamma- og log-normal-fordelte frailty-variable. I disse modeller er både `rx` og `sex` taget med. Angiveligt lader det til, at rotter af hankøn er mere resistente end kvinderne. Tidligere blev det bemærket, at Cox proportional hazard-modellen ikke var i stand til at give tilfredsstillende estimater, når der var tale om data, der var simuleret med frailty-variable. Det blev endvidere noteret, at gamma- og log-normal-fordelte frailty-variable præsterede tæt på samme estimater. I dette tilfælde ses dog, at alle tre modeller giver stort set ens estimater. Noget kunne tyde på, at kuldene intet har at sige. De estimerede varianser er henholdsvis $\hat{\theta} = 0.47$ og $\sigma^2 = 0.39$. Dette er forholdsvist lave varianser, hvorfor det ikke lader til, at der er stor forskel på kuldene. Da rotterne i hvert kuld har samme køn, er det muligt, at kønnet kan være med til at 'neutralisere' effekten af en eventuel frailty.

En oversigt over `status` viser, at kun 2 rotter af hankøn oplever udviklingen af en tumor. De resterende 148 er altså censoreringer. For hunkønnet er de tilsvarende tal 40 med tumor og 110 censoreringer. Derfor tilpasses modellerne blot på data for rotter af hunkøn.

Tabel 4.5.2: Summary af tilpassede reducerede modeller.

	coef	exp(coef)	se(coef)	z	p
Cox					
rx	0.90	2.47	0.32	2.85	0.00
Gamma					
rx	0.91	2.50	0.32	2.83	0.00
Log-normal					
rx	0.91	2.49	0.32	2.83	0.00

Tabel 4.5.2 viser ligeledes, at estimaterne bliver tilsvarende på trods af, at det kun er et enkelt køn, der betragtes. Det kunne tyde på, at ingen af kuldene er mere (eller mindre) 'resistente' over for at udvikle tumor, hvis de bliver udsat for det givne stof.

Slutteligt kan et interaktionsled tilføjes for at se, hvorvidt effekten af stoffet er forskellig kønnene imellem.

Tabel 4.5.3: Summary af tilpassede reducerede modeller.

	coef	exp(coef)	se(coef)	z	p
Cox					
rx	0.90	2.46	0.32	2.83	0.00
sexm	-2.24	0.11	0.74	-3.01	0.00
rx:sexm	-16.94	0.00	2957.71	-0.01	1.00
Gamma					
rx	0.91	2.48	0.32	2.82	0.00
sexm	-2.30	0.10	0.76	-3.04	0.00
rx:sexm	-	-	-	-	-
Log-normal					
rx	0.91	2.48	0.32	2.82	0.00
sexm	-2.25	0.11	0.75	-2.98	0.00
rx:sexm	-	-	-	-	-

`R`-funktionen `coxph` formår ikke at udregne interaktionsledende for frailty-modellerne grundet singularitet. Tabel 4.5.3 indeholder derfor intet output for dette led i de to modeller. Cox proportional hazard-modellen fik et estimat for interaktionsleddets på trods af fejlmelding om singularitet. Dette er muligvis også grunden til, at det givne estimat har en p-værdi på 1.00. Dette skyldes formentligt, at der ingen rotter af hankøn, der har modtaget stoffet og udviklet en tumor. Al data er altså censoreret i dette tilfælde, hvilket giver problemer med estimation af hazard-funktionerne. Frailty-variablenes varians er på trods af fejlmelding blevet estimeret lavt igen med henholdsvis $\theta = 0.48$ og $\sigma^2 = 0.40$.

5. Afrunding

I Kapitel 4 blev et mindre simulationsstudie opstillet. Her simuleredes 200 datasæt med 200 observationer hver. Observationerne indeholdt et tidspunkt for en begivenhed eller en censorerings indtræffelse, og hvorvidt dette var en hændelse eller censorering. Ydermere blev to binære variable simuleret hver med en tilknyttet effekt på henholdsvis 2 og -0.6. Hver observation blev også tildelt en gruppe, som havde hver deres gamma-fordelte frailty med varians, $\theta = 3$.

En parametrisk og Cox proportional hazard-model blev tilpasset på datasættene. Herudover blev også EM-algoritmen og PPL-metoden anvendt til at tilpasse en delt frailty-model med gamma-fordelte frailty-variable på datasættene. Sidstnævnte metode benyttes også til at tilpasse en delt frailty-model med log-normal-fordelte frailty-variable. Resultaterne viste, at modellerne, der ikke tog højde for frailty, numerisk set underestimerede effekterne for de bineære variable.

Fremadrettet kan det måske vises matematisk, at dette er en generel 'egenskab' ved proportional hazard-modeller, når disse tilpasse på frailty data. Det lader endvidere til, at det ikke kommer sig så nøje, hvilken fordeling, der antages for frailty-variablene, da modellerne med henholdsvis gamma- og log-normal-fordelte frailty-variable gennemsnitligt set lå meget tæt. Dette resultat holdt også, hvis blot Minimum Square Error betragtedes. Endvidere blev det vist, at en modificeret version af EM-algoritmen teoretisk giver samme estimater som PPL-metoden, når gamma-fordelingen er antaget for frailty-variablene. Dette viste sig ikke at holde nøjagtigt, hvilken umiddelbart kan skyldes, at det er to funktioner fra forskellige `R`-pakker, der anvendes.

Mange kilder anvender tælleprocesser til at udlede teorien for overlevelsesanalyse og frailty-modeller, herunder Nielsen et al. [1992]. Tælleprocesserne giver mulighed for at bevise asymptotiske resultater, der blandt andet kan bruges til at lave tests, hvilket dette projekt ikke har fokuseret på.

