



AALBORG UNIVERSITET
STUDENTERRAPPORT

AALBORG UNIVERSITET

SPECIALE

Statistisk analyse af Hodgkin lymfekræft

Sofie Haugaard Andersen

Under vejledning af
Christophe Biscio



AALBORG UNIVERSITET
STUDENTERRAPPORT

School of Engineering and Science

Institut for matematiske fag

Skjernvej 4a

9000 Aalborg

<http://www.math.aau.dk/>

Titel:

Statistisk analyse af Hodgkin lymfekræft

Tema:

Overlevelsesanalyse

Projekt:

Speciale

Projektperiode:

Forår 2021

Projektgruppe:

f21matspec_1

Gruppemedlemmer:

Sofie Haugaard Andersen

Vejleder:

Christophe Biscio

Sidetal: 52

Afsluttet den 03/06-2021

Synopsis:

Rammen for dette speciale er overlevelsessanalyse. I specialet undersøges et datasæt omhandlende Hodgkin lymfekræft, som er stillet til rådighed af Hæmatologisk afdeling på Aalborg Universitetshospital. Datasættet indeholder observation for 1271 patienter samt 48 variable. Der arbejdes ud fra problemformuleringen:

Hvordan er det muligt at lave en model, som kan prædikterer sandsynligheden for overlevelse, når en patient rammes af Hodgkin lymfekræft? Og findes der er forbedring af IPS-modellen eller en anden metode til prædiktation af overlevelsessandsynlighederne, som stadig er let at anvende i praksis?

Problemformuleringen besvares ved brug af teori om overlevelsesanalyse, herunder censurering, Nelson-Aalen estimator, Kaplan-Meier estimator, samt Cox proportional hazard modeller.

Teorien kommer til anvendelse i specialets databehandling, hvor det forsøges at lave en model, som giver bedre prædiktioner og er let at anvende i praksis.

Forord

Dette projekt er udarbejdet i foråret 2021 af en statistik-matematik specialestuderende fra Aalborg universitet. Projektet skrives indenfor emnet overlevelsesanalyse. Målgruppen for dette projekt er statistik-matematik studerende.

Til reference af litteratur bruges Vancouver-modellen, som findes i en liste umiddelbart efter konklusionen. I starten af hvert afsnit klargøres, hvilken primære kilde der er brugt til at skrive afsnittet. Er der brugt flere kilder i afsnittet vil disse refereres umiddelbart efter den givne paragraf.

Definitioner, sætninger og eksempler er nummereret efter hvert delafsnit, hvor eksempelvis 2.1.3 betyder kapitel 2, afsnit 1, boks nummer 3. Ligninger er nummereret efter kapitel. Når der refereres til disse, gøres det ved at skrive "ifølge Sætning x" eller "ifølge ligning (x)". Et bevis afsluttes med ■, og et eksempel afsluttes med □. Begge dele i nedre højre hjørne.

Alle figurer er lavet i RStudio og der vil derfor ikke være referencer til disse.

Abstract

In this master's thesis, the aim is to see how it is possible to build models, which can predict the survival probability for a patient suffering from Hodgkin Lymphoma and to build a model, which is better than the one used in practice now, but still easy to use in practice.

At first theory about survival analysis is presented. This is theory about censoring, the Nelson-Aalen estimator, the Kaplan-Meier estimator and Cox Proportion Hazard models. This is used in the analysis of the data frame on Hodgkin Lymphoma, which is provided by Hematology department on Aalborg Universitetshospital.

A Cox Proportional Hazard model is fitted by variable selection by LASSO, and another Cox model is fitted by the variables, which is significant for the survival time and status of the patients in the data frame. The one used in practice now is the International Prognosis Score (IPS), which is based on a point system. One point for each statement that is fulfilled by the patient. Based on the variables from the LASSO Cox model and the Cox model with the significant variables a modification of the IPS model is made with some new statements. This is done to give a better prediction of the survival probability, without making a model, that is more complicated than the IPS model. The advantages of the IPS model is that it is easy to use in practice, because of the point system, and the ability to read the survival probability on a graph or in a table. With the IPS model there is no calculations, other than counting the number of points.

The modified IPS model and the model with the significant variables are tested on a test data frame and there is not a big difference between the two models ability to predict. Therefore it is concluded that the modified IPS model is the best one, mainly because it is easy to use in practice, which the model with the significant variables is not.

Indholdsfortegnelse

Indholdsfortegnelse	v
Kapitel 1 Indledning	1
Kapitel 2 Teori	3
2.1 Overlevelsesanalyse	3
2.2 Censurering	6
2.3 Nelson-Aalen	8
2.4 Kaplan-Meier	12
2.5 Cox proportional hazard model	14
Kapitel 3 Dataanalyse	17
3.1 Introduktion	17
3.2 Manglende data	18
3.3 Forklarende variable	19
3.4 Modeller	23
3.5 Variabel selektion med LASSO	25
3.6 IPS-model	28
3.7 Test af modeller	33
3.8 Et reduceret datasæt	36
Kapitel 4 Konklusion	40
Litteratur	42
Appendiks A Appendiks	43
A.1 Tælleprocesser og martingaler	43
Appendiks B LASSO	47

Indledning 1

Lymfekræft er en kræftsygdom i lymfeknuderne celler. Disse lymfeknuder er en del af immunforsvaret og findes overalt i kroppen. Der findes over 400 lymfeknuder i kroppen, og disse er en del af lymfekarsystemet. Lymfeknuderne findes i større samlinger i halsen, armhulerne, brystkassen og lyskerne. Da kræften opstår i en af disse lymfeceller, så kan sygdommen opstå hvor som helst i kroppen.

Lymfekræft opdeles i to overordnede kategorier, non-Hodgkin og Hodgkin. Non-Hodgkin udgør cirka 80 % af alle tilfælde, det vil sige, at cirka 20% af tilfældene er Hodgkin. Hodgkin lymfekræft rammer ofte mennesker mellem 20-30 år, og derefter falder risikoen for at få Hodgkin lymfekræft, indtil 50-års alderen hvor risikoen igen stiger.[6]

I dette projekt arbejdes der med et datasæt, som indeholder observationer for patienter med Hodgkin lymfekræft. Datasættet indeholder 48 oplysninger om 1271 patienter. Disse oplysninger er blandt andet alder, køn, værdier målt i blodet, samt hvor kræften befinder sig i kroppen. Projektet er lavet i samarbejde med Hæmatologisk afdeling på Aalborg Universitetshospital.

Det ønskes at finde en model, som kan forudsige overlevelsessandsynligheder for nydiagnosticerede patienter. En sådan model findes allerede, den kaldes den internationale prognose score (IPS), og bruges i praksis nu. Modellen er opbygget på et pointsystem. Det ønskes at undersøge hvor meget information der mistes ved brug af IPS-modellen, og derudover hvilke variable der muligvis vil kunne forbedre forudsigelserne signifikant ved tilføjelse til den nuværende IPS-model, uden at modellen bliver for vanskelig at bruge i praksis. Ydermere ønskes det at undersøge, om de variable der allerede er i IPS-modellen er signifikante for forudsigelserne. Dette leder frem til følgende problemformulering:

Hvordan er det muligt at lave en model, som kan prædikterer sandsynligheden for overlevelse, når en patient rammes af Hodgkin lymfekræft? Og findes der er forbedring af IPS-modellen eller en anden metode til prædiktion af overlevelsessandsynlighederne, som stadig er let at anvende i praksis?

Til besvarelse af problemformuleringen anvendes overlevelsesanalyse, herunder Kaplan-Meier estimatorer, Nelson-Aalen estimatorer og Cox proportional hazard modeller. Denne teori introduceres i Kapitel 2. I Kapitel 3 anvendes denne teori på data, hvor der laves flere Cox proportional hazard modeller på forskellige måde, blandt andet ved brug af LASSO. Det undersøges blandt andet, om variablene i IPS-modellen er signifikante. Herefter sammenlignes modellerne med IPS-modellen, for at se om de er bedre til at prædiktere overlevelsessandsynligheden. Dis-

se modeller evalueres herefter på et test-datasæt. Efter litteraturlisten findes et appendiks, som indeholder teori om tælleprocesser og martingale, samt LASSO metoden.

2.1 Overlevelsesanalyse

Dette afsnit er baseret på [4, s.21-36].

Lad T være tid til en specifik hændelse. Denne hændelse kan eksempelvis være død, udvikling af en sygdom eller nedbrud af udstyr. Fordelingen af T kan beskrive ved fire funktioner; *overlevelsesfunktionen*, *hazard raten*, *tæthedsfunktionen*, og *mean residual life*. Hvis en af de fire funktioner er kendt, så kan de tre andre bestemmes entydigt.

Overlevelsesfunktion

Overlevelsesfunktionen beskriver sandsynligheden for at et individ overlever til tid t .

Definition 2.1.1 (Overlevelsesfunktion)

Overlevelsesfunktionen er defineret som:

$$S(t) = P(T > t). \quad (2.1)$$

$S(t)$ er en kontinuert, strengt aftagende funktion, når T er en kontinuert stokastisk variabel. I dette tilfælde er overlevelsesfunktionen komplementet til den kumulative fordelingsfunktion, $S(t) = 1 - F(t)$, hvor $F(t) = P(T \leq t)$. Overlevelsesfunktionen kan også skrives som:

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx,$$

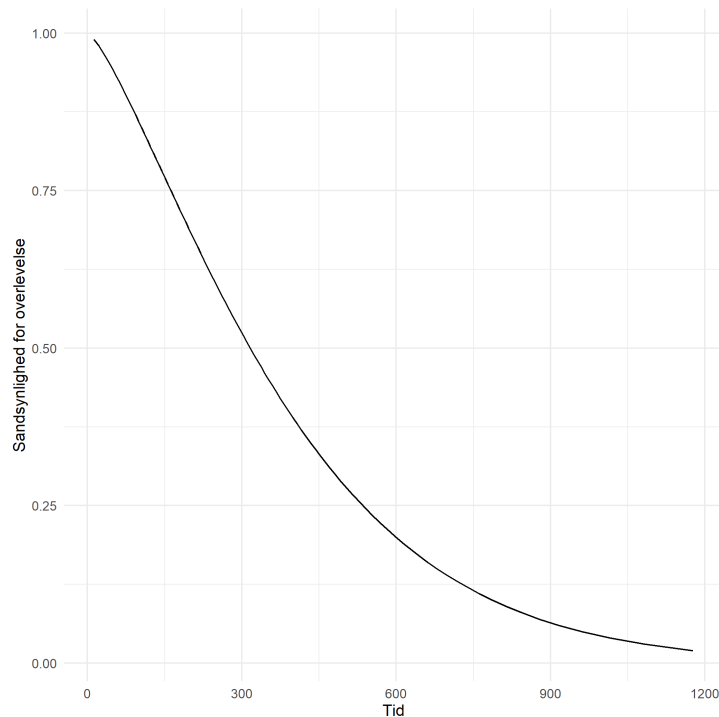
hvor $f(x)$ er tæthedsfunktionen. Den afledede af $S(t)$ er givet ved:

$$\frac{dS(t)}{dt} = \frac{d(1 - F(t))}{dt} = -\frac{dF(t)}{dt} = -f(t), \quad (2.2)$$

det vil dermed også sige at $f(t) = -\frac{dS(t)}{dt}$.

Overlevelsesfunktionens kurve kan være af mange forskellige typer, men fælles er, at de alle er monotone, ikke-voksende funktioner, som er lig 1 til tid 0, og 0 når tiden går mod uendelig. Et eksempel på en overlevelseskurve er en Weibull kurve.

På Figur 2.1 er der plottet en Weibull overlevelseskurve for patienter med lungekræft. Datasættet `lung` er fra R-pakken `KMsurv`. Kurven viser sandsynligheden for overlevelse i forhold til antallet af dage fra diagnosen bliver stillet.



Figur 2.1: En Weibull overlevelseskurve.

Når T er en diskret stokastisk variabel så er overlevelsesfunktionen givet ved:

$$S(t) = \sum_{t_j > t} p(t_j),$$

hvor $p(t_j)$ er probability mass funktion, $p(t_j) = P(T = t_j)$. Når T er diskret, så er overlevelsesfunktionens kurve en ikke-voksende step funktion.

Hazard rate

Hazard raten bruges ofte til at modellere et individs risiko for at dø, som funktion af individets alder, t . Den kan også bruges til at modellere andre tidsafhængige hændelser. Mere specifikt bruges hazard raten til at finde perioder med højest og lavest risiko for at en hændelse sker.

Definition 2.1.2 (Hazard rate)

Hazard raten er givet ved:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t). \quad (2.3)$$

Bemærk, at hvor overlevelseskurven er en funktion som starter i 1 og aftager over tid, så er hazard raten en funktion som kan være alle ikke-negative funktioner.

Hvis T er kontinuert så kan hazard raten skrives som:

$$\begin{aligned}
 h(t) &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \frac{P(T \geq t | t \leq T < t + \Delta t) P(t \leq T < t + \Delta t)}{P(T \geq t)} \\
 &= \lim_{\Delta t \rightarrow 0} \frac{1 \cdot P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} \\
 &= \frac{f(t)}{P(T \geq t)} = \frac{f(t)}{S(t)},
 \end{aligned}$$

hvor Bayes formel, $P(A|B) = P(B|A)P(A)/P(B)$, anvendes i anden lighed, $P(t \leq T < t + \Delta t | T \geq t) = 1$ i tredje lighed og $\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = f(t)$ i fjerde lighed.

Hazard raten kan også skrives som:

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{S(t)} \frac{dF(t)}{dt} = \frac{1}{S(t)} \frac{d(1 - S(t))}{dt} = -\frac{1}{S(t)} \frac{dS(t)}{dt} = -\frac{d}{dt} \ln S(t).$$

Når T er diskret så er hazard raten givet ved:

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_{j-1})}, \quad (2.4)$$

hvor $S(t_0) = 1$. Sættes $p(t_j) = S(t_{j-1}) - S(t_j)$ ind i formlen i ligning (2.4) så er $h(t_j) = 1 - S(t_j)/S(t_{j-1})$.

En anden vigtig funktion er den kumulative hazard funktion, som er defineret ved følgende definition:

Definition 2.1.3 (Kumulative hazard funktion)

Den kumulative hazard funktion er givet ved:

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

For en diskret tid er den kumulative hazard givet ved:

$$H(t) = \sum_{t_j \leq t} h(t_j).$$

Sammenhængen mellem den kumulative hazard og overlevelsesfunktionen er givet ved

$$H(t) = \int_0^t h(u) du = \int_0^t -\frac{d}{du} \ln S(u) du = -\ln S(t).$$

Deraf er $S(t) = \exp(-H(t))$.

Mean Residual Life

Mean residual life er gennemsnitstiden til en hændelse, når hændelsen ikke er sket til tid t .

Definition 2.1.4 (Mean residual life)

Mean residual life er defineret som:

$$mrl(t) = \mathbb{E}[T - t | T > t].$$

Mean life er defineret som $\mu = mrl(0)$, hvilket svarer til det totale areal under overlevelseskurven. Mean residual life er arealet under overlevelseskurven til højre for t og divideret med $S(t)$.

Når T er kontinuert stokastisk variabel så kan mean residual life skrives som:

$$mrl(t) = \frac{\int_t^\infty S(u) du}{S(t)},$$

og $\mu = \mathbb{E}[T] = \int_0^\infty S(u) du$.

2.2 Censurering

Dette afsnit er baseret på [4, s.63-72].

Censurering er nødvendigt, når eksempelvis et studie slutter før alle individer har opnået hændelsen, eller hvis et individ forlader studiet før tid, uden at hændelsen er sket.

Det er flere former for censurering. I dette afsnit ses der på højrecensurering, venstrecensurering og intervalcensurering.