Den afsluttende model, der blev betragtet i projektet, var den delte frailty-model. Fremadrettet kunne denne model udvides til at tage højde for en gruppes indbyrdes korrelation.

Litteratur

A. P. Dempster, N. M. Laird og D. B. Rubin, *Maximum Likelihood from Incomplete Data via the EM algortihm. Journal of the Royal Statistical Society*, 39(1), 1–38, 1977.

Luc Duchateau og Paul Janssen, *The Frailty Model*, Springer, 2008.

John P. Klein og Melvin L. Moeschberger, *Survival Analysis - Techniques for Censored and Truncated Data*, Springer, 2005.

Torben Martinussen, *Causal inference and survival analysis*, 2017. DSTS todagsmøde 31/10-1/11 2017 i Aarhus.

Geoffrey J. McLachlan og Thriyambakam Krishnan, *The EM Algorithm and Extensions*, John Wiley, 2008, 2. udgave.

Kristoffer Segerstrøm Mørk og Mikkel Findinge, *Statistiske metoder - EM-algoritmen, Faktoranalyse og Bayesianske netværk*, `http://www.aau.dk/digitalAssets/343/343753_p1---statistiske-metoder.pdf`, 2016. [Online; accessed 02-04-2018].

Gert G. Nielsen, Richard D. Gill, Per Kragh Andersen og Thorkild I. A. Sørensen, *A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. Scandinavian Journal of Statistics*, 19(1), 25–43, 1992.

P. Olofsson og M. Andersson, *Probability, statistics, and stochastic processes*, Wiley, 2012, 2. udgave.

Andreas Wienke, *Frailty Models in Survival Analysis*, Chapman & Hall, 2011.

A. Afledede

A.1 Proportional hazard-model

Nedenstående er de dobbelt afledede for en proportional hazard-model, hvor baseline-hazard-funktionen er givet som Weibull-hazard-funktionen.

$$
\begin{aligned}
\frac{\partial^2}{\partial\alpha^2}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\frac{D}{\alpha^2} - \lambda\sum_{i=1}^{n}\log(t_i)^2 t_i^{\alpha}\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta}),\\
\frac{\partial^2}{\partial\lambda^2}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\frac{D}{\lambda^2},\\
\frac{\partial^2}{\partial\alpha\partial\lambda}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\sum_{i=1}^{n}\log(t_i) t_i^{\alpha}\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta}),\\
\frac{\partial^2}{\partial\alpha\partial\beta_k}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\lambda\sum_{i=1}^{n}\log(t_i) t_i^{\alpha} z_{ik}\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta}),\\
\frac{\partial^2}{\partial\lambda\partial\beta_k}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\sum_{i=1}^{n} t_i^{\alpha} z_{ik}\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta}),\\
\frac{\partial^2}{\partial\beta_j\partial\beta_k}\ell(\alpha,\lambda,\boldsymbol{\beta}) &= -\lambda\sum_{i=1}^{n} t_i^{\alpha} z_{ij} z_{ik}\exp(\mathbf{z}_i^{\top}\boldsymbol{\beta}).
\end{aligned}
$$

B. Simulation af event-tider

Dette afsnit betragter simulationer af event-tider, der følger en Cox proportional hazard-model med Weibull-hazard-funktionen som $h_0(t)$. Til dette anvendes den inverse transformationsmetode.

Proposition B.0.1 (Den inverse transformationsmetode)
Lad F være en kontinuert og monotont voksende. Lad ydermere $U \sim \text{unif}[0,1]$ og definer $Y = F^{-1}(U)$. Da har Y den kumulative fordelingsfunktion F.

Bevis
Betragt fordelingsfunktionen for Y, $F_Y(x)$, hvor x er i billedet af Y. Da $F_U(u) = u$ for $0 \leq u \leq 1$ er

$$F_Y(x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F_U(F(x)) = F(x). \qquad \blacksquare$$

Bemærk, at $S(x) = 1 - F(x)$ og anvend, at hvis $U \sim \text{unif}[0,1]$, så er $1 - U \sim \text{unif}[0,1]$. Da giver den inverse transformationsmetode samt overlevelsesfunktionen for en Cox proportional hazard-model, (2.9), at

$$U = \exp(-H_0(X)\exp(\mathbf{z}^\top \boldsymbol{\beta})),$$

hvor $U \sim \text{unif}[0,1]$, samt X er event-tiderne. Isoleres X fås

$$X = H_0^{-1}\left(-\frac{\log U}{\exp(\mathbf{z}^\top \boldsymbol{\beta})}\right).$$

Da Weibull-hazard-funktionen betragtes som baseline hazard-funktion fås af 2.18:

$$H_0(x)^{-1} = \left(\frac{t}{\lambda}\right)^{1/\alpha}.$$

Dermed kan event-tider genereres ved

$$X = \left(-\frac{\log U}{\lambda \exp(\mathbf{z}^\top \boldsymbol{\beta})}\right)^{1/\alpha}.$$

Dette kan udvides til at simulere event-tider for en delt frailty-model, hvor baseline hazard-funktionen er en Weibull-hazard-funktion. I dette tilfælde simuleres først frailty-variablene $w_i > 0$ for $i = 1, \ldots, n$. Dernæst kan frailty betragtes som værende en konstant faktor, hvorfor

$$U = \exp\left(-w_i H_0(X)\exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})\right).$$

Dette giver da

$$X = H_0^{-1}\left(-\frac{\log U}{w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta})}\right).$$

C. R-kode

C.1 R-koden for simulering af event-tider

```
##Funktion til at simulere datasaet med Weibull som baseline i coxph-model eller delt frailty
WeibDataSim <- function(N, alpha, lambda, beta, rateC, frailpar = -1, frailn = -1, probx=-1, seed=-1)
{
  #Hvis fordeling af observerede parametre og seed ikke er givet
  if(probx == -1){
    probx = rep(0.5,length(beta))
  }
  if(seed != -1){
    set.seed(seed)
  }
  #Hvis frailty-variable skal medtages simuler
  frailties = rep(1, N)
  frailg = rep(1, N)

  if(frailpar>0){
    if(frailn == -1){
      frailn = 1
    }
    #Laver frailn antal grupper med frailties som tilhoerende frailty-vaerdier
    frailg = rep(c(1:frailn), times=1, each = N/frailn)
    frailties = rep( rgamma(frailn, shape=1/frailpar,  scale = frailpar),
                     times=1, each = N/frailn)
  }

  #Hver kovariat har N observationer i en matrix
  x <- matrix(ncol=length(beta), nrow=N)

  #Kovariater er bernoulli-fordelte
  for(i in 1:length(beta))
    x[,i] <- sample(x=c(0, 1), size=N, replace=TRUE, prob=c(1-probx[i], probx[i]))

  #Generer Weibull event-tider
  v <- runif(n=N)

  event <- (- log(v) / (lambda * frailties * exp(x %*% beta)))^(1 / alpha)
```

```
36   for(i in 1:length(v)){
37     while(event[i]==0){
38       event[i] <- (- log(runif(1)) / (lambda * frailties[i] *
             exp(x[i,] %*% beta)))^(1 / alpha)
39     }
40   }
41   #Generer censoreringstider fra exp(rateC).
42   C <- rexp(n=N, rate=rateC)
43
44   #status (delta) - skal censorerings- eller event-tid
        vaelges
45   tid <- pmin(event, C)
46   status <- as.numeric(event <= C)
47
48   #Lav data.frame med observeret tid, delta og kovariater
49   data.frame(tid=tid,
50              status=status,
51              x=x,
52              gruppe=frailg)
53 }
```

C.2 R-koden for estimation af (Cox) proportional hazard-model

```
#Load libraries
library(survival)
library(ggplot2)
library(plyr)
#Set seed saa resultater er mulige at genskabe
set.seed(42)
#Hvor mange simulationer samt hvilke vektorer er af interesse
n = 200
beta = c(2,-0.6)
Weib_Par <- matrix(nrow = n, ncol=2)
colnames(Weib_Par) <- c("alpha", "lambda")
Weib_est <- matrix(nrow = n, ncol=2)
colnames(Weib_est) <- c("Beta1", "Beta2")
Cox_est <- matrix(nrow = n, ncol=2)
colnames(Cox_est) <- c("Beta1", "Beta2")
censored <- c()
#Lav n simulationer
for(i in 1:n){
  #Simuler data
  df = WeibDataSim(N=200, alpha=3, lambda=2, beta=beta, rateC
      = 0.1)[,-5]
  #Tilpas Weibull-baseline-hazard paa funktion samt Cox
     proportional hazard model
  Weibull <- survreg(Surv(tid,status)~x.1+x.2, df, dist="
     weibull")
  CoxP <- coxph(Surv(tid,status)~x.1+x.2,df)
  #Estimerede Weibull-parametre
  alpha = 1/Weibull$scale
  b = exp(Weibull$coef[1])
  lambda = 1/(b^alpha)
  Weib_Par[i,] <- c(alpha, lambda)
  #Estimerede Beta
  Weib_est[i,] <- -Weibull$coef[2:3]*alpha
  Cox_est[i,] <- CoxP$coef
  #Hvor mange non-censored observationer
  censored[i] <- sum(df$status)
}
#Hvor mange event-tider observeres
censored <- as.data.frame(censored)
colnames(censored) <- "Observeret"

pdf("Weibevent.pdf",width=10,height=4)
qplot(censored$Observeret, geom="histogram",
      xlab = "Antal ikke-censorerede tider",
      ylab = "Antal simulationer",
      fill=I("cyan"), col=I("black"))
```

```
44 dev.off()
45
46 #Alpha og lambda
47 Parametre <- matrix(c(Weib_Par[,1],Weib_Par[,2], rep(c("Alpha
      ","Lambda"), each=n)), ncol=2)
48
49 Parametre <- as.data.frame(Parametre)
50 colnames(Parametre) <-  c("Estimat", "Parameter")
51 Parametre$Estimat <- as.numeric(as.character(Parametre$
      Estimat))
52
53 #Plotter histogram
54 pdf("Weibpar.pdf",width=10,height=4)
55 mup <- ddply(Parametre, "Parameter", summarise, grp.mean=mean
      (Estimat))
56 ggplot(Parametre, aes(x=Estimat, color=Parameter, fill=
      Parameter)) +
57   geom_histogram(position="identity", alpha = 0.5)+
58   geom_vline(data=mup, aes(xintercept=grp.mean, color=
        Parameter),
59              linetype="dashed")+
60   theme(legend.position="none")+
61   labs(x = "Estimater", y = "Antal")+
62   geom_vline(data=mup, aes(xintercept=2), size=1.2,
63              linetype="dashed")+
64   geom_vline(data=mup, aes(xintercept=3), size=1.2,
65              linetype="dashed")
66 dev.off()
67
68 ###Opstiller matricer til histogrammer til sammenligning
69 Beta_1 = matrix(c(Weib_est[,1],Cox_est[,1], rep(c("
      Parametrisk","Partiel"), each=n)), ncol=2)
70
71 Beta_1 <- as.data.frame(Beta_1)
72 colnames(Beta_1) <-  c("Estimat", "Metode")
73 Beta_1$Estimat <- as.numeric(as.character(Beta_1$Estimat))
74
75
76 Beta_2 = matrix(c(Weib_est[,2],Cox_est[,2], rep(c("
      Parametrisk","Partiel"), each=n)), ncol=2)
77
78 Beta_2 <- as.data.frame(Beta_2)
79 colnames(Beta_2) <-  c("Estimat", "Metode")
80 Beta_2$Estimat <- as.numeric(as.character(Beta_2$Estimat))
81
82 #Plotter histogram
83 pdf("Beta1.pdf",width=10,height=4)
84 mu1 <- ddply(Beta_1, "Metode", summarise, grp.mean=mean(
      Estimat))
```

```
85 ggplot(Beta_1, aes(x=Estimat, color=Metode, fill=Metode)) +
86   geom_histogram(position="identity", alpha = 0.5)+
87   geom_vline(data=mu1, aes(xintercept=grp.mean, color=Metode)
       ,
88              linetype="dashed")+
89   theme(legend.position="none")+
90   labs(x = "Estimater", y = "Antal")+
91   geom_vline(data=mu1, aes(xintercept=beta[1]), size=1.2,
92              linetype="dashed")
93 dev.off()
94 #Plotter histogram
95 pdf("Beta2.pdf",width=10,height=4)
96 mu2 <- ddply(Beta_2, "Metode", summarise, grp.mean=mean(
     Estimat))
97 ggplot(Beta_2, aes(x=Estimat, color=Metode, fill=Metode)) +
98   geom_histogram(position="identity", alpha = 0.5)+
99   geom_vline(data=mu2, aes(xintercept=grp.mean, color=Metode)
       ,
100             linetype="dashed")+
101  theme(legend.position="none")+
102  labs(x = "Estimater", y = "Antal")+
103  geom_vline(data=mu2, aes(xintercept=beta[2]), size=1.2,
104             linetype="dashed")
105 dev.off()
106
107 #Udregner MSE
108 c(sum((Weib_est[,1]-beta[1])^2)/n, sum((Cox_est[,1]-beta[1])
     ^2)/n)
109 c(sum((Weib_est[,2]-beta[2])^2)/n, sum((Cox_est[,2]-beta[2])
     ^2)/n)
```