Hvis hændelsen er lig død, så for et specifikt individ i et studie er det antaget, at der er en livstid T og en bestemt censureringstid C . Antag, at alle T er uafhængige og identisk fordelte med tæthedsfunktion $f(t)$ og overlevelsesfunktion $S(t)$. Et individs præcise livstid T kendes kun hvis $T \leq C$. Hvis $T > C$ så overlever individet, og hændelsestidspunktet er censureret af C .

Først kigges på højrecensurering, hvor en hændelse kun er observeret, hvis den sker før en fast tid, t_{obs} . Censureringstiden kan variere fra individ til individ. Hvis et individ går ind i studiet til tid, t_{start} , så er individets censureringstid $C = t_{obs} - t_{start}$. Dette kaldes *type-1 censurering*.

Hvis censureringstiden C ikke er den samme for alle individer så kaldes censureringen *progressiv type-1 censurering*.

Hvis forskellige individer kommer ind i studiet på forskellige tidspunkter, så kaldes censureringen for *generaliseret type-1 censurering*. I dette tilfælde har hvert individ deres egen faste censureringstid.

En anden type af højrecensurering er *type-2 censurering*, hvor studiet starter på samme tid for n individer og studiet fortsætter indtil hændelsen er observeret for r individer. r er et forudbestemt heltal, som er mindre end n . Et studie med denne type censurering ses typisk ved test af levetid for udstyr, eksempelvis lyspære eller batterier.

En tredje type af højrecensurering er *competing risk censurering*. Denne type censurering bruges, når en anden hændelse sker før studiets hændelse. Det vil sige, at T ikke er observeret, og C er tiden til den anden hændelse sker. Denne type censurering er svær i praksis, undtagen når C og T er uafhængige. Et eksempel på denne censurering er, at studiets hændelse er forværring eller død af en bestemt sygdom, men individet dør i en bilulykke.

T er venstrecensureret, hvis T er mindre end C . Det vil sige, at hvis hændelsen allerede er sket for et individ inden individet er observeret i studiet til tid C . Individet har altså oplevet hændelsen før tid C , men det præcise tidspunkt er ukendt. X kendes kun hvis $T \geq C$.

Et eksempel på venstrecensurering er hvis deltagerne i et studie undersøges for tumor i hjernen, og et individ allerede har en tumor ved studiets start. Det betyder, at hændelsestidspunktet er ukendt og individet bliver derfor venstrecensureret.

Intervalcensurering forekommer når det eksakte hændelsestidspunkt ikke kendes, men at det vides at det ligger i et bestemt interval, $(L_i, R_i]$. Det kunne eksempelvis være at drægtige kvæg blev observeret hver 7. dag, og der på observationsdagen blev optalt hvor mange kvæg der havde kælvet. Her kendes altså kun intervallet hvori kalven er født, ikke det eksakte hændelsestidspunkt, derfor vil der være tale om intervalcensurering.

Venstre- og højrecensurering er specielle tilfælde af intervalcensurering, hvor begyndelsen af intervallet er nul for venstrecensurering, og slutningen af intervallet er uendelig for højrecensurering.

Likelihood for censureret data

Beregningen af likelihood funktionen kompliceres af censurering. Per definition er likelihood funktionen den betingede sandsynlighed for data givet modellen, uanset om den er parametrisk eller ikke-parametrisk. Det antages sædvanligvis at dataene er uafhængige givet parametrene. Likelihood funktionen for forskellige typer af censurering kan deles op i følgende komponenter; $f(x)$ er de eksakte livstider, $S(C_r)$ er højrecensurerede observationer, $1 - S(C_l)$ er venstrecensurerede observationer og $S(L) - S(R)$ er intervalcensurerede observationer. Dermed kan likelihood funktionen skrives som:

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} (1 - S(C_l)) \prod_{i \in I} (S(L_i) - S(R_i)), \quad (2.5)$$

hvor D er mængden af hændelsestidspunkter, R er mængden af højrecensurerede observation, L er mængden af venstrecensurerede observationer og I er mængden af intervalcensurerede observationer.

Data som involverer højrecensurering kan repræsenteres som par af stokastiske variable, (T, δ) , hvor δ indikerer om hændelsestidspunktet er observeret, $\delta = 1$, eller ikke-observeret, $\delta = 0$. T er lig X hvis hændelsestidspunktet er observeret, og C_r hvis observationen er højrecensureret, det vil sige at $T = \min(X, C_r)$.

Dermed kan likelihood funktionen for type-1 censurering konstrueres som følger. For $\delta = 0$ så ses det at:

$$\begin{aligned} P(T, \delta = 0) &= P(T = C_r | \delta = 0) P(\delta = 0) = P(\delta = 0) \\ &= P(X > C_r) = S(C_r). \end{aligned}$$

For $\delta = 1$ så ses det at:

$$\begin{aligned} P(T, \delta = 1) &= P(T = X | \delta = 1) P(\delta = 1) \\ &= P(X = T | X \leq C_r) P(X \leq C_r) \\ &= \left[\frac{f(t)}{1 - S(C_r)} \right] (1 - S(C_r)) = f(t). \end{aligned}$$

Disse udtryk kan kombineres til:

$$P(t, \delta) = f(t)^\delta S(t)^{1-\delta}.$$

Likelihood funktionen for et stokastisk par, (T_i, δ_i) , $i = 1, \dots, n$, er da givet ved:

$$L = \prod_{i=1}^n P(t_i, \delta_i) = \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i},$$

Da $f(t_i) = h(t_i)S(t_i)$ og $S(t_i) = \exp(-H(t_i))$ så kan likelihood funktionen skrives som:

$$L = \prod_{i=1}^n h(t_i)^{\delta_i} \exp(-H(t_i)). \quad (2.6)$$

2.3 Nelson-Aalen

Dette afsnit er baseret på [1, s.3-4].

Nelson-Aalen estimatoren er en ikke-parametrisk estimator, som kan bruges til at estimere den kumulative hazard i tilfælde af censureret data.

Antag, at $L \leq n$ er antallet af forskellige observerede hændelsestidspunkter, hvor n er antal af alle observationer. Lad $T_1 < T_2 < \dots < T_L$ være de ordnede tider for hændelserne, og D_k antallet af observerede hændelser til tid T_k . Lad d_l være antallet af hændelser i intervallet $[t_{l-1}, t_l)$, og r_l være antallet af individer i risiko til tid t_{l-1} . For små Δt :

$$H(t + \Delta t) - H(t) \approx h(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t),$$

hvor definition af hazard raten i ligning (2.3) anvendes.

Når $r_l > 0$ så vil estimatet for $H(t_l) - H(t_{l-1})$ være d_l/r_l , og summeres der over de delintervaller, $[t_{l-1}, t_l)$, hvor $r_l > 0$, så vil den kumulative hazard estimeres ved:

$$\hat{H}(t) = \sum_{t_l \leq t} \frac{d_l}{r_l}.$$

Dette leder til følgende definition.

Definition 2.3.1 (Nelson-Aalen estimator)

Nelson-Aalen estimatoren for den kumulative hazard er givet ved:

$$\hat{H}(t) = \sum_{t^* \leq t} \frac{d(t^*)}{r(t^*)},$$

hvor t^* er et tidspunkt hvor mindst en hændelse er sket, $d(t^*)$ er antal hændelser til tid t^* og $r(t^*)$ er antal individer i risiko til tid t^* .

Nelson-Aalen estimatoren har to primære anvendelser i dataanalyse. Den første er at vælge mellem parametriske modeller for tid til hændelse. Plottet af en Nelson-Aalen estimator indtegnes på en speciel type af papir, og hvis en given parametriske model passer godt på data, så vil grafen være approksimativt lineær. Den anden anvendelse af Nelson-Aalen estimatoren er at give estimer for hazard raten $h(t)$. Disse estimer er hældningen af Nelson-Aalen estimatoren.[4, s.94-95]

Martingale

Dette afsnit er baseret på [4, s.84-86].

Nelson-Aalen estimatoren kan undersøges ved brug af martingale teori, som er et godt teoretisk redskab, som har en del gode egenskaber, eksempelvis en central grænseværdisætning, som findes i Sætning A.1.4.

Det ønskes at udlede den ikke-parametriske estimation af den kumulative hazard baseret på højrecensureret data, det vil sige Nelson-Aalen estimatoren, samt at tjekke om den centrale grænseværdis sætning er opfyldt for denne estimator.

Fra Appendix A.1 vides det at $M(t) = N(t) - \Lambda(t)$, og dermed er $dN(t) = Y(t)h(t) + dM(t)$.

Hvis $Y(t)$ er ikke-nul, så:

$$\frac{dN(t)}{Y(t)} = h(t)dt + \frac{dM(t)}{Y(t)}. \quad (2.7)$$

Hvis $dM(t)$ er støj, så er $\frac{dM(t)}{Y(t)}$ også støj. Da værdien af $Y(t)$ er fast givet tiden før t , så fås, ved brug af ligning (A.1):

$$\mathbb{E} \left[\frac{dM(t)}{Y(t)} \middle| \mathbf{F}_{t-} \right] = \frac{\mathbb{E}[dM(t) | \mathbf{F}_{t-}]}{Y(t)} = 0.$$

Det betingede varians af støjen kan findes, ved brug af ligning (A.2):

$$\text{Var} \left[\frac{dM(t)}{Y(t)} \middle| \mathbf{F}_{t-} \right] = \frac{\text{Var}[dM(t) | \mathbf{F}_{t-}]}{Y(t)^2} = \frac{d\langle M \rangle(t)}{Y(t)^2}.$$

Lad nu $J(u)$ være en funktion, som indikerer om $Y(t)$ er positiv, og definer $\frac{0}{0} = 0$, så fås ved integration af ligning (2.7):

$$\int_0^t \frac{J(u)}{Y(u)} dN(u) = \int_0^t J(u)h(u)du + \int_0^t \frac{J(u)}{Y(u)} dM(u). \quad (2.8)$$

Integralet $\int_0^t \frac{J(u)}{Y(u)} dN(u) = \hat{H}(t)$ er Nelson-Aalen estimatoren for $H(t)$, da $Y(u)$ er antal individer i risiko. Integralet $W(t) = \int_0^t \frac{J(u)}{Y(u)} dM(u)$ er et stokastisk integral for den prædikterbare proces $\frac{J(u)}{Y(u)}$ med hensyn til en martingale $M(u)$, og derfor er $W(t)$ også en martingale. Dette integral kan tænkes på som tilfældig støj eller en statistisk usikkerhed i estimatet.

Det stokastiske led $\int_0^t J(u)h(u)du = H^*(t)$ er for højrecensureret data lig $H(t)$, da $Y(t)$ er positiv for $t \leq \max\{T_1, \dots, T_n\}$, altså højrecensureret data. Hvis den statistiske usikkerhed $W(t)$ ignoreres så er $\hat{H}(t)$ den ikke-parametriske estimator for $H^*(t)$.

Da $W(t) = \hat{H}(t) - H^*(t)$ er en martingale, så er $\mathbb{E}[\hat{H}(t)] = \mathbb{E}[H^*(t)]$. Det skal dog bemærkes, at $H^*(t)$ er en stokastisk variabel, og dens forventede værdi er ikke generelt lig $H(t)$. Den prædikterbare proces for $W(t)$ findes, ved brug af ligning (A.3), til at være:

$$\begin{aligned} \langle W \rangle(t) &= \int_0^t \left[\frac{J(u)}{Y(u)} \right]^2 d\langle M \rangle(t) = \int_0^t \left[\frac{J(u)}{Y(u)} \right]^2 Y(u)h(u)du \\ &= \int_0^t \left[\frac{J(u)}{Y(u)} \right] h(u)du. \end{aligned}$$

$Y(t)/n$ og $N(t)/n$ er stikprøve gennemsnit og for et stor stikprøve vil den tilfældige variation i dem være lille. For et stort n , antages at $Y(t)/n$ vil være tæt på den deterministiske funktion $y(t)$. Lad $Z^{(n)}(t) = \sqrt{n}W(t) = \sqrt{n}(\hat{H}(t) - H^*(t))$. Da $H^*(t)$ er tæt på at være lig $H(t)$ for store n , så er denne proces tæt på at være lig $\sqrt{n}(\hat{H}(t) - H(t))$.

Den betingede varians af springene i $Z^{(n)}(t)$ givet filtrationen kan udregnes til følgende:

$$\begin{aligned}\text{Var}[dZ^{(n)}(t)|\mathbf{F}_{t-}] &= \text{Var}[\sqrt{n}dW(t)|\mathbf{F}_{t-}] = n\text{Var}\left[\frac{dM(t)}{Y(t)}|\mathbf{F}_{t-}\right] \\ &= n\frac{d\langle M \rangle(t)}{Y(t)^2} = n\frac{\lambda dt}{Y(t)^2} \\ &= n\frac{Y(t)h(t)dt}{Y(t)^2} = \frac{h(t)dt}{Y(t)/n},\end{aligned}$$

hvilket for store n konvergerer mod $\frac{h(t)dt}{y(t)}$. For store n så vil $Z^{(n)}$ have mange spring, men disse spring vil være små og af orden $1/\sqrt{n}$.

Dette betyder at for store stikprøver er $Z^{(n)}$ næsten kontinuert og har en prædikerbar variation meget tæt på:

$$\langle Z^{(n)} \rangle \approx \int_0^t \frac{h(u)du}{y(u)}. \quad (2.9)$$

Det viser sig, at der kun er en begrænsende proces, $Z^{(\infty)}$ som er en martingale med kontinuert udfaldssti og en deterministisk prædikerbar varians $\langle Z^{(\infty)} \rangle$ som er præcis lig (2.9). Denne begrænsende proces har uafhængigt increments og normalfordelte endelige-dimensionelle fordelinger.

En proces har uafhængige increments, hvis der for enhver mængde af ikke-overlappende intervaller (t_{i-1}, t_i) gælder, at de stokastiske variable $Z^{(\infty)}(t_i) - Z^{(\infty)}(t_{i-1})$ er uafhængige.

Den begrænsende proces har normalfordelte endelig-dimensionelle fordelinger hvis simultanfordelingen af $[Z^{(\infty)}(t_1), \dots, Z^{(\infty)}(t_k)]$ er multivariat normale for enhver k .

For processen $Z^{(\infty)}$ har $[Z^{(\infty)}(t_1), \dots, Z^{(\infty)}(t_k)]$ en k -variater normalfordeling med middelværdi 0 og kovariansmatrix med indgangene:

$$\text{cov}[Z^{(\infty)}(t), Z^{(\infty)}(t_s)] = \int_0^{\min(s,t)} \frac{h(u)du}{y(u)}.$$

Denne konvergens gør, at det er muligt at finde konfidensintervaller for den kumulative hazard til en fast tid, da $\sqrt{n}(\hat{H}(t) - H^*(t))$ vil have en approksimativ normalfordeling med middelværdi 0 og varians:

$$\sigma(Z^{(\infty)}) = \int_0^t \frac{h(u)du}{y(u)}.$$

Et estimat for variansen kan findes fra:

$$n \int_0^t \frac{dN(u)}{Y(u)^2}$$

da $y(t)$ kan estimeres ved $Y(t)/n$ og $h(t)$ ved $dN(t)/Y(t)$.

2.4 Kaplan-Meier

Dette afsnit er baseret på [4, s.92-93,99].

Kaplan-Meier estimatoren er en ikke-parametrisk estimator, som bruges til at estimere ukendte funktioner, eksempelvis overlevelseshfunktionen eller den kumulative hazard. En af fordelene ved Kaplan-Meier er, at den ikke har antagelser omkring type af fordeling, men ulempen er at det kræves, at observationer er uafhængige og identisk fordelte.

Definition 2.4.1 (Kaplan-Meier estimatoren for overlevelseshfunktionen)

Kaplan-Meier estimatoren for $S(t)$ er givet ved:

$$\hat{S}(t) = \hat{P}(X > t) = \prod_{t^* \in D: t^* \leq t} \left(1 - \frac{d(t^*)}{r(t^*)}\right), \quad (2.10)$$

hvor t^* er et tidspunkt hvor mindst en hændelse er sket, $d(t^*)$ er antallet af hændelser til tid t^* , og $r(t^*)$ er antallet af individer i risiko til tid t^* .