C.3 R-koden for model- og metode-sammenligning

```
##Simuler data 1000 gange og tilpas model med coxph, EM og PPL
#Load libraries
library(frailtyEM)
library(survival)
library(ggplot2)
library(plyr)
#Sande parametre
beta = c(2,-0.6)
theta = 3 #Variansen
#Weibull-parametre til simulation
alpha = 3
lambda = 2

F_N = 50 #Antal grupper (forskellige frailties)
I_N = 4 #Antal individer pr. gruppe
N = F_N*I_N #Antal observationer i data
rateC = 0.1 #Eksponentialraten for censoreringsfordelingen
k = 200

#Hvilke estimater faas
Est_1 = matrix(ncol = 5, nrow = k)
Est_2 = Est_1
Est_theta = matrix(ncol = 3, nrow=k)

#Simuler og tilpas model. Vaelger seed, saa resultat kan genskabes
set.seed(121293)

for(i in 1:k){
  A = WeibDataSim(N=N, alpha = alpha, lambda = lambda, beta = beta, rateC = rateC, frailpar = theta, frailn=F_N)

  Weib_fit <- survreg(Surv(tid,status)~x.1+x.2, A, dist="weibull")
  alpha_est = 1/Weib_fit$scale

  Cox_fit <- coxph(Surv(tid, status)~ x.1 + x.2, A)
  EM_fit <- emfrail(Surv(tid,status)~x.1+x.2+cluster(gruppe), dist=emfrail_dist(dist = "gamma"), data=A)
  PPL_fit <- coxph(Surv(tid, status)~ x.1 + x.2 + frailty(gruppe, dist="gamma"), A)
  Gaus_fit <- coxph(Surv(tid, status)~ x.1 + x.2 + frailty(gruppe, dist="gaussian"), A)

  Est_1[i,] <- c(-alpha_est*Weib_fit$coef[2], Cox_fit$coef[1], EM_fit$coef[1], PPL_fit$coef[1], Gaus_fit$coef[1])
```

```
40   Est_2[i,] <- c(-alpha_est*Weib_fit$coef[3], Cox_fit$coef
        [2], EM_fit$coef[2], PPL_fit$coef[2], Gaus_fit$coef[2])
41   Est_theta[i,] <- c(1/exp(EM_fit$logtheta),
42                      PPL_fit$history$`frailty(gruppe, dist =
                           "gamma")`$theta,
43                      Gaus_fit$history$`frailty(gruppe, dist =
                            "gaussian")`$theta)
44   }
45
46 ###Opstiller matricer til histogrammer til sammenligning
47 Beta_1 = matrix(c(Est_1[1:(5*k)],
48                   rep(c("Parametrisk","Partiel","EM-Gamma","
                        PPL-Gamma", "PPL-Gaussisk"), each=k)),
                        ncol=2)
49
50 Beta_1 <- as.data.frame(Beta_1)
51 colnames(Beta_1) <-  c("Estimat", "Metode")
52 Beta_1$Estimat <- as.numeric(as.character(Beta_1$Estimat))
53
54
55 Beta_2 = matrix(c(Est_2[1:(5*k)],
56                   rep(c("Parametrisk","Partiel","EM-Gamma","
                        PPL-Gamma", "PPL-Gaussisk"), each=k)),
                        ncol=2)
57
58 Beta_2 <- as.data.frame(Beta_2)
59 colnames(Beta_2) <-  c("Estimat", "Metode")
60 Beta_2$Estimat <- as.numeric(as.character(Beta_2$Estimat))
61
62
63 Theta = matrix(c(Est_theta[1:(2*k)],
64                   rep(c("EM-Gamma","PPL-Gamma"), each=k)),
                        ncol=2)
65 Theta <- as.data.frame(Theta)
66 colnames(Theta) <-  c("Estimat", "Metode")
67 Theta$Estimat <- as.numeric(as.character(Theta$Estimat))
68
69 #Plotter histogram
70 pdf("Beta1.pdf",width=10,height=4)
71 mu1 <- ddply(Beta_1, "Metode", summarise, grp.mean=mean(
       Estimat))
72 ggplot(Beta_1, aes(x=Estimat, color=Metode, fill=Metode)) +
73   geom_histogram(position="identity", alpha = 0.5)+
74   geom_vline(data=mu1, aes(xintercept=grp.mean, color=Metode)
        ,
75              linetype="dashed")+
76     theme(legend.position="right")+
77   labs(x = "Estimater", y = "Antal")+
78   geom_vline(data=mu1, aes(xintercept=beta[1]), size=1.2,
```

```
79                                                    linetype="dashed
                                                         ")
80 dev.off()
81 #Plotter histogram
82 pdf("Beta2.pdf",width=10,height=4)
83 mu2 <- ddply(Beta_2, "Metode", summarise, grp.mean=mean(
       Estimat))
84 ggplot(Beta_2, aes(x=Estimat, color=Metode, fill=Metode)) +
85   geom_histogram(position="identity", alpha = 0.5)+
86   geom_vline(data=mu2, aes(xintercept=grp.mean, color=Metode)
         ,
87              linetype="dashed")+
88   theme(legend.position="right")+
89   labs(x = "Estimater", y = "Antal")+
90   geom_vline(data=mu2, aes(xintercept=beta[2]), size=1.2,
91              linetype="dashed")
92 dev.off()
93
94 #Plotter histogram
95 pdf("Theta.pdf",width=10,height=4)
96 mu3 <- ddply(Theta, "Metode", summarise, grp.mean=mean(
       Estimat))
97 ggplot(Theta, aes(x=Estimat, color=Metode, fill=Metode)) +
98   geom_histogram(position="identity", alpha = 0.5)+
99   geom_vline(data=mu3, aes(xintercept=grp.mean, color=Metode)
         ,
100             linetype="dashed")+
101  theme(legend.position="right")+
102  labs(x = "Estimater", y = "Antal")+
103  geom_vline(data=mu3, aes(xintercept=3), size=1.2,
104             linetype="dashed")
105 dev.off()
106
107 #Udregner MSE
108 colSums((Est_1[,]-beta[1])^2)/N #Parametrisk, Partiel, EM_
       Gamma, PPL_Gamma, PPL_Gaussisk
109 colSums((Est_2[,]-beta[2])^2)/N #Parametrisk, Partiel, EM_
       Gamma, PPL_Gamma, PPL_Gaussisk
```

C.4 R-koden for dataanalyse af rottedata

```
#Hent pakker og se data
library(survival)
library(xtable)
data(rats)
View(rats)

#Hvor mange er censorerede eller har oplevet tumor
table(rats$status)

#Tilpas modeller
Cox_fit <- coxph(Surv(time, status)~rx+sex, rats)
Frail_fit_gamma <- coxph(Surv(time, status)~ rx+sex + frailty
   (litter, dist="gamma"), rats)
Frail_fit_gaus<- coxph(Surv(time, status)~ rx+sex + frailty(
   litter, dist="gaussian"), rats)

Frail_fit_gamma$history$`frailty(litter, dist = "gamma")`[1]
Frail_fit_gaus$history$`frailty(litter, dist = "gaussian")
   `[1]

xtable(Cox_fit)
xtable(Frail_fit_gamma)
xtable(Frail_fit_gaus)

#Hvordan fordeler status sig paa hvert af koennene
table(rats[which(rats$sex=='m'),]$status) #0 - 148, 1 - 2
table(rats[which(rats$sex=='f'),]$status) #0 - 110, 1 - 40

#Tilpas reducerede modeller
rats2 <- rats[which(rats$sex=='f'),]
Cox_fit2 <- coxph(Surv(time, status)~rx, rats2)
Frail_fit_gamma2 <- coxph(Surv(time, status)~ rx + frailty(
   litter, dist="gamma"), rats2)
Frail_fit_gaus2 <- coxph(Surv(time, status)~ rx + frailty(
   litter, dist="gaussian"), rats2)


Frail_fit_gamma2$history$`frailty(litter, dist = "gamma")`[1]
Frail_fit_gaus2$history$`frailty(litter, dist = "gaussian")
   `[1]

xtable(Cox_fit2)
xtable(Frail_fit_gamma2)
xtable(Frail_fit_gaus2)

#Proev fuld data med interaktionsled
Cox_fit3 <- coxph(Surv(time, status)~rx*sex, rats)
```

```
42 Frail_fit_gamma3 <- coxph(Surv(time, status)~ rx*sex +
      frailty(litter, dist="gamma"), rats)
43 Frail_fit_gaus3 <- coxph(Surv(time, status)~ rx*sex + frailty
      (litter, dist="gaussian"), rats)
44
45 Frail_fit_gamma3$history$`frailty(litter, dist = "gamma")`[1]
46 Frail_fit_gaus3$history$`frailty(litter, dist = "gaussian")
      `[1]
47
48 xtable(Cox_fit3)
49 xtable(Frail_fit_gamma3)
50 xtable(Frail_fit_gaus3)
```