Bemærk, hvis ikke der er censurerede observationer så reduceres (2.10) til $1 - \hat{F}_n(t)$, hvor $\hat{F}_n(t)$ er den empiriske fordelingsfunktion, $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}$.

Et plot af Kaplan-Meier estimatoren er en række af faldene horisontale skridt, som går mod den sande overlevelseshfunktion, når antallet af observationer er stort nok.

En fordel ved Kaplan-Meier kurven er, at den kan tage højde for højrecensureret data.

Kaplan-Meier estimatoren for den kumulative hazard er givet ved:

$$\hat{H}(t) = -\ln(\hat{S}(t)).$$

Både Kaplan-Meier og Nelson-Aalen estimatoren er baseret på antagelser om ikke-informative censurering, hvilket betyder at kendskabet til censureringstiden for et individ ikke giver information om individets sandsynlighed for at overleve på et fremtidigt tidspunkt, hvis individet havde fortsat i studiet. Denne antagelse brydes hvis individer med dårlige prognoser systematisk censureres. Dette ville lede til estimering af en forkert funktion, hvilket vil lede til forkerte konklusioner.

Martingale

Kaplan-Meier estimatoren kan også undersøges med martingale teori, som tillader at bestemme den asymptotiske opførelse ved censurering. For at estimere overlevelseshfunktionen, huskes det at for en kontinuert stokastisk variabel så er $S(t) = \exp(-H(t))$, og for en diskret stokastisk variabel så er $S(t) = \prod_{s=0}^t (1 - d\hat{H}(s))$. Her er $S(t)$ produkt integralet af $1 - d\hat{H}(t)$, dermed

er estimatoren for overlevelsesfunktionen givet ved:

$$\hat{S}(t) = \prod_{s=0}^t (1 - d\hat{H}(s)) = \prod_{s=0}^t \left(1 - \frac{dN(s)}{Y(s)}\right),$$

hvor $dN(t) > 0$. Det viser sig, at $\hat{S}(t)/S(t) - 1 = - \int_0^t \frac{\hat{S}(s^-)}{\hat{S}(s)Y(s)} dM(s)$ er et stokastisk integral, med hensyn til martingalen $M(s)$, af en prædikerbar proces, og estimatoren er derfor også selv en martingale.[4, s.86]

Konsistens af Kaplan-Meier

Dette afsnit er baseret på [1, s.112-121].

En estimator af endelig-dimensionelle parametre er konsistent hvis den euklidiske afstand mellem parameteren og dens estimator konvergerer mod 0 i sandsynlighed, eller med sandsynlighed 1 for stærk konsistent. En ikke-parametrisk estimator af en funktion, for eksempel overlevelsesfunktionen eller den kumulative hazard, er konsistent hvis afstanden målt i en metrik mellem funktionen og dens estimator konvergerer mod 0 i sandsynlighed. Supremumsnormen kan bruges som metrik.

Hvis en estimator er konsistent ved supremumsnormen siges estimatoren at være uniform konsistent. En estimator som er konsistent ved denne norm vil altså være uniformt tæt på den sande overlevelsesfunktion på hele intervallet for en tilstrækkelig stor stikprøve. Dette er nyttigt, da det ofte interessant at se på opførslen af overlevelsesfunktionen i et interval, i stedet for i få tidspunkter.

Sætning 2.4.2 (Konsistens af Kaplan-Meier)

Lad T være stokastisk variabel for hændelsestider med kontinuert fordelingsfunktion $F(s) = P(T \leq s)$, hvor $s \in [0, t]$, og kumulativ hazard funktion $H(s) = \int_0^s dF(v)/(1 - F(v))$. Hvis $t \in (0, \infty]$ er valgt således at $Y(t) \xrightarrow{P} \infty$ når $n \rightarrow \infty$ så

$$\sup_{0 \leq s \leq t} \left| \int_0^s \frac{dN(v)}{Y(v)} - H(s) \right| \xrightarrow{P} 0 \text{ når } n \rightarrow \infty,$$

og

$$\sup_{0 \leq s \leq t} |\hat{F}(s) - F(s)| \xrightarrow{P} 0 \text{ når } n \rightarrow \infty, \quad (2.11)$$

hvor $1 - \hat{F}$ er Kaplan-Meier estimatoren.

Beviset for sætningen kan findes i [1, s.117-120].

Når $F^n \equiv F$ så indikerer Sætning 2.4.2, at når $Y(t) \xrightarrow{P} \infty$ når $n \rightarrow \infty$ for ethvert $t \in [0, u]$, så er Kaplan-Meier estimatoren, $\hat{S}(t)$, en uniform konsistent estimator af $S(t)$ over $[0, u]$ så længe at $h(t) = \frac{-\frac{\partial}{\partial u} P(T \geq u, U \geq t)|_{u=t}}{P(T \geq t, U \geq t)}$, når $P(X > t) > 0$, gælder.

2.5 Cox proportional hazard model

Dette afsnit er baseret på [4, s.244-245, 253, 258].

Det er ofte interessant at sammenligne to eller flere gruppers tid til hændelse. I tilfældet hvor grupperne ligner hinanden, bortset fra behandlingen under studiet, så kan ikke-parametriske metoder anvendes direkte. Ofte er det dog sådan, at individerne i grupperne har flere karakteristika, som kan have betydning for udfaldet. Disse karakteristika kan være observerede demografiske variable, som alder eller køn, adfærdsmæssige variable, som spisevaner eller fysisk aktivitetsniveau, eller fysiologiske variable, som blodtryk eller blodsukkerniveau. Disse variable kan bruges, som kovariater til at forklare responsvariabel. Efter justering af disse forklarende variable, så burde sammenligningen mellem grupperne være mindre biased og mere præcis.

Lad X være tid til hændelse. Data indeholder en triple $(T_j, \delta_j, \mathbf{Z}_j(t))$, for $j = 1, \dots, n$, hvor T_j er tiden på studiet for individ j , δ_j er hændelsesindikator for individ j og $\mathbf{Z}_j(t) = (Z_{j1}(t), \dots, Z_{jp}(t))^T$ er en vektor af kovariater for individ j til tid t , som kan påvirke overlevelsesfunktionen for X . $Z_{jk}(t)$, for $k = 1, \dots, p$, kan være tidsafhængige kovariater, hvis værdi skifter over tid, eksempelvis løbende kontrol af blodtryk eller nuværende status på sygdom. Kovariater kan også være konstante, eksempelvis køn eller race. Det antages her, at kovariaterne er konstante.

Lad $h(t|\mathbf{Z})$ være hazard funktionen til tid t for et individ med risiko vektor $\mathbf{Z}$. Så er Cox modellen som følger:

$$h(t|\mathbf{Z}) = h_0(t) \exp(\boldsymbol{\beta}^T \mathbf{Z}),$$

hvor $h_0(\cdot)$ er en fuldstændig uspecificeret ikke-negativ funktion. Den skal være ikke-negativ, da en hazard funktion beskriver risikoen for, at et individ dør, og dette kan ikke være negativt. $h_0(\cdot)$ kaldes baseline hazard. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ er en parametervektor, og $\exp(\boldsymbol{\beta}^T \mathbf{Z})$ er en kendt funktion. Bemærk, at $\exp$ kan skiftes ud med en anden kendt funktion c . $\exp(\boldsymbol{\beta}^T \mathbf{Z}) = \exp(\sum_{k=1}^p \beta_k Z_k)$ kan bruges, da denne sikre at hazard raten er positiv.

Modellen kombinerer fleksibilitet, gennem den ikke-parametriske $h_0(\cdot)$, med muligheden for at indføre effekter af kovariater, gennem leddet $\exp(\boldsymbol{\beta}^T \mathbf{Z})$.

Cox modellen kaldes ofte en proportional hazard model, fordi det er muligt at se på ratioen mellem to individers kovariate værdier $\mathbf{Z}$ og $\mathbf{Z}^*$. Ratioen er givet ved:

$$\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z}^*)} = \frac{h_0(t) \exp(\sum_{k=1}^p \beta_k Z_k)}{h_0(t) \exp(\sum_{k=1}^p \beta_k Z_k^*)} = \exp\left(\sum_{k=1}^p \beta_k (Z_k - Z_k^*)\right), \quad (2.12)$$

hvilket er en konstant betinget $\mathbf{Z}$ og $\mathbf{Z}^*$. Det vil sige, at hazard raterne er proportionale. Ligning (2.12) kaldes den relative risiko for, at et individ med risiko faktor $\mathbf{Z}$ har en hændelse sammenlignet med et individ med risiko faktor $\mathbf{Z}^*$. Det kunne eksempelvis være, at Z_1 indikerer behandlingstype, $Z_1 = 1$ for behandling og $Z_1 = 0$ for placebo, og derudover er alle kovariater

ens. Så er:

$$\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z}^*)} = \exp(\beta_1),$$

risikoen for at få en hændelse, hvis individet modtager behandlingen relativ til risikoen for få en hændelse, når individet modtager placebo.

Standard metoder kan ikke bruges til at estimere β , når der ikke benyttes en parametrisk model for $h_0(\cdot)$. Derfor introduceres den partielle likelihood funktion, som kan bruges til at estimere komponenterne af β ved at maksimere med hensyn til β , selv når $h_0(\cdot)$ er fuldstændig uspecificeret.

Inden den partielle likelihood defineres så antages det, at alle t_i er forskellige og $D \subseteq \{1, \dots, p\}$ defineres som $D = \{i | \delta_i = 1\}$. Det vil sige, at D er indeks mængden for hændelsestidspunkter. For $t \geq 0$ defineres $R(t) = \{i | t_i \geq t\}$ som risikomængden. Det vil sige at $R(t)$ er indeksmængden for individer i risiko til tid t .

Den partielle likelihood funktion er givet ved:

$$L(\beta) = \prod_{i=1}^D \frac{\exp\left(\sum_{k=1}^p \beta_k Z_{(i)k}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^p \beta_k Z_{jk}\right)},$$

hvor $Z_{(i)k}$ er den k 'te kovariat associeret med individet hvis hændelsestidspunkt er t_i . Bemærk, at tælleren af den partielle likelihood funktionen kun afhænger af information fra de individer der har oplevet hændelsen, og nævneren indeholder information fra de individer der ikke har oplevet hændelsen endnu, dette inkluderer de individer, som senere bliver censureret.

Den partielle log-likelihood findes til at være:

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^D \left[\log \left[\exp \left(\sum_{k=1}^p \beta_k Z_{(i)k} \right) \right] - \log \left[\sum_{j \in R(t_i)} \exp \left(\sum_{k=1}^p \beta_k Z_{jk} \right) \right] \right] \\ &= \sum_{i=1}^D \sum_{k=1}^p \beta_k Z_{(i)k} - \sum_{i=1}^D \log \left[\sum_{j \in R(t_i)} \exp \left(\sum_{k=1}^p \beta_k Z_{jk} \right) \right]. \end{aligned}$$

At estimere β ved at maksimere $L(\beta)$ eller $\ell(\beta)$ afhænger ikke af $h_0(\cdot)$, eller de faktiske hændelsestidspunkter, kun deres rækkefølge, og de censurerede observationer optræder kun i risikomængden.

Den partielle likelihood for Cox modellen kan findes som profil likelihood funktion fra den

censurerede data likelihood, ligning (2.6), givet ved:

$$\begin{aligned} L(\beta, h_0(t)) &= \prod_{j=1}^n h(t_j | \mathbf{Z}_j)^{\delta_j} S(t_j | \mathbf{Z}_j) = \prod_{j=1}^n h_0(t_j)^{\delta_j} (\exp(\beta^T \mathbf{Z}_j))^{\delta_j} \exp(-H(t_j | \mathbf{Z}_j)) \\ &= \prod_{j=1}^n h_0(t_j)^{\delta_j} (\exp(\beta^T \mathbf{Z}_j))^{\delta_j} \exp\left(-\int_0^{t_i} h_0(u) du \exp(\beta^T \mathbf{Z}_j)\right), \end{aligned} \quad (2.13)$$

da $S(t) = \exp(-H(t))$ og $H(t) = \int_0^t h(u) du$.

Hold β fast og betragt maksimeringen af denne likelihood, som en funktion af $h_0(t)$. Antag, at $h_0(t_j) > 0$ for $j \in D$, og at $h_0(t) = 0$ mellem hændelsestidspunkter. Bemærk, at likelihood funktionen ikke giver information om $h_0(t)$ for t større end $\max_i t_i$. Så er den funktion, som skal maksimeres givet ved:

$$\begin{aligned} L(\beta, h_0(t)) &= \left(\prod_{l \in D} h_0(t_l) \exp(\beta^T \mathbf{Z}_l) \right) \exp\left(-\sum_{i=1}^n \exp(\beta^T \mathbf{Z}_i) \sum_{t_l \leq t_i} h_0(t_l)\right) \\ &= \left(\prod_{l \in D} h_0(t_l) \exp(\beta^T \mathbf{Z}_l) \right) \exp\left(-\sum_{l \in D} h_0(t_l) \sum_{i \in R(t_l)} \exp(\beta^T \mathbf{Z}_i)\right). \end{aligned}$$

Ved differentiation fås:

$$\frac{\partial}{\partial h_0(t_l)} \log L(\beta, h_0(t_l)) = \frac{1}{h_0(t_l)} - \sum_{j \in R(t_l)} \exp(\beta^T \mathbf{Z}_j).$$

Dette udtryk sættes lig nul og $h_0(t_l)$ isoleres:

$$\hat{h}_0(t_l) = \frac{1}{\sum_{j \in R(t_l)} \exp(\beta^T \mathbf{Z}_j)}.$$

Ved at indsætte dette i likelihood funktionen opnås profil likelihood funktionen for β :

$$L_p(\beta) = L(\beta, \hat{h}_0(t_l)) = \left(\prod_{l \in D} \frac{\exp(\beta^T \mathbf{Z}_l)}{\sum_{j \in R(t_l)} \exp(\beta^T \mathbf{Z}_j)} \right) \exp(-|D|),$$

hvilket er den partielle likelihood for Cox modellen.

Ud af det ovenstående fås også Breslow estimatoren for den kumulative hazards baseline for H_0 :

$$\hat{H}_0(t) = \sum_{l \in D: t_l \leq t} \frac{1}{\sum_{j \in R(t_l)} \exp(\beta^T \mathbf{Z}_j)}$$

3.1 Introduktion

Datasættet indeholder 1271 observationer af Hodgkin lymfekræft, hver observation indeholder 48 variable, som er demografiske variable, som alder og køn.

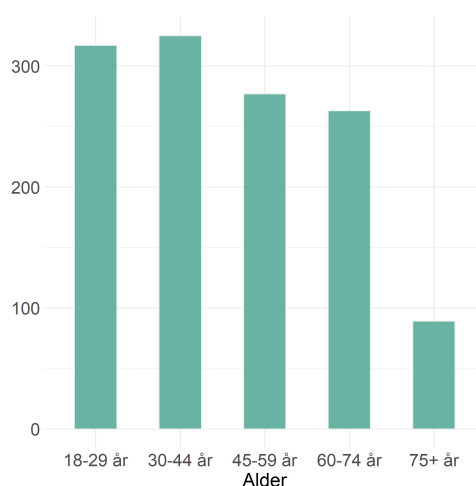
To af variablene kræver en smule forklaring, da det er kategoriske variable, som beskriver stadiet af sygdommen og patientens velbefindende. Disse to variable, Ann Arbor stadiet og ECOG scoren, kunne tænkes at have en stor betydning for overlevelsessandsynligheden.