D. EM-algoritmen

D.1 Den grundlæggende teori

Dette kapitel er baseret på Dempster et al. [1977] og har til formål at give et overordnet blik af EM-algoritmen. Derfor udelades visse tekniske detaljer. Betragt to stokastiske vektorer $\mathbf{Z} \in \mathcal{Z} \subseteq \mathbb{R}^n$ og $\mathbf{Y} \in \mathcal{Y} \subseteq \mathbb{R}^m$. Lad *de observerede data* betegne realiseringen $\mathbf{z}$ af $\mathbf{Z}$. Lad endvidere $\mathbf{y}$ være en realisering af $\mathbf{Y}$, som indirekte er observeret gennem $\mathbf{z}$. *De fulde data* betegnes $\mathbf{y}$. Lad $\mathcal{Y}(\mathbf{z}) \subseteq \mathcal{Y}$ være mængden, der afhænger af de observerede data, $\mathbf{z}$, sådan at $\mathbf{y}$ kan ses som en realisering af $\mathcal{Y}(\mathbf{z}) \subseteq \mathcal{Y}$.

Betragt de parametriserede tæthedsfunktioner $f_{\mathbf{Z}}(\mathbf{z};\boldsymbol{\theta})$ og $f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})$ for henholdsvis $\mathbf{Z}$ og $\mathbf{Y}$, hvor $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ er parameteren. Da $\mathbf{y} \in \mathcal{Y}(\mathbf{z})$, er

$$f_{\mathbf{Z}}(\mathbf{z};\boldsymbol{\theta}) = \int_{\mathcal{Y}(\mathbf{z})} f_{\mathbf{Y}}(\mathbf{y},\boldsymbol{\theta})\, d\mathbf{y}_{\mathbf{z}},$$

hvor $\mathbf{y}_{\mathbf{z}}$ er indgangene i $\mathbf{y}$, der varierer over $\mathcal{Y}(\mathbf{z})$. Den betingede tæthedsfunktion, $f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta})$, hvor $\mathbf{Y}$ betinges af $\mathbf{Z}$, bliver da

$$f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}) = \frac{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{f_{\mathbf{Z}}(\mathbf{z};\boldsymbol{\theta})}.$$

Lad ℓ_{obs} og ℓ være log-likelihood-funktionerne til de tilsvarende tæthedsfunktioner for $\mathbf{Z}$ og $\mathbf{Y}$. Da er

$$\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{z}) = \ell(\boldsymbol{\theta};\mathbf{y}) - \log(f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta})). \tag{D.1}$$

Fremadrettet betragtes parametrene $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$.

Definition D.1.1 ()
Funktionerne $Q, Q_{\text{bet}} \colon \Theta \times \Theta \to \mathbb{R}$ er givet ved henholdsvis

$$Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathbb{E}\left[\ell(\boldsymbol{\theta}_1;\mathbf{Y}) \mid \mathbf{z},\boldsymbol{\theta}_2\right] \tag{D.2}$$

og

$$Q_{\text{bet}}(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \mathbb{E}\left[\log f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z};\boldsymbol{\theta}_1) \mid \mathbf{z},\boldsymbol{\theta}_2\right]. \tag{D.3}$$

Det følger af (D.1), (D.2) og (D.3), at

$$\ell_{\text{obs}}(\boldsymbol{\theta}_1;\mathbf{z}) = Q(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) - Q_{\text{bet}}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2). \tag{D.4}$$

I det følgende defineres *Expectation Maximization-algoritmen*, som forkortet er EM-algoritmen, samt en mere generel algoritme. Disse er iterative algoritmer, der forsøger at maksimere $\ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{z})$ med hensyn til $\boldsymbol{\theta}$.

Definition D.1.2
Lad $\boldsymbol{\theta} \in \Theta$ være en parameter. EM-algoritmen er for iteration $t = 0, 1, \ldots$ defineret ved følgende trin:

E: Lad $\boldsymbol{\theta}^{(t)}$ være givet, og bestem $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$.

M: Sæt $\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$.

Definér endvidere afbildningen $M \colon \Theta \to \Theta$ ved

$$M(\boldsymbol{\theta}^{(t)}) = \boldsymbol{\theta}^{(t+1)}.$$

M-trinet i EM-algoritmen består i at finde det $\boldsymbol{\theta}$, der maksimererer $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ for $t = 0, 1, \ldots$. Det kan imidlertid være udregningstungt at skulle finde maksimum ved hver iteration. Derfor betragtes en *generaliseret EM-algoritme*.

Definition D.1.3
Lad $M \colon \Theta \to \Theta$. Da er en iterativ algoritme specificeret ved M en generaliseret EM-algoritme (GEM-algoritme), hvis

$$Q(M(\boldsymbol{\theta}), \boldsymbol{\theta}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta})$$

for alle $\boldsymbol{\theta} \in \Theta$.

Dermed kræver GEM-algoritmen blot, at $\boldsymbol{\theta}^{(t+1)}$ ikke får $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ til at aftage, hvor EM-algoritmen krævede, at $\boldsymbol{\theta}^{(t+1)}$ maksimerede denne. Det bemærkes, at EM-algoritmen opfylder

$$Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) = Q(M(\boldsymbol{\theta}^{(t)}), \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}),$$

hvorfor EM-algoritmen er en GEM-algoritme. Fremadrettet betragtes nogle resultater for GEM-algoritmerne.