Ann Arbor stadiet beskriver det stadiet, som lymfekræften er i. Ann Arbor stadiet bestemmes af placeringen af tumoren: Stadiet 1 indikerer, at lymfekræften er placeret i et enkelt område, eksempelvis en lymfeknude, og stadiet 4 indikerer, at kræften er spredt til et eller flere af de sekundære lymfoide organer, eksempelvis milten.[8]

ECOG scoren går fra 0 til 5. Denne variabel beskriver patientens generelle velbefindende og aktiviteter i sit daglige liv. En score på 0 betyder, at patienten er fuldt aktiv og kan klare alle aktiviteter, som før patienten blev syg, og en score på 4 betyder at patienten er fuldstændigt sengeliggende og ude af stand til at tage sig af sig selv.[9]

Datasættet indeholder flere variable, som er målt i blodet, samt indikatorer for hvilke områder der er involveret i lymfekræften. Ud af de 1271 observationer er 286 patienter døde.

Aldersfordelingen af patienterne kan ses på Figur 3.1. Det kan ses, at der er under 100 patienter med Hodgkin lymfekræft, som er over 75 år. Der er flest patienter i de to yngste aldersgrupper.



Figur 3.1: Aldersfordeling af patienter

Da patienterne indtræder i studiet, når de får konstateret Hodgkin lymfekræft, så er indtrædelsestidspunktet forskelligt for hver patient. Patienterne censureres ved studiets afslutning. Der-

med indeholder data højrecensurering, mere bestemt generaliseret type-1 censurering. Der er en del positiv korrelation på op til 0.58 mellem den forklarende variabel nodal og nogle af indikatorerne for hvilke områder der er involveret i kræften. Nodal beskriver antallet af involverede knudepunkter. Den største korrelation er mellem IgA og IgG, som er immunglobiner, der er en positiv korrelation på 0.604.

3.2 Manglende data

Et af problemerne ved datasættet er, at der for en stor del af observationerne er manglende værdier for nogle af de forklarende variable. Det betyder, at der for variabelen Beta2Microglobulin eksempelvis mangler værdier for cirka 60% af patienterne.

En mulighed er at fjerne variabelen fra datasættet, på grund af for stor andel af manglende data, dette gøres for nogle af variablene i afsnit 3.8. Antallet af manglende værdier og procenten af manglende værdier for alle variable kan ses i Tabel 3.1. Det ses blandt andet, at 33 af forklarende variable i datasættet har værdier for alle observationer, og at 2 af de forklarende variable mangler 8 værdier, hvilket svarer til at de mangler værdier for 0.63% af observationerne.

For variable med få manglende værdier kunne et gennemsnit af de observerede værdier indsættes, som et estimat på de manglende værdiers plads. For kategoriske variable kunne den mest observerede værdi indsættes. Dette kaldes imputation ved middelværdi.

Antal NA	Antal NA i %	Antal va- riable
0	0	33
4	0.31	1
8	0.63	2
10	0.79	1
16	1.26	1
23	1.81	1
25	1.97	1
30	2.36	1
81	6.37	1
117	9.21	1
118	9.28	1
128	10.07	1
364	28.64	1
411	32.34	1
762	59.96	1

Tabel 3.1: Første søjle viser antaller af NA for en variabel. Anden søjle viser hvad dette antal NA svarer til i procent, og sidste søjle er hvor mange af variablene der har det givne antal NA.

Fordelen ved at bruge et gennemsnit af de observerede værdier er at det ikke ændrer på middelværdien, men ulempen er, at det dæmper korrelationen mellem de involverede variable. Altså hvis en forklarende variabel er korreleret med en anden forklarende variable, så tages der ikke hensyn til dette ved at indsætte et gennemsnit. [5, s.61-62]

Det vælges, at den mest repræsenterede indsættes på pladserne, hvor der mangler værdier for de kategoriske variable, samt et gennemsnit for de numeriske variable, indsættes i de respektive søjler. Eksempelvis indsættes der på de tomme plads for den kategoriske variabel ECOG værdien 0, da dette er den mest repræsenterede værdi, og for den numeriske variabel tumordiameter indsættes gennemsnittet af de observerede værdier, som er 5.74. Et gennemsnit for Beta2Microglobulin er også indsat, til trods for den højde andel af manglende værdier. Dermed er der nu et fuldt datasæt til at estimere parametrene for de forklarende variable i de kommende modeller.

3.3 Forklarende variable

I dette afsnit vil der blive set nærmere på nogle af de kategoriske variable, og hvorvidt disse er signifikante i forhold til at beskrive tid til hændelse. En Kaplan-Meier kurve, altså en overlevelseskurve, laves med funktionen `survfit` fra R-pakken `survival`, ud fra et `Surv`-element lavet på observationstid og status, som beskrives af aldersgrupperne i datasættet. Dette gøres med følgende R-kode:

```
1 sfit_alder<-survfit(Surv(OS, status)~ aldersgruppe, data = Data)
2 alderplot<-ggsurvplot(sfit_alder, pval=TRUE, ggtheme=theme_minimal(),
3   legend.title="Alder", legend.labs=c("18-29_aar",
4   "30-44_aar", "45-59_aar", "60-74_aar", "75+_aar"))
```

Funktionen `ggsurvplot` fra `survminer`-pakken anvendes til at plotte Kaplan-Meier kurverne.

Der tilpasses altså en Kaplan-Meier kurve, hvor tid til hændelse, beskrives ud fra aldersgrupperne. Dette plot kan ses i højresiden af Figur 3.2.

Konfidensintervallerne for modellerne er log-transformerede og udregnet på følgende måde:

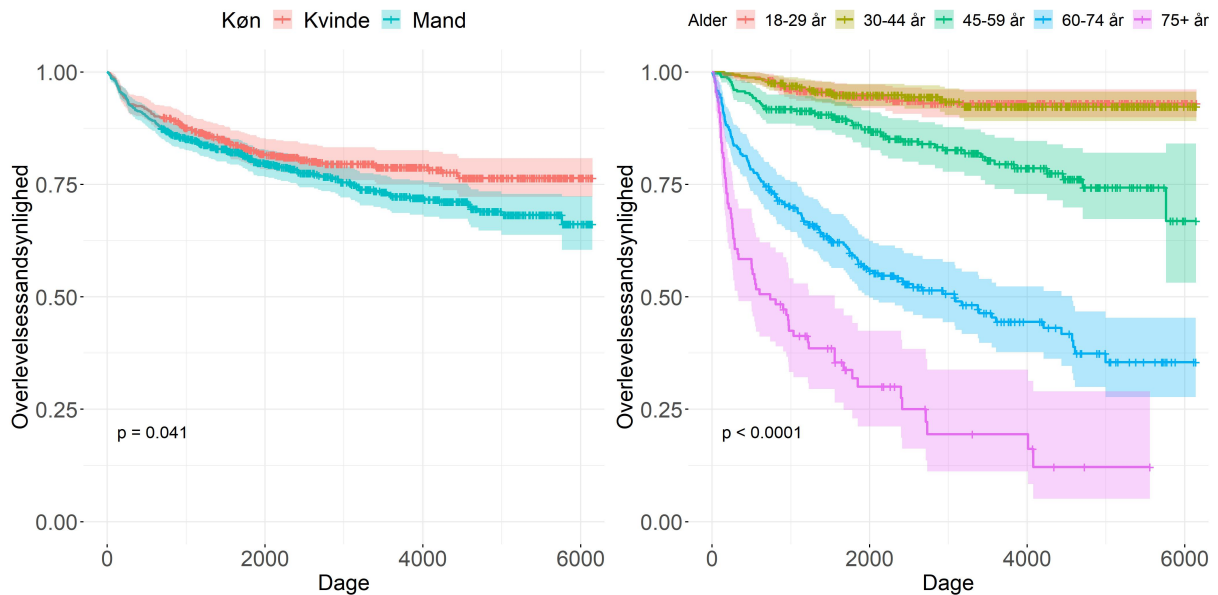
$$[\hat{S}(t_0)^{1/\theta}, \hat{S}(t_0)^\theta] \text{ hvor } \theta = \exp \left\{ \frac{Z_{1-\alpha/2} \sigma_S(t_0)}{\log[\hat{S}(t_0)]} \right\},$$

hvor $Z_{1-\alpha/2}$ er $1 - \alpha/2$ -percentilen af en standardnormalfordeling, og $\sigma_S^2 = \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$. [4, s.105]

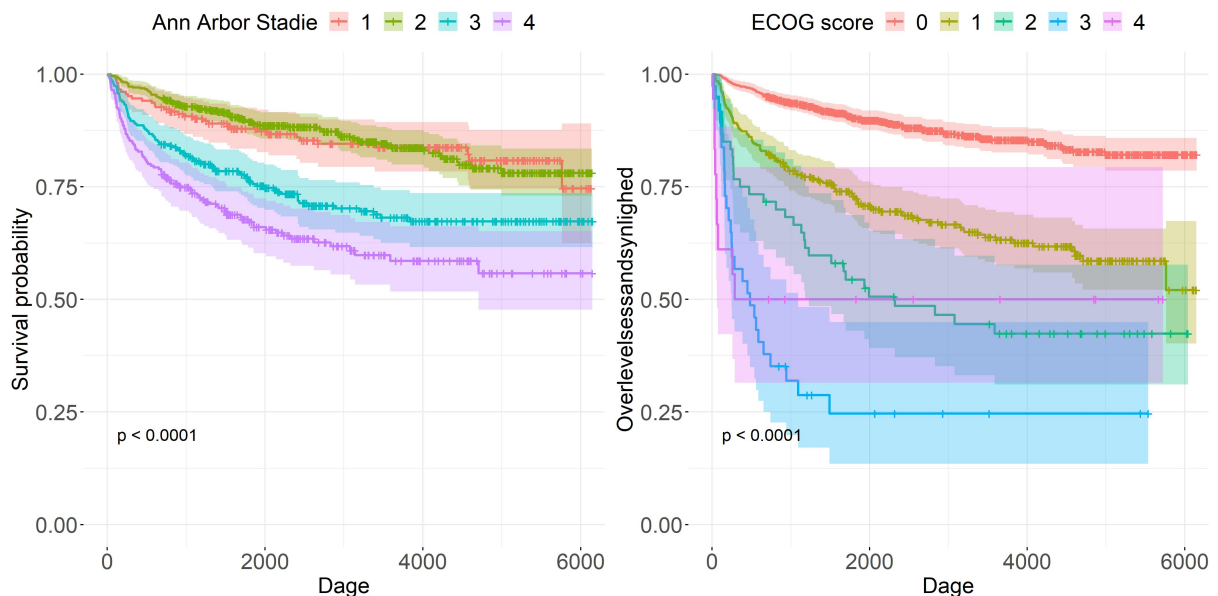
Ud fra Kaplan-Meier kurverne i højresiden af Figur 3.2 ses det, at jo ældre patienten er, jo mindre sandsynlighed for overlevelse. Særligt for de 75+ årige ses der et hurtigt fald, det vil sige, at der fra diagnosticeringstidspunktet til hændelsen typisk ikke går ret lang tid før hændelsen

indtræffer, samt at dette er aldersgruppen der har den laveste sandsynlighed for overlevelse. Der er dog ikke en markant forskel på to af aldersgrupperne, 18 – 29 år og 30 – 45 år.

Det samme kan laves i forhold til køn, og dette plot ses på venstre siden af Figur 3.2, hvor det kan ses at overlevelsessandsynligheden ved denne type kræft er mindre for mænd end for kvinder.



Figur 3.2: Kaplan-Meier kurver for køn og aldersgrupper med konfidensintervaller.



Figur 3.3: Kaplan-Meier kurver for Ann Arbor stadie og ECOG scoren med konfidensintervaller.

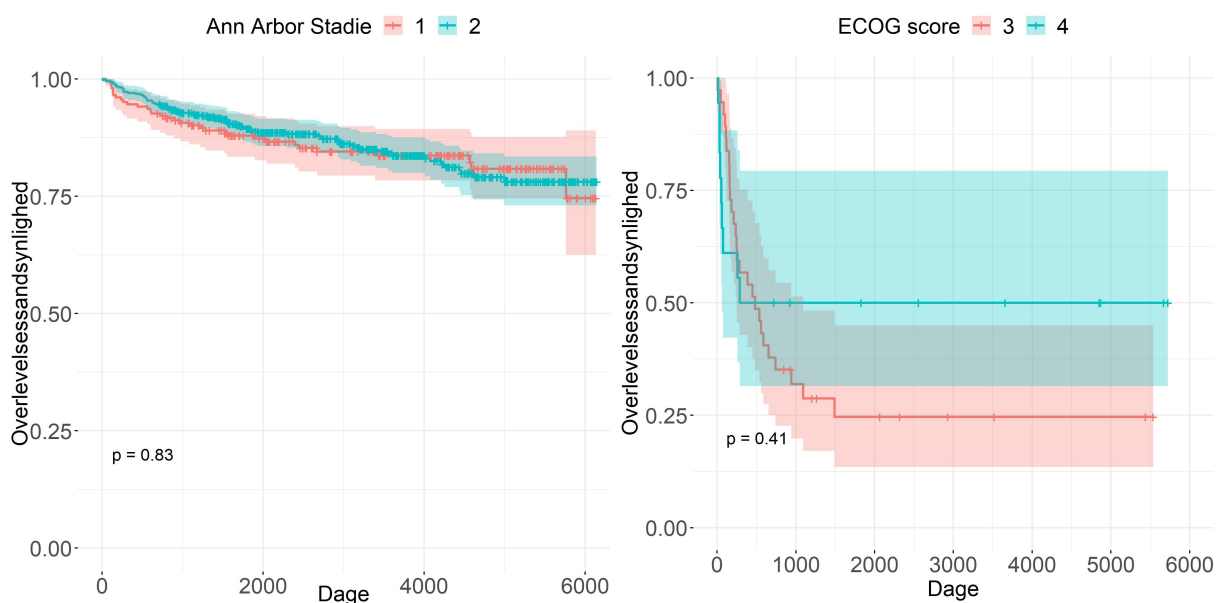
På Figur 3.3 ses Kaplan-Meier kurverne for Ann Arbor stadiet og ECOG scoren. For Ann Arbor stadiet kan det ses, at hvis patienten er på stadie 4, så er overlevelsessandsynligheden mindre

end på stadie 1, 2 og 3. For ECOG scoren ses der et hurtigt fald for patienter med score 3 og 4, dog er der kun 18 patienter med score 4, hvor af 9 har en ret kort tid til hændelse, og de sidste 9 bliver censureret efter længere tid, denne gruppe har også et tilsvarende stort konfidensinterval.

På alle fire plot i Figur 3.2 og 3.3 kan p -værdien ses, og den er for alle fire forklarende variable mindre end 0.05, og kan derfor ses som signifikante i forhold til at forklare tid til hændelse.

På de to plot i Figur 3.3 krydser nogle kurver hinanden. For Ann Arbor stadiet er det kurverne for stadie 1 og 2, og for ECOG er det kurven for score 4 som krydser både score 2 og 3. Det kan tyde på at proportional hazard antagelsen ikke er opfyldt for disse, hvilket er vigtigt da det ønskes at lave en Cox model for data. For ECOG kan det undersøges om score 4 er signifikant forskellig fra eksempelvis score 3, hvis ikke kunne der laves en gruppe hvor de to score samles. Det samme kan gøres for Ann Arbor stadie.