Lemma D.1.4
Lad $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta \times \Theta$. Da er

$$Q_{\text{bet}}(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) \leq Q_{\text{bet}}(\boldsymbol{\theta}_2; \boldsymbol{\theta}_2),$$

hvor ligheden gælder, hvis og kun hvis $f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_1) = f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_2)$ næsten overalt.

Bevis
Betragt

$$\begin{aligned} Q_{\text{bet}}(\boldsymbol{\theta}_2; \boldsymbol{\theta}_2) - Q_{\text{bet}}(\boldsymbol{\theta}_1; \boldsymbol{\theta}_2) &= \mathbb{E}\left[\log\left(f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_2)\right) \,\middle|\, \mathbf{z}; \boldsymbol{\theta}_2\right] \\ &\quad - \mathbb{E}\left[\log\left[f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_1)\right] \,\middle|\, \mathbf{z}; \boldsymbol{\theta}_2\right] \\ &= \mathbb{E}\left[-\log\left(\frac{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_1)}{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z}; \boldsymbol{\theta}_2)}\right) \,\middle|\, \mathbf{z}; \boldsymbol{\theta}_2\right]. \end{aligned}$$

Jensens ulighed kan nu benyttes, da logaritmen er en konkav funktion, hvorfor den negative logaritme må være konveks. Dermed fås

$$\begin{aligned} Q_{\text{bet}}(\boldsymbol{\theta}_2;\boldsymbol{\theta}_2) - Q_{\text{bet}}(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) &= \mathbb{E}\left[-\log\left(\frac{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z};\boldsymbol{\theta}_1)}{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z};\boldsymbol{\theta}_2)} \right) \,\middle|\, \mathbf{z};\boldsymbol{\theta}_2 \right] \\ &\geq -\log\left(\mathbb{E}\left[\frac{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z};\boldsymbol{\theta}_1)}{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{Y} \mid \mathbf{z};\boldsymbol{\theta}_2)} \,\middle|\, \mathbf{z};\boldsymbol{\theta}_2 \right] \right) \\ &= -\log\left(\int_{\mathcal{Y}(\mathbf{z})} \frac{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_1)}{f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_2)} f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta}_2)\, d\mathbf{y}_{\mathbf{z}} \right) \\ &= -\log 1 = 0. \end{aligned}$$

Dermed er lemmaet bevist. Ligheden i Jensens ulighed gælder, hvis den stokastiske variabel, der transformeres, er konstant næsten overalt. Dermed følger lemmaets hvis og kun hvis relationen af ovenstående udregninger. ■

Sætning D.1.5
Følgende ulighed gælder for enhver GEM-algoritme:

$$\ell_{\text{obs}}(M(\boldsymbol{\theta});\mathbf{z}) \geq \ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{z}) \quad \text{for alle } \boldsymbol{\theta} \in \Theta.$$

Ligheden gælder, hvis og kun hvis to betingelser er opfyldt:

1. $Q(M(\boldsymbol{\theta});\boldsymbol{\theta}) = Q(\boldsymbol{\theta};\boldsymbol{\theta})$.
2. $f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z}; M(\boldsymbol{\theta})) = f_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y} \mid \mathbf{z};\boldsymbol{\theta})$ næsten overalt.

Bevis
Af definitionen for GEM-algoritmerne er $Q(M(\boldsymbol{\theta}),\boldsymbol{\theta}) - Q(\boldsymbol{\theta},\boldsymbol{\theta}) \geq 0$. Endvidere giver Lemma D.1.4, at $Q_{\text{bet}}(\boldsymbol{\theta};\boldsymbol{\theta}) - Q_{\text{bet}}(M(\boldsymbol{\theta});\boldsymbol{\theta}) \geq 0$. Dermed er

$$\ell_{\text{obs}}(M(\boldsymbol{\theta});\mathbf{z}) - \ell_{\text{obs}}(\boldsymbol{\theta};\mathbf{z}) = (Q(M(\boldsymbol{\theta});\boldsymbol{\theta}) - Q(\boldsymbol{\theta};\boldsymbol{\theta})) + (Q_{\text{bet}}(\boldsymbol{\theta};\boldsymbol{\theta}) - Q_{\text{bet}}(M(\boldsymbol{\theta});\boldsymbol{\theta})) \geq 0,$$

hvor hvis og kun hvis relationen også følger af definitionen for GEM-algoritmerne samt Lemma D.1.4. ■

Sætning D.1.5 viser, at likelihood-funktionen stiger, når (G)EM-algoritmen anvendes.

D.2 Den partielle likelihood-funktion med offset

I det følgende udledes en partiel likelihood-funktion for den delte frailty-model

$$h(t \mid \mathbf{z}_{ij}, W_i = w_i) = w_i h_0(t) \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}),$$

hvor W_i antages at være observeret, og $h_0(\cdot)$ er uspecificeret. Dette svarer til den partielle likelihood-funktion med et offset led.

Lad $(t_{ij}, \delta_{ij}, \mathbf{z}_{ij}, w_i)$ for $i = 1, \ldots, n$ og $j = 1, \ldots, n_i$ være realiseringer for det j'te individ i den i'te gruppe. Antag, at ingen t_{ij} er ens. En profil likelihood tilgang kan som ved Cox proportional hazard-modellen i Afsnit 2.3 også anvendes i dette tilfælde. Dette gøres ved at finde et udtryk, der maksimerer $h_0(\cdot)$ i forhold til $\boldsymbol{\beta}$. Af (2.8) bliver likelihood-funktionen for den i'te gruppe

$$L_i(\boldsymbol{\beta}, h_0(\cdot)) = \prod_{j=1}^{n_i} \left(h_0(t_{ij})\, dt_{ij} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}} \exp\left(-H_0(t_{ij}) w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right). \tag{D.5}$$

Lad $D = \{ij \mid \delta_{ij} = 1\}$ være indeksmængden af event-tider. Antag, at $h_0(t_{ij})dt_{ij} = a_{ij} > 0$ i et lille interval $[t_{ij}, t_{ij} + dt_{ij}[$ for $ij \in D$, samt $h_0(t)dt = 0$ andetsteds. Det er derfor kun tider i D, der bidrager til den kumulative hazard-funktion. Notationen i det følgende lettes ved at sætte $a_{ij} = 0$ for $ij \notin D$. Den kumulative baseline hazard-funktion er integralet af hazard-funktionen, hvilken kun er forskellig fra nul på meget små intervaller. Den kumulative baseline hazard-funktion approksimeres da ved

$$H_0^*(t) = \sum_{t_{ij} \le t} a_{ij}. \tag{D.6}$$

Denne approksimation anvendes i (D.5), hvilket giver

$$L_i(\boldsymbol{\beta}) = \prod_{j=1}^{n_i} \left(a_{ij} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right)^{\delta_{ij}} \exp\left(- \sum_{kl:\, t_{kl} \le t_{ij}} a_{kl} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \right).$$

Da det er likelihood-funktionen for den i'te gruppe, der er opstillet, kan den fulde likelihood-funktion for alle grupper findes ved at opstille et produkt af alle disse likelihood-funktioner. Det betyder, at der er to ydre produkter for $i = 1$ og $j = 1$ op til henholdsvis n og n_i. Produktet af eksponentialfunktioner svarer til at summe potenserne. Betragt derfor følgende lighed

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{kl: t_{kl} \le t_{ij}} a_{kl} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) &= \sum_{k=1}^{n} \sum_{l=1}^{n_i} \sum_{ij: t_{kl} \le t_{ij}} a_{kl} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n_i} a_{kl} \sum_{ij \in R(t_{kl})} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}), \end{aligned}$$

hvor $R(t) = \{t_{ij} \mid t \le t_{ij}\}$. Dermed kan den fulde likelihood-funktion skrives som

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left(a_{ij} w_i \exp(\mathbf{z}_{ij}^\top \boldsymbol{\beta}) \exp\left(-a_{ij} \sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^\top \boldsymbol{\beta}) \right) \right)^{\delta_{ij}}, \tag{D.7}$$

da $a_{ij} = 0$, hvis $\delta_{ij} = 0$. Tages logaritmen af (D.7) fås log-likelihood-funktionen

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \left[\log(a_{ij}) + \log(w_i) + \mathbf{z}_{ij}^{\top}\boldsymbol{\beta} - a_{ij} \sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^{\top}\boldsymbol{\beta}) \right]. \tag{D.8}$$

Differentieres (D.8) i forhold til a_{ij} fås udtrykket

$$\frac{\delta_{ij}}{a_{ij}} - \delta_{ij} \sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^{\top}\boldsymbol{\beta}).$$

Sættes dette lig med nul, kan a_{ij} isoleres

$$\hat{a}_{ij} = \frac{1}{\sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^{\top}\boldsymbol{\beta})}. \tag{D.9}$$

Dette giver et estimat for den kumulative hazard-funktion. Indsættes $\hat{a}_{ij}$ i (D.8), opnås den partielle log likelihood-funktion for den delte frailty-model med observerede frailty-variable (eller offset led)

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \delta_{ij} \left[-\log \left(\sum_{kl \in R(t_{ij})} w_k \exp(\mathbf{z}_{kl}^{\top}\boldsymbol{\beta}) \right) + \log(w_i) + \mathbf{z}_{ij}^{\top}\boldsymbol{\beta} \right]. \tag{D.10}$$

D.3 Middelværdi for logaritmen af en gamma-fordelt variabel

Lad $W \sim \Gamma(\alpha, \lambda)$, da er

$$\mathbb{E}_W[\log W] = \int_0^\infty \log(w) \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} \exp(-w\lambda)\, dw.$$

Substituér $u = \lambda w$ i integralet. Dette giver

$$\begin{aligned}\mathbb{E}_W[\log W] &= \int_0^\infty \log\left(\frac{u}{\lambda}\right) \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{u}{\lambda}\right)^{\alpha-1} \exp(-u) \frac{1}{\lambda}\, du \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \log\left(\frac{u}{\lambda}\right) u^{\alpha-1} \exp(-u)\, du.\end{aligned}$$

Da $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \exp(-u)\, du$, og $\frac{d}{d\alpha} u^{\alpha-1} = \log(u) u^{\alpha-1}$, kan følgende omskrivninger laves:

$$\begin{aligned}\mathbb{E}_W[\log W] &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \log\left(\frac{u}{\lambda}\right) u^{\alpha-1} \exp(-u)\, du \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \log(u) u^{\alpha-1} \exp(-u)\, du - \frac{\log\lambda}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} \exp(-u)\, du \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \frac{d}{d\alpha} u^{\alpha-1} \exp(-u)\, du - \frac{\log\lambda}{\Gamma(\alpha)} \Gamma(\alpha) \\ &= \frac{1}{\Gamma(\alpha)} \frac{d}{d\alpha} \int_0^\infty u^{\alpha-1} \exp(-u)\, du - \log\lambda \\ &= \frac{1}{\Gamma(\alpha)} \frac{d}{d\alpha} \Gamma(\alpha) - \log\lambda \\ &= \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log\lambda \\ &= \psi(\alpha) - \log\lambda,\end{aligned}$$

hvor $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$. Funktionen $\psi(\cdot)$ kaldes digamma-funktionen.