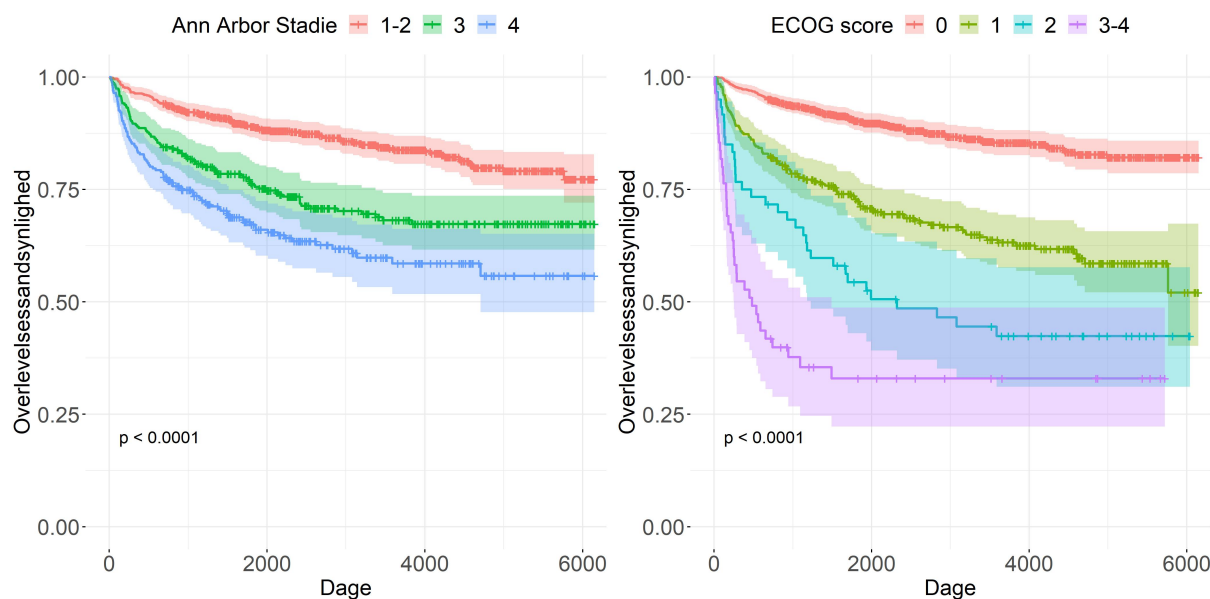
I Figur 3.4 er Kaplan-Meier kurverne for Ann Arbor stadie 1 og 2 plottet. Det kan ses, at en model med disse to kurver, har en p -værdi på 0.83, hvilket ikke er signifikant, derfor kan de to stadier sættes sammen i en gruppe. I figuren ses også Kaplan-Meier kurverne for ECOG score 3 og 4, her gør det samme sig gældende med en p -værdi på 0.41. Derfor sættes disse to score værdier også sammen i en gruppe.



Figur 3.4: Kaplan-Meier kurver for Ann Arbor stadie 1 og 2, og ECOG score 3 og 4 med konfidensintervaller.

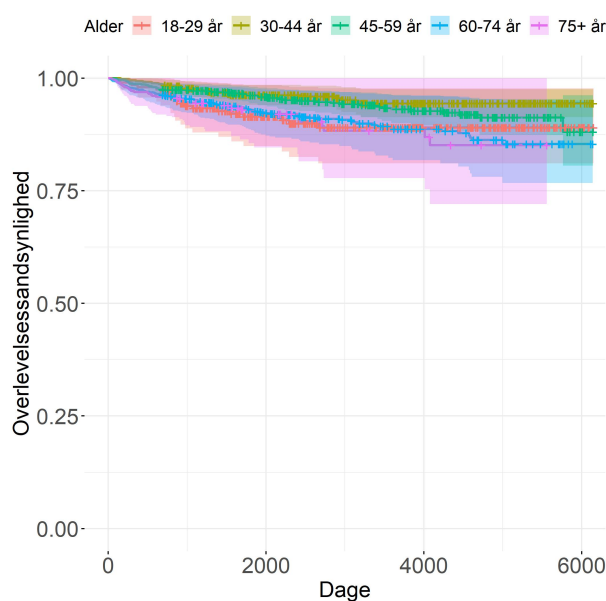
De andre værdi i de to forklarende variable er som før. Der laves nye overlevelsesfunktioner for de to forklarende variable, og disse er plottet i Figur 3.5.

På Figur 3.5 kan det ses, at både ECOG og Ann Arbor stadiet stadig er signifikant for respons variabelen. De krydsende kurver er nu væk, og det vil sige, at de opfylder proportional hazard antagelsen, som er at disse overlevelseskurver skal være proportionale med hinanden.



Figur 3.5: Kaplan-Meier kurver for den nye opdeling af Ann Arbor stadie og ECOG scoren med konfidensintervaller.

En Cox-model kan laves ud fra disse fire forklarende variable, samt tilstedeværelsen af B-symptomer, antallet af knuder i lymfatisk væv og antallet af knuder i lymfatisk celler i andre organer. Disse kategoriske variable er blevet undersøgt på samme vis som ovenfor, og er også fundet signifikante i forhold til at beskrive tid til hændelse. En enkelt kategorisk variabel er udeladt, da den ikke blev fundet signifikant. Denne variabel fortæller, om indholdet af LDH i blodet er forhøjet eller normalt. Modellens Kaplan-Meier kurver kan ses på Figur 3.6.



Figur 3.6: Kaplan-Meier kurver for model for de kategoriske forklarende variable.

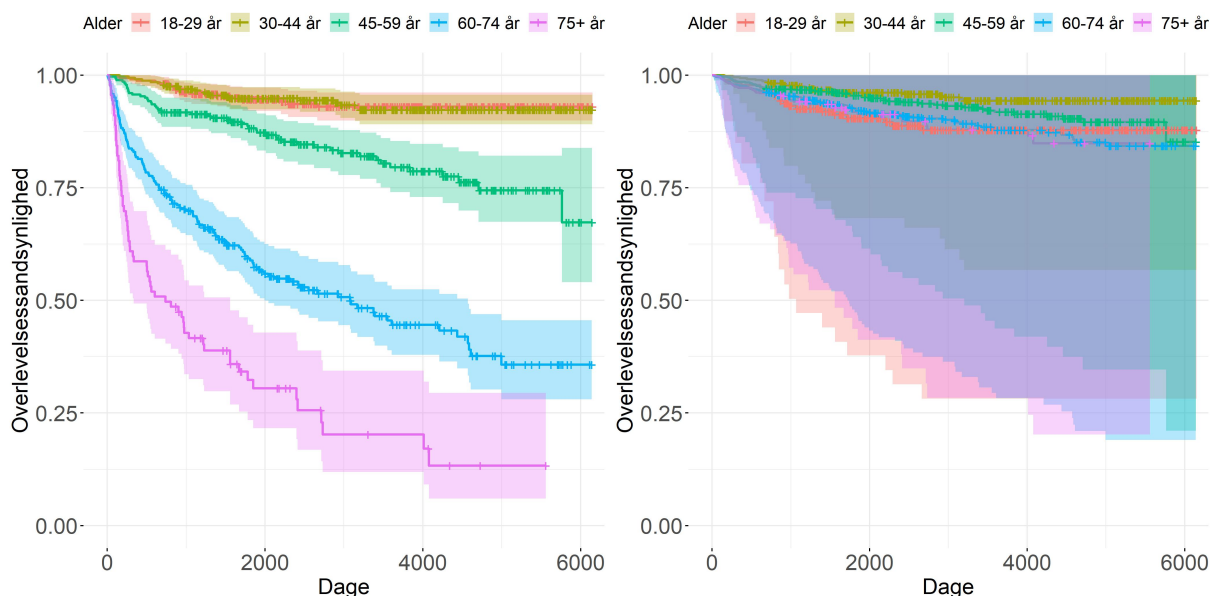
AIC-scoren kan udregnes ved $AIC = 2k - 2\log(\hat{L})$, hvor k er antal parametre i modellen

og $\hat{L}$ er den maksimale værdi af likelihood funktionen for modellen, og i tilfælde af en Cox proportional hazard model, så er det den partielle likelihood funktion. [4, s.253] AIC-scoren kan tjekkes for at se om en model passer på data, i forhold til antallet af parametre. Det er modellen med lavest score som er bedst. For modellen med de kategoriske variable er AIC scoren 2698.529.

På Figur 3.6 kan det ses, at konfidensintervallerne overlapper for alle aldersgrupper, hvilket tyder på at der ikke er så meget forskel i overlevelsessandsynligheden for de forskellige grupper. Derfor kunne det være nyttigt, at have en måde hvor på det er muligt at udvælge de forklarende variable, som ikke er kategoriske. En sådan metode kunne være LASSO, som præsenteres i afsnit 3.5.

3.4 Modeller

Inden LASSO anvendes vil det være en fordel at have lidt flere modeller at sammenligne med. Derfor laves nu en model med alle variable og en uden.



Figur 3.7: I venstresiden af figuren ses Kaplan-Meier kurver for nul-modellen, og i højresiden af figuren ses Kaplan-Meier kurver for den fulde model.

Der kan tilpasses en Cox model, med funktionen `coxph` fra `survival`-pakken i R, kun ud fra stratificering af aldersgrupper på datasættet, det giver en nul-model, hvor Kaplan-Meier kurverne er plottet i Figur 3.7.

I højresiden af figuren ses Kaplan-Meier kurverne for en Cox model hvor alle forklarende variable er med. Den fulde model har nogle meget brede konfidensintervaller, hvilket vil sige, at der er en sandsynlighed for, at der ikke er forskel på overlevelsessandsynligheden for alle alders-

grupperne, hvor imod at nul-modellen kun har to overlappende konfidensintervaller, dem for de to yngste aldersgrupper. Da alle konfidensintervallerne er store og overlapper hinanden i den fulde model er der en stor usikkerhed for hver af gruppernes overlevelsessandsynlighed. Dette er der ikke tilfældet i nul-modellen, og dermed tyder det på, at en model med færre forklarende variable er at foretrække.

AIC scoren for nul-modellen er 2788.17 og for den fulde model er $AIC = 2703.512$. Begge disse modeller har en højere AIC score end den kategoriske model fra afsnit 3.3, det vil altså sige, at disse modeller ikke beskriver data bedre. Det kunne derfor tyde på, at en model med færre parametre, dog flere end nulmodellen, er at foretrække, derfor laves en model ved at fjerne de insignifikante forklarende variable i en Cox model en efter en indtil alle variable i modellen er signifikante. Det kommer der følgende model ud af:

```
Call:
coxph(formula = Surv(OS, status) ~ Alder + Albumin + Leukocytter.mia.L. +
      CreatineMikroMol + Thrombocytter.mia.L. + Tonsilla.Palatina +
      Pleuraeksudat.med.lymfomceller + Hud + NyECOG + NyAnn + strata(aldersgruppe),
      data = data)
```

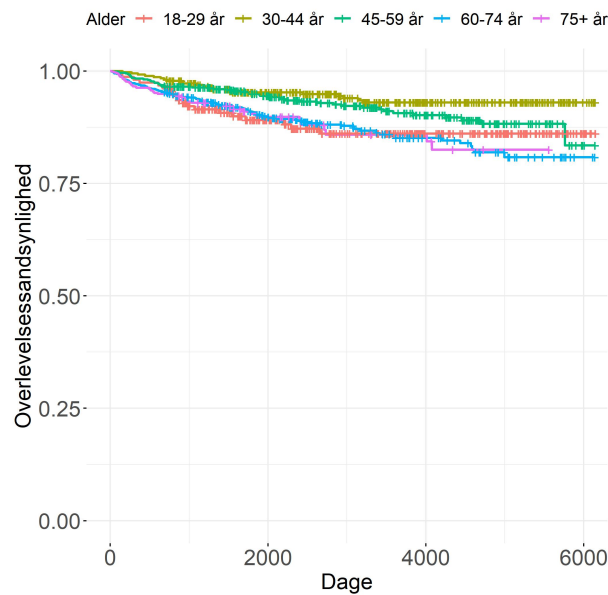
	coef	exp(coef)	se(coef)	z	p
Alder	0.0480900	1.0492651	0.0141352	3.402	0.000669
Albumin	-0.0394890	0.9612806	0.0101792	-3.879	0.000105
Leukocytter.mia.L.	0.0205402	1.0207526	0.0090409	2.272	0.023092
CreatineMikroMol	0.0077257	1.0077556	0.0016030	4.820	1.44e-06
Thrombocytter.mia.L.	-0.0013797	0.9986212	0.0004454	-3.098	0.001948
Tonsilla.Palatina	-0.8058275	0.4467181	0.3600569	-2.238	0.025217
Pleuraeksudat.med.lymfomceller	-1.3507975	0.2590336	0.3399041	-3.974	7.07e-05
Hud	-1.0650037	0.3447266	0.5299031	-2.010	0.044451
NyECOG1	0.4975259	1.6446472	0.1478663	3.365	0.000766
NyECOG2	0.5281694	1.6958252	0.2258215	2.339	0.019342
NyECOG3-4	0.8446753	2.3272220	0.2329063	3.627	0.000287
NyAnn3	0.3549712	1.4261396	0.1525429	2.327	0.019964
NyAnn4	0.6184103	1.8559753	0.1632839	3.787	0.000152

```

Likelihood ratio test=151.3 on 13 df, p=< 2.2e-16
n= 1271, number of events= 286

```

Kaplan-Meier kurverne for denne model kan ses på Figur 3.8. Denne model har en AIC score på 2663.898, og er indtil videre den model med lavest AIC, og derfor også den model der passer bedst på data i forhold til antallet af forklarende variable.



Figur 3.8: *Kaplan-Meier kurver for modellen kun med signifikante forklarende variable.*

Fælles for de tre ovenstående modeller er, at der er stratificeret efter aldersgruppe, og det kunne være at der var en bedre metode til at opdele. I afsnit 3.6 præsenteres den internationale prognose score, som er opbygget på et pointsystem, og det kunne være at en modifikation af dette system ville give en bedre model.

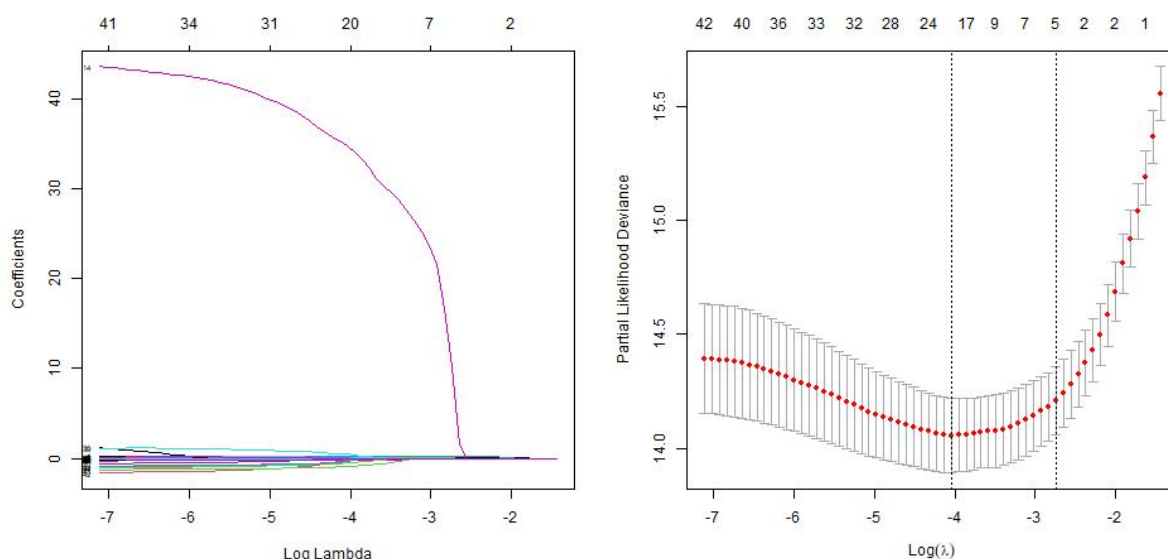
3.5 Variabel selektion med LASSO

Inden IPS-modellen introduceres så forsøges det at lave variabel selektion med LASSO metoden.

En måde hvor på de forklarende variable kan udvælges er ved brug af LASSO, som er en metode der har den fordel at den kan formindske parametrene for nogle af de forklarende variable til nul, hvilket giver en model der er lettere at fortolke. En introduktion til variabel selektion med LASSO metoden kan findes i Appendiks B.

LASSO anvendes på datasættet, hvor 33 ud af 48 variable er kategoriske variable, resten er numeriske. Det ønskes at lave en Cox model, hvor tid til hændelse og hændelse beskrives ud fra de forklarende variable. Et plot for LASSO udført på data kan ses på venstresiden af Figur 3.9. Figuren læses fra højre mod venstre. Hver af stierne repræsenterer hvordan koefficienterne for de forklarende variable opfører sig til forskellige værdier af λ . Stierne som starter først, altså dem som ikke er lig 0 først, er de forklarende variable som har betydning for responsvariablen. På LASSO plottet i Figur 3.9 kan det ses at en sti skiller sig meget ud, og har en ret høj værdi, dette er stien for Beta2Microglobulin. De andre stier ligger meget samlet over og under nul.

Det kan ses på figuren, at når $\log(\lambda) = -4 \Rightarrow \lambda = 0.018$ indeholder modellen 20 forklarende variable.



Figur 3.9: I venstresiden af figuren ses LASSO udført på data, og på højresiden ses en 10-fold krydsvalidering udført på data.

For at finde det optimale λ bruges krydsvalidering, da det er for en Cox model så bruges deviance som prædiktionsfejls mål. [2, s.20] Deviance er givet ved $-2 \cdot \ell(\beta)$, hvor ℓ er den partielle log-likelihood for den tilpassede model.

Først opdeles data, sådan at et `Surv`-element laves på variablene tid til hændelse og status, og resten bliver et data element med de forklarende variable. `Surv`-elementet er responsvariablen i funktionen `glmnet`, som er en funktion som tilpasser en Cox model med straffet maksimum likelihood, og det er standard funktionen, som bruges til at lave LASSO i R. Funktionen `cv.glmnet` anvendes til at udføre 10-fold krydsvalidering, en sådan krydsvalidering kan ses på højresiden af Figur 3.9. Da alle krydsvalideringer er forskellige laves en funktion, som udfører krydsvalideringen l antal gange, og som derefter tager middelværdien på de l $\lambda_{\min}$ og l λ_{1se} . $\lambda_{\min}$ er den værdi som giver den mindste krydsvalideringsfejl, det svarer til mindste kvadratersfejl, og λ_{1se} er den største værdi af λ indenfor en standard afvigelse fra $\lambda_{\min}$, det er den mest regulerede model. Herefter laver funktionen to Cox modeller med `glmnet`, hvor hver af de to middelværdierne bruges som λ i hver sin model. Funktionen printer for begge modeller de udvalgte forklarende variable, samt deres β -koefficient og det samlede antal variable i hver model. Jo højere valg af l , jo flere krydsvalideringer tages gennemsnittet over. Det vil altså sige, at et stort l giver et bedre bud på λ , end et lille l .

Anvendes denne funktion på data for forskellige værdier af l fås resultaterne som er opsummeret i Tabel 3.2.

Antal krydsvalidering	$\lambda_{\min}$	Antal signifikante variable	λ_{1se}	Antal signifikante variable
5	0.02455	16	0.05797	6
10	0.02292	16	0.05976	6
100	0.02275	16	0.05872	6
500	0.02272	16	0.05747	6
1000	0.02270	16	0.05760	6

Tabel 3.2: 10-fold krydsvalidering for en Cox model på data udført henholdsvis $l = \{5, 10, 100, 500, 1000\}$ gange. Middelværdien af de l værdier af $\lambda_{\min}$ og λ_{1se} udregnes.

Da λ -værdierne i Tabel 3.2 er et gennemsnit af de λ som er fundet ved krydsvalidering, må det λ fundet ved brug af 1000 krydsvalideringer være et godt bud på det optimale λ . $\lambda_{\min}$ vælges, da dette giver en model med lidt flere parametre.

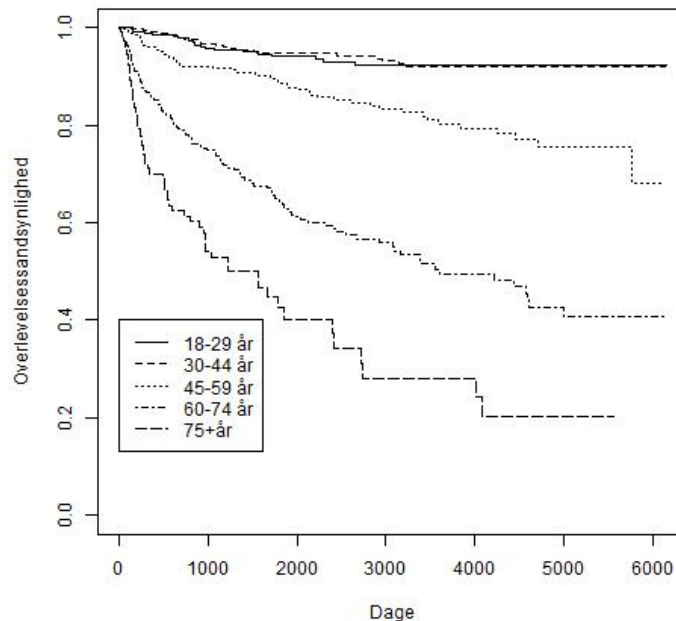
De 16 signifikante variable for LASSO med $\lambda_{\min} = 0.02270$ kan ses i Tabel 3.3. I tabellen ses at Beta2Microglobulins parameter har en meget høj værdi, hvilket også kunne ses på venstresiden af Figur 3.9, hvor den også var den første parameter som var vigtig. Det var også den parameter som havde flest manglende værdier fra start, så cirka 60 % af værdierne for denne parameter er estimeret ved et gennemsnit af de øvrige.

Forklarende variabel	β	Forklarende variabel	β
Alder	0.04538	Creatine	0.00484
Ann Arbor stadie	0.15684	Thrombocytter	-0.00012
ECOG	0.28391	Tonsilla Palatina	-0.28049
B-symptomer	0.04676	Knoglemarv	-0.18563
Albumin	-0.02616	Mamma	-0.22736
Lymfocytter	0.00706	Tyktarm	0.24854
Beta2Microglobulin	33.38235	Pleuraeksudat med lymfoceller	-0.77022
IgM	-0.01103	Aldersgruppe	0.1568429

Tabel 3.3: Udvalgte forklarende variable for Cox regressions model med $\lambda = 0.02165$, samt deres parameterværdi, β .

Det kan derudover også ses, at nogle af de forklarende variable som blev valgt i afsnit 3.3 også er med i blandt de forklarende variable, som er valgt ved hjælp af LASSO. Det vil sige, at alder, Ann Arbor stadie, ECOG, B-symptomer og aldersgrupper er de samme for modellerne, men

LASSO har derudover en række forklarende variable, som stammer fra målinger i blodet og parametre der indikerer spredning til bestemte organer.



Figur 3.10: *Kaplan-Meier kurver for modellen lavet med LASSO.*

I Figur 3.10 ses Kaplan-Meier kurverne for modellen lavet med LASSO. Det kan ses, at overlevelsessandsynligheden er ret lav for de patienter som er 75 år eller over, når observationstiden kommer over 4000 dage så er der kun 20 procent af patienterne der er i live for dem over 75 år, hvor det for aldersgruppen 45-59 år er omkring 80 procent der stadig er i live.

Denne Cox model lavet med LASSO straf har en AIC score på 2695.938, hvilket er lavere end kategoriske model fra afsnit 3.3, samt nul-modellen og den fulde model fra afsnit 3.4, men LASSO modellen har en højere AIC score end modellen kun med signifikante variable, og det er derfor stadig denne model der beskriver data bedst.

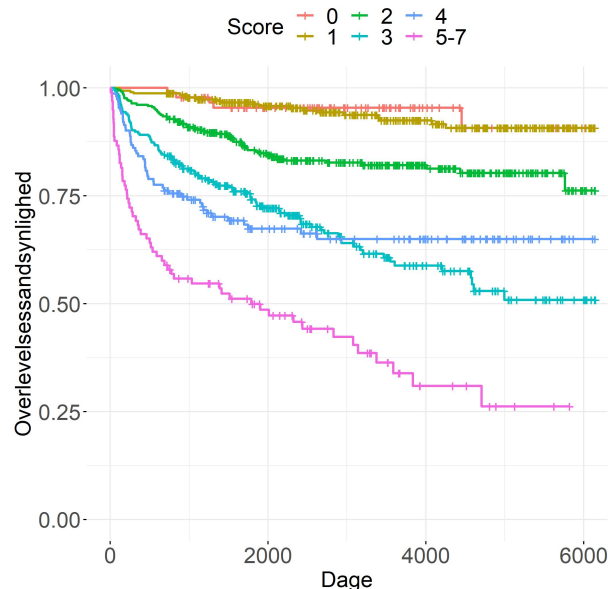
3.6 IPS-model

Den internationale prognose score (IPS) for Hodgkin lymfekræft er opbygget på et pointsystem. IPS-modellen er et redskab der hjælper læger med at forudsige en overlevelsessandsynlighed for deres patienter med Hodgkin lymfekræft. For hvert af følgende udsagn der er opfyldt, så får patienten et point:

- Ældre end 45 år
- Mand
- Ann Arbor stadie 4

- Albumin niveauet i blodet er mindre end 4g /dl
- Hæmoglobin niveauet i blodet er mindre end 10.5g/dl
- Antallet af hvide blodlegemer er større end eller lig 15.000/mikroL
- Det absolutte antal af lymfocytter i blodet er mindre end 600/mikroL og/eller under 8% af antallet af hvide blodlegemer

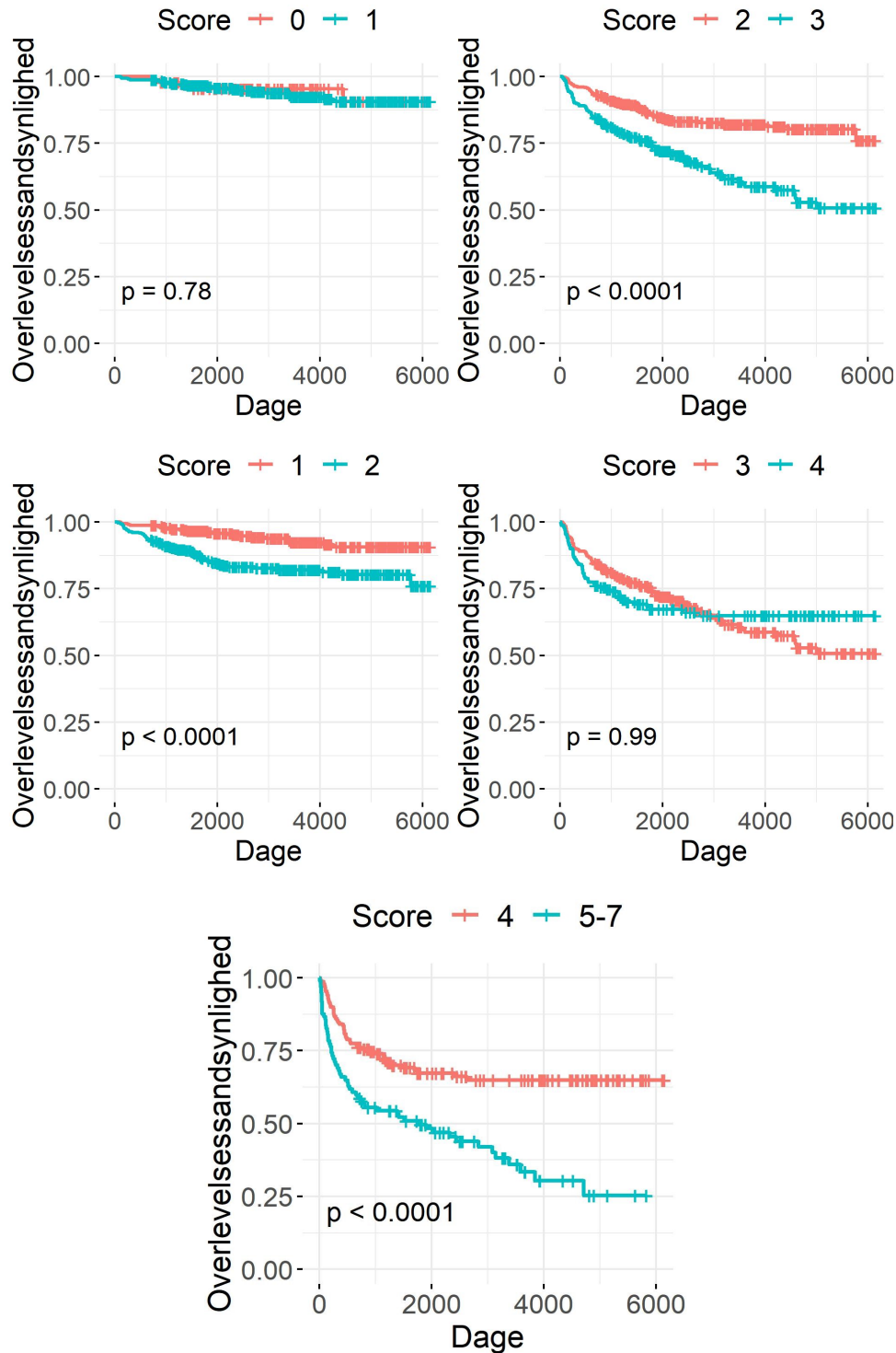
Dermed er den maksimale score 7. Dette pointsystem anvendes på datasættet. Observationerne opdeles i grupper efter hvor mange point de har, da der ikke er ret mange patienter med 5, 6 og 7 point, slås disse tre grupper sammen til en. Eksempelvis er der kun en patient med 7 point. Kaplan-Meier kurverne for disse grupper plottes, og kan ses på Figur 3.11.



Figur 3.11: *Kaplan-Meier kurver for IPS-modellen.*

Det kan ses på Figur 3.11, at jo flere point, jo lavere overlevelsessandsynlighed. Eksempelvis er overlevelsessandsynligheden for gruppen med 5-7 point efter 5 år omkring 50%, hvor den for dem med 2 point er omkring 85%.

Det kunne også være interessant at se om forskellen mellem grupperne parvist er signifikante. Det viser sig, at forskellen mellem at have 0 eller 1 point ikke er signifikant og forskellen på at have 3 eller 4 point heller ikke er signifikant. Dette kan ses på Figur 3.12. Det kan derudover ses på både Figur 3.11 og 3.12, at Kaplan-Meier kurverne for 3 point og 4 point krydser hinanden, det vil sige, at de ikke overholde proportional hazard antagelsen.



Figur 3.12: Kaplan-Meier kurver for grupperne parvist.

Sammenlignes de forklarende variable i IPS-modellen med de forklarende variable i modellen lavet med LASSO og modellen med de signifikante parametre, så kan det ses, at ECOG, thrombocytter samt om variablene tonsilla palatina og pleuraeksudat med lymfeceller indikerer, at disse steder er involveret i kræften, er med i de to andre modellen, men de tre forklarende variable er ikke med i IPS-modellen. Derudover så er hæmoglobin og køn ikke med i nogle af

de 2 forrige modeller, men med i IPS-modellen. De tre modeller har alder, Ann Arbor stadie og albumin tilfælles.

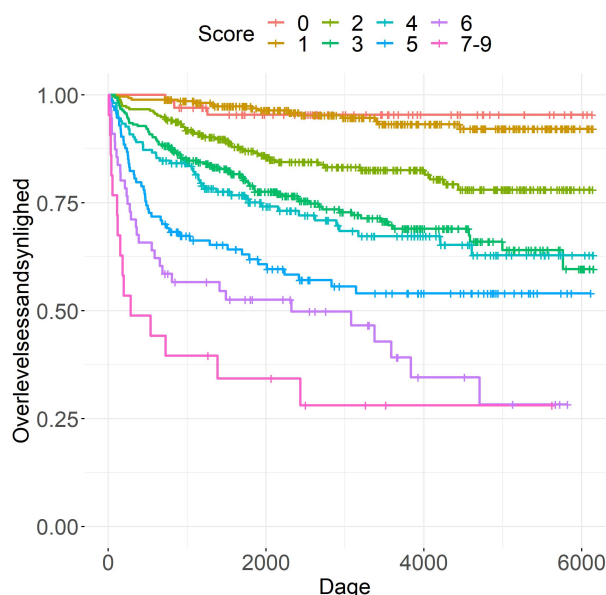
På baggrund af dette laves en ny model med nogle af de forklarende variable i LASSO modellen og modellen med de signifikante variable kombineres med IPS-modellen, således at modellen stadig bygger på et pointsystem, da dette har en god anvendelse i praksis. Værdien for thrombocytter, altså blodplader, vælges ud fra normalværdien, som er mellem 145 – 400 for mænd og kvinder. [7]

I den nye model tilføjes følgende udsagn til IPS-modellen:

- ECOG score 3-4
- Er pleuraeksudat med lymfocytter involveret i kræften?
- Er tonsilla palatina involveret i kræften?
- Antallet af blodplader er større end 400 mia/L

Det vil altså sige, at den maksimale score nu er 11. Det er der dog ikke nogle af patienterne i datasættet der har.

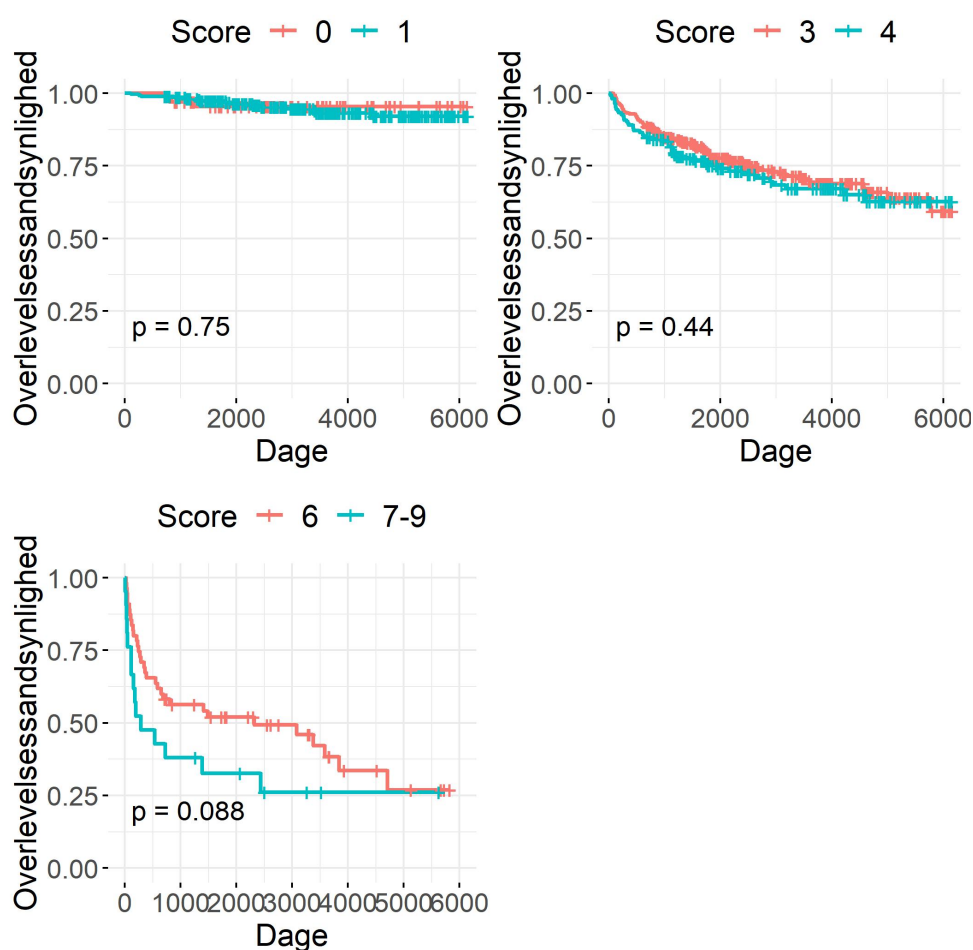
For denne model er der igen ikke ret mange tilfælde med 7, 8 og 9 point, henholdsvis 16, 4 og 1. Disse grupper sammensættes derfor. Kaplan-Meier kurverne for den nye model plottes, og disse kan ses på Figur 3.13.



Figur 3.13: *Kaplan-Meier kurver for den nye model.*

I forhold til IPS-modellen er der selvfølgelig flere kurver, da den maksimale score er højere. Det kan ses, at der for den nye model stadig er nogle kurver der krydser hinanden, samt at der ikke er den store forskel på at have 0 point og 1 point, derudover ser det ud til, at forskellen på overlevelsessandsynligheden for 3 og 4 point heller ikke er så stor. Dette undersøges på samme vis, som for IPS-modellen. Kun de parvise forskelle som ikke er signifikante plottes i Figur 3.14. På figuren ses det, at forskellen mellem 0 og 1 point ikke er signifikant, samt forskellen

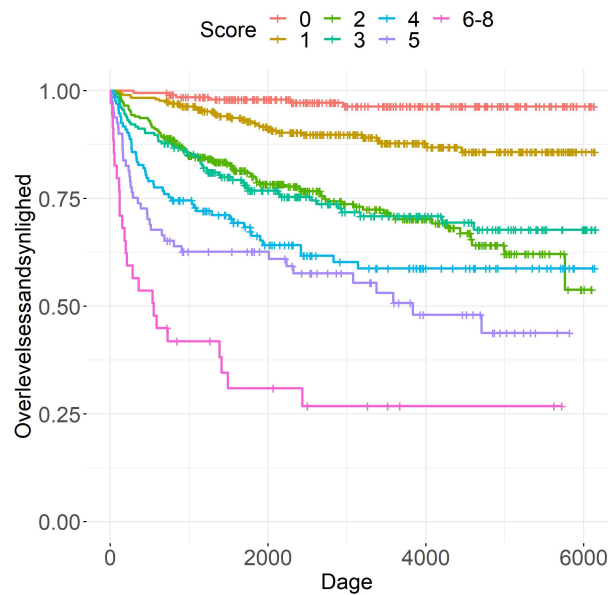
på 3 og 4 point. Derudover kan det ses at forskellen på at have 6 og 7-9 point heller ikke er signifikant, hvilket var mindre åbenlyst end for de to andre grupper, da der er ret stor forskel på kurverne fra dag 0 til cirka dag 4500.



Figur 3.14: Kaplan-Meier kurver for de ikke-signifikante forskelle på point-grupperne for den nye model.

Den nye model har en AIC score på 2684.026, hvor IPS-modellen har en AIC score på 2821.882, det vil altså sige, at den nye model beskriver data bedre.

Da hverken LASSO eller modellen med de signifikante variable havde den forklarende variable køn med, kunne det forsøges at fjerne denne fra modellen. Det giver en model med en AIC score på 2720.045, hvilket er bedre end IPS-modellen, men dårligere end den nye model. En af fordelene ved denne model er dog, at forskellen mellem at have 0 og 1 point nu er signifikant med en p -værdi på 0.0015. Det betyder, at der nok var en del af de patienter med 1 point, som kun opfyldte det udsagn, at de var mand. Kaplan-Meier kurverne for denne model kan ses på Figur 3.15. På figuren ses, at kurverne for 0 og 1 point ikke længere ligger oveni hinanden, samt at kurverne for 2 og 3 point næsten svarer til kurver for 3 og 4 point i Figur 3.13. Sammenlignes kurver for de to højeste antal point så er der ret stor forskel på disse i Figur 3.13 og Figur 3.15.



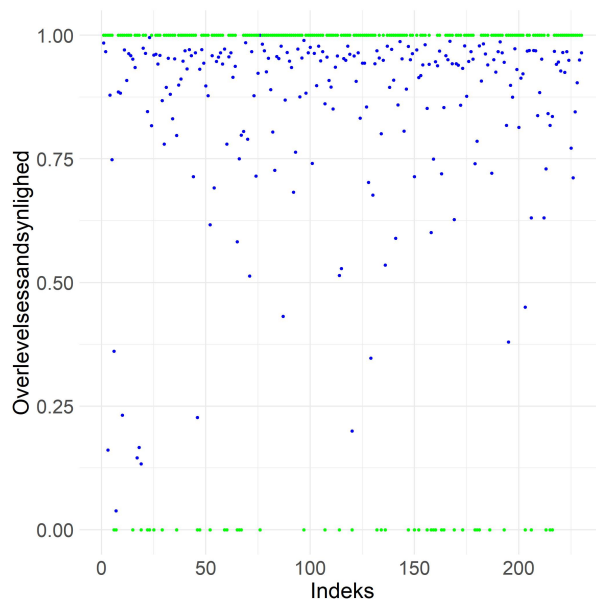
Figur 3.15: *Kaplan-Meier kurver for modellen uden køn.*

Fjernes variablen hæmoglobin også, fås en model med en AIC score på 2809.324, hvilket er også er højere end den nye model. Disse to model er derfor ikke bedre, og derfor fortsættes der med den nye modificerede IPS-model, hvor både køn og hæmoglobin indgår.

3.7 Test af modeller

Ud fra de forrige afsnit vælges det at teste de to modeller med lavest AIC score, det vil sige den modificerede IPS-modellen med en AIC score på 2684.026 og modellen med de signifikante parametre med en AIC score på 2662.898. Testen udføres ved at anvende de to modeller på et test-datasæt, som fra starten har være udeholdt af det data, som modellerne er tilpasset på.

Først testes modellen med de signifikante parametre ved brug af funktionen `predict` på modellen på test-datasættet. Funktionen prædikerer overlevelsessandsynligheder for hver af patienterne i test-datasættet. Denne overlevelsessandsynlighed for hver af patienterne plottes sammen med patientens virkelige status. Dette kan ses på Figur 3.16, hvor de grønne prikker er om patientens virkelige status er død, 0, eller censureret, 1, (dette er omvendt af resten af databehandlingen) og de blå prikker er den estimerede overlevelsessandsynlighed.



Figur 3.16: De prædikterede overlevelsessandsynligheder med modellen med signifikante variable for test-datasættet (blå), samt patienternes oplyste status, 0 =død, og 1 =censureret(grønne).

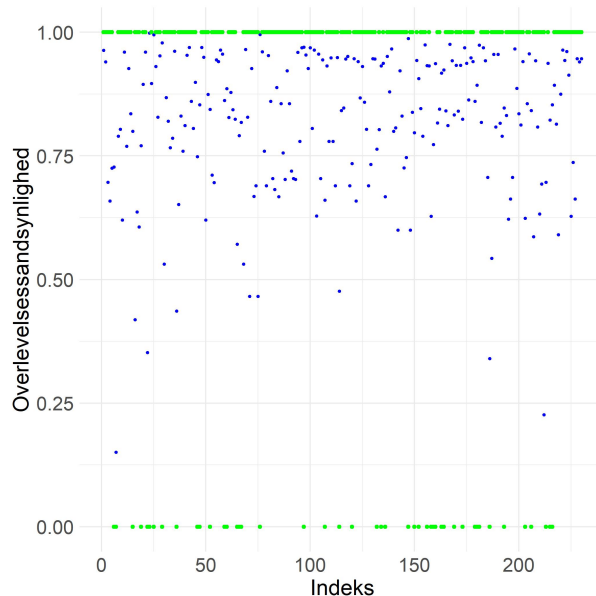
På Figur 3.16 kan det ses, at der er en del flere patienter som er døde, end patienter med lave overlevelsessandsynligheder. Middelværdien af forskellen på den prædikterede overlevelsessandsynlighed og status på patienten udregnes for alle 230 patienter, denne middelværdi bliver 0.238. Det vil altså sige, at denne model i gennemsnit er 23.8% fra det rigtige resultat.

De gennemsnitlige prædikterede overlevelsessandsynligheder for de 5 aldersgrupperne er givet i Tabel 3.4, hvor det kan ses at der er en mindre overlevelsessandsynlighed, jo ældre aldersgruppen er. De gennemsnitlige overlevelsessandsynligheder for mænd og kvinder kan laves på samme vis, og er henholdsvis 0.835 og 0.881, det vil altså sige, at den gennemsnitlige prædikterede overlevelsessandsynlighed er større for kvinder end mænd.

Aldersgruppe	18-29 år	30-44 år	45-59 år	60-74 år	75+ år
Gennemsnitlig overlevelssandsynlighed	0.948	0.943	0.873	0.646	0.531

Tabel 3.4: Overlevelsessandsynligheden for de 5 aldersgrupper med modellen med signifikante variable.

Derefter testes den modificerede IPS-model på samme vis, det giver Figur 3.17. Det samme gør sig gældende for denne model, der ses også færre lave overlevelsessandsynligheder, end døde patienter. Middelværdien for denne model er 0.296, altså højere end for den forrige model. Det vil altså sige, at denne model er længere fra det rigtige resultat.



Figur 3.17: De prædikterede overlevelsessandsynligheder med den modificerede IPS-model for test-datasættet (blå), samt patienternes oplyste status, 0 =død, og 1 =censureret(grønne).

For denne model kan der opstilles en tabel over overlevelsessandsynligheder estimeret med den modificerede IPS-model for et halvt år, et år og to år kan ses i Tabel 3.5. Her kan det også ses at forskellen på overlevelsessandsynligheden ikke er ret stor, når patienten har 0 eller 1 point, hvilket også var det som blev fundet ikke-signifikant i afsnit 3.6.

Score	1/2 år	1 år	2 år
0	0.992	0.986	0.980
1	0.991	0.983	0.975
2	0.969	0.947	0.920
3	0.947	0.909	0.865
4	0.939	0.896	0.856
5	0.904	0.837	0.764
6	0.850	0.752	0.649
7-9	0.754	0.611	0.473

Tabel 3.5: Overlevelsessandsynligheden for patienter med de forskellige scorer for et halvt år, et år og to år.

Fordelen ved den modificerede IPS-model er, at den er let at bruge i praksis. En læge kan altså blot tælle point sammen, og tjekke en graf eller et skema for at se overlevelsessandsynligheden for sin patient, hvor imod der i modellen med de signifikante variable skal laves udregninger, og den er derfor ikke lige så brugervenlig, som den modificerede IPS-model.

Ulempen ved den modificerede IPS-model er, at usikkerheden for den givne overlevelsessandsynlighed er større end for modellen med de signifikante variable, da denne vil give et mere

specifikt estimat ud fra eksempelvis variabelen albumin, hvor den vil bruge det konkrete tal gange parameterværdien, hvor IPS-modellen kun ser på om niveauet af albumin er mindre end 4g/dl.

3.8 Et reduceret datasæt

I Tabel 3.1 kunne det ses, at flere af de forklarende variable havde mange manglende værdier, derfor forsøges det også at lave et datasæt hvor de fleste af disse værdier fjernes medmindre de synes vigtige i forhold til sygdommen. De forklarende variable med flest manglende værdier er Beta2Microglobulin, tumordiameter, Creatine, LDH, IgG, IgA og IgM; disse variable fjernes fra data. Albumin har 81 manglende værdier, men da den har været signifikant i alle de forrige modeller, samt at den er en variabel i IPS-modellen, så vælges det at beholde denne. Det giver et datasæt med 41 variable. Herefter fjernes de patienter, som har manglende værdier i nogle af de tilbageværende variable, hvor der tidligere blev estimeret en værdi for disse patienter. Det reducerer antallet af patienter i datasættet fra 1271 til 1133.

Der laves nu en model for dette datasæt, som kun indeholder signifikante variable på samme måde, som i afsnit 3.4. Det giver følgende model:

Call:

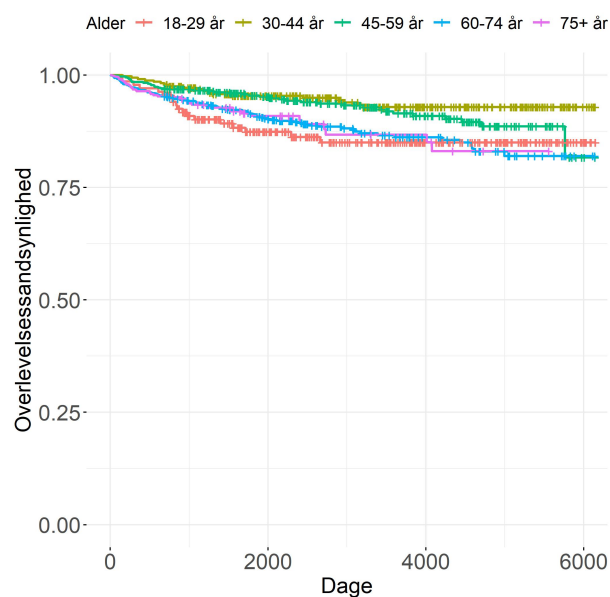
```
coxph(formula = Surv(OS, status) ~ Alder + Albumin + Lymfocytter.mia.L. +
      Tonsilla.Palatina + Lungehili + Pleuraeksudat.med.lymfomceller +
      Knogler + NyECOG + NyAnn + strata(aldersgruppe), data = nydata)
```

	coef	exp(coef)	se(coef)	z	p
Alder	0.05325	1.05470	0.01525	3.493	0.000478
Albumin	-0.04277	0.95813	0.01085	-3.944	8.03e-05
Lymfocytter.mia.L.	0.02583	1.02617	0.01134	2.278	0.022748
Tonsilla.Palatina	-1.07053	0.34283	0.39467	-2.712	0.006679
Lungehili	-0.33971	0.71198	0.16695	-2.035	0.041868
Pleuraeksudat.med.lymfomceller	-1.21849	0.29568	0.34090	-3.574	0.000351
Knogler	0.56589	1.76102	0.26520	2.134	0.032856
NyECOG1	0.35663	1.42851	0.16030	2.225	0.026099
NyECOG2	0.40156	1.49415	0.23666	1.697	0.089737
NyECOG3-4	0.94421	2.57077	0.25542	3.697	0.000218
NyAnn3	0.27600	1.31785	0.16418	1.681	0.092749
NyAnn4	0.60478	1.83085	0.18933	3.194	0.001402

Likelihood ratio test=126.7 on 12 df, p=< 2.2e-16

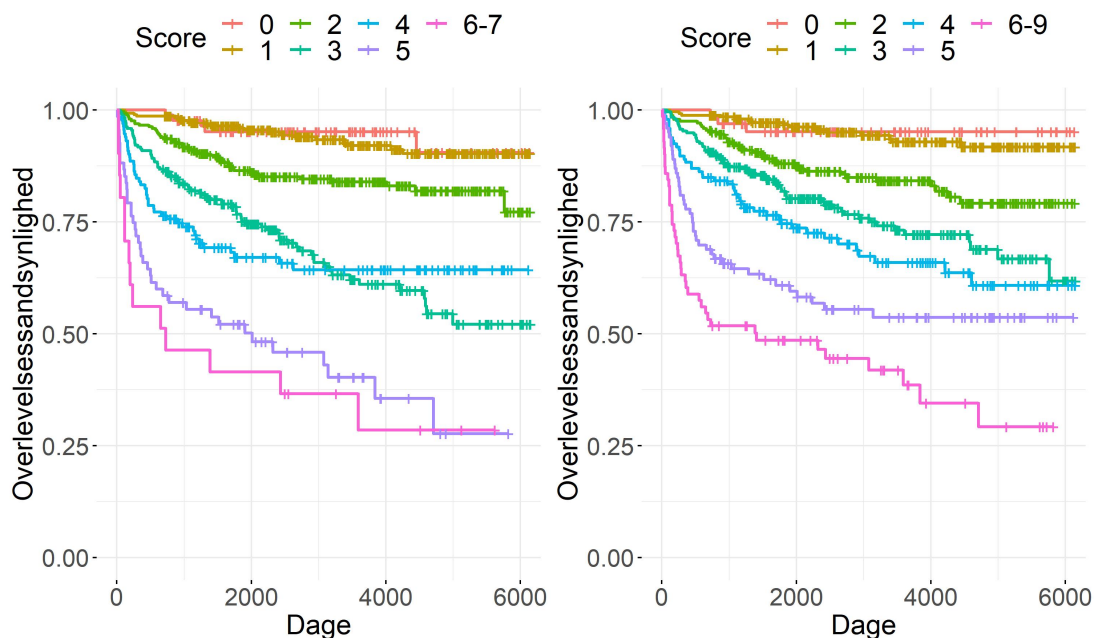
n= 1133, number of events= 244

Overlevelseskurverne for denne model kan ses på Figur 3.18, hvor det kan ses at den yngste aldersgruppe ret hurtigt får den laveste overlevelsessandsynlighed, samt at aldersgrupperne 30–44 og 45–59 år er dem der har de højeste overlevelsessandsynligheder fra starten. Denne model har en AIC score på 2212.734, men denne kan ikke sammenlignes med AIC scoren for de forrige modeller, da denne er lavet på et reduceret datasæt.



Figur 3.18: Kaplan-Meier kurverne for modellen med signifikante variable på det reducerede datasæt.

Sammenlignes de forklarende variable i denne model med de forklarende variable i modellen lavet på det fulde datasæt, så er der forskel i de signifikante variable. I denne model er lymfocytter, lungehili og knogler signifikante, hvilket de ikke er i modellen i afsnit 3.4, men de har pleuraeksudat med lymfomceller, tonsilla palatina, alder, albumin, ECOG score og Ann Arbor stadie tilfælles.

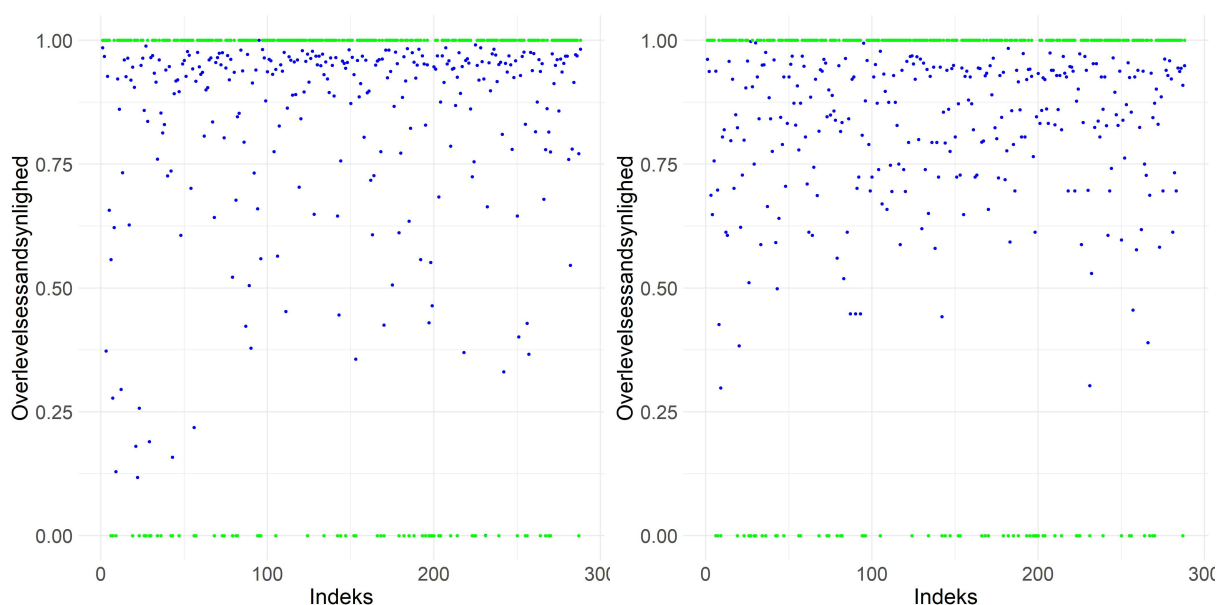


Figur 3.19: Kaplan-Meier kurver for IPS-modellen (venstre) og den modificerede IPS-model (højre) på det reducerede datasæt.

Anvendes IPS-modellen på det reducerede datasæt, så fås overlevelseskurverne i det venstre

plot i Figur 3.19. IPS-modellen har en AIC score på 2272.645. Anvendes den modificerede IPS-model fra afsnit 3.6, så fås Kaplan-Meier kurverne i højre side af Figur 3.19. Den modificerede IPS-model har en AIC score på 2263.268. Kaplan-Meier kurverne for score 3 og 4 i den modificerede IPS-model krydser ikke hinanden, hvilket de gør for IPS-modellen. Som i afsnit 3.6 er der stadig ingen signifikant forskel på at have 0 eller 1 point.

Modellen med de signifikante variable og den modificerede IPS-model testes på et udeholdt datasæt, som i forrige afsnit. På Figur 3.20 ses de prædikterede overlevelsessandsynligheder og status på patienterne. Det kan ses, at der for modellen med de signifikante variable er flere patienter der har en lavere overlevelsessandsynlighed end i den modificerede IPS-model. Middelværdien for forskellen på de prædikterede sandsynligheder og status for modellen med de signifikante variable er 0.254, og for den modificerede IPS-model er middelværdien 0.315. Det vil altså sige, at også for modellerne lavet på det reducerede datasæt, er det modellen med de signifikante variable der kommer tættest på.



Figur 3.20: De prædikterede overlevelsessandsynligheder med den modificerede IPS-model (højre) og modellen med de signifikante variable (venstre) for test-datasættet (blå), samt patienternes oplyste status, 0 =død, og 1 =censuret(grønne).

Som i forrige afsnit kan der opstilles en tabel med de gennemsnitlige prædikterede overlevelsessandsynlighederne for de 5 aldersgrupper i testdatasættet for modellen med de signifikante variable, dette kan ses i Tabel 3.6. Det ses, at der stadig er mindre sandsynlighed for overlevelse, jo ældre aldersgruppe.

Aldersgruppe	18-29 år	30-44 år	45-59 år	60-74 år	75+ år
Gennemsnitlig overlevelssandsynlighed	0.947	0.937	0.867	0.617	0.540

Tabel 3.6: *Overlevelsessandynligheden for de 5 aldersgrupper med modellen med signifikante variable.*

Der kan også opstilles en tabel med de prædikterede overlevelsessandynligheder for et halvt, et og to år for den modificerede IPS-model. Dette gøre i Tabel 3.7. Der ses at jo højere score, jo lavere sandsynlighed for overlevelse, samt at jo længere ud i fremtiden prædiktionen er, jo lavere bliver sandsynligheden.

Score	1/2 år	1 år	2 år
0	0.992	0.986	0.979
1	0.990	0.982	0.974
2	0.972	0.951	0.928
3	0.953	0.919	0.882
4	0.936	0.890	0.841
5	0.899	0.830	0.758
6-9	0.833	0.726	0.621

Tabel 3.7: *Overlevelsessandynligheden for patienter med de forskellige scorere for et halvt år, et år og to år.*

Konklusion 4

Ved brug af teori om overlevelsesanalyse, herunder Cox proportional hazard modeller, blev det muligt at prædiktere sandsynligheder for overlevelse, når en patient rammes af Hodgkin lymfekræft. Ved brug af LASSO var det muligt at lave variabel selektion, og denne model sammen med en model, som kun indeholdte de signifikante forklarende variable, gav en indikation om hvilke variable der kunne have betydning for overlevelsessandsynligheden. Det viste sig, at kategoriske variable, som ECOG score, Ann Arbor stadie og alder var vigtige i forhold til at estimere overlevelsessandsynligheden, nogle placeringer af kræften blev også fundet signifikante, eksempelvis om kræften var involveret i tonsilla palatina. Målingen af albumin og thrombocyter i blodet blev også fundet signifikant i både LASSO modellen og modellen med signifikante variable.

Med denne viden blev det muligt at modificere den nuværende IPS-model ved at tilføje fire udsagn. Disse udsagn var om ECOG scoren var 3 – 4, om de to placeringer tonsilla palatina og pleuraeksudat med lymfomceller var involveret i kræften, samt om antallet af thrombocyter/blodplader var større end 400 mia/L. Dette gav en modificeret IPS-model med en lavere AIC score end den nuværende, derfor ville denne model bedre kunne beskrive sandsynligheden for overlevelse.

Den modificerede IPS-model og modellen med signifikante variable blev begge testet på et testdatasæt, hvor det blev fundet ud af, at selvom om den modificerede IPS-model ramte lidt mere ved siden af en modellen med de signifikante variable, så havde den en stor fordel i forhold til anvendelse i praksis. Den modificerede IPS-model var altså let at anvende for læger, som skulle finde en overlevelsessandsynlighed for deres patienter, da denne kunne aflæses i en tabel eller på en graf, hvorimod der skulle beregninger til for modellen med de signifikante variable.

Et problem ved datasættet var, at der var en del manglende værdier. I afsnit 3.2 blev dette taget hånd om ved at indsætte middelværdien for de numeriske variable, og den mest repræsenterede i det kategoriske variable. Dette indførte bias i datasættet, men det gav et fuldt datasæt til at arbejde med. En anden måde at løse problemet med manglende data blev introduceret i afsnit 3.8, hvor de variable med flest manglende værdi blev fjernet. De resterende manglende værdier blev fjernet ved at fjerne de patienter, som havde manglende værdier for nogle af de resterende variable. Dette gav et fuldt datasæt, hvor der blev lavet en model kun med signifikante variable, som blev forskellig fra den lavet på datasættet med de estimerede værdier, samt en modificeret IPS-model, på samme vis som for det estimerede datasæt. Modellerne lavet i afsnit 3.8 blev også testet, og stadig havde den modificerede IPS-model en lavere AIC score end den nuværende

IPS-model.

Det vil derfor kunne konkluderes, at den modificerede IPS-model er et godt bud på en bedre udgave af den nuværende IPS-model, som stadig er let at anvende i praksis, da der stadig er et overskueligt antal parametre at tjekke for patienterne og overlevelsessandsynlighederne stadig kan aflæses på en tabel eller en graf efter at have sammenlagt point for sin patient.

- [1] Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. Wiley, 1st edition, 1991. ISBN 0-471-52218-X.
- [2] Trevor Hastie. *Package 'glmnet'*. Tilgængelig på <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>. Set den d. 25/5-2021.
- [3] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 1st edition, 2015. ISBN 978-1-4987-1216-3.
- [4] John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition, 2003. ISBN 978-1-4419-2985-3.
- [5] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2nd edition, 2002. ISBN 0-471-18386-5.
- [6] Leukæmi og MDS LyLe-Patientforeningen for Lymfekræft. *Lymfekræft - hvad er det?* Tilgængelig på <https://lyle.dk/lymfekraeft/>. Set den d. 18/5-2021.
- [7] Sundhed.dk. *B-Trombocytter (blodplader)*. Tilgængelig på <https://www.sundhed.dk/borger/patienthaandbogen/undersogelser/blod-og-urinproever/b-thrombocytter-blodplader/>. Set den d. 12/5-2021.
- [8] Wikipedia. *Ann Arbor staging*, . Tilgængelig på https://en.wikipedia.org/wiki/Ann_Arbor_staging. Set den d. 10/5-2021.
- [9] Wikipedia. *Performance status*, . Tilgængelig på https://en.wikipedia.org/wiki/Performance_status. Set den d. 10/5-2021.

A.1 Tælleprocesser og martingaler

Dette afsnit er baseret på [4, s.79-87].

Definition A.1.1 (Tælleproces)

En kontinuert stokastisk tidsproces $N = \{N(t)\}_{t \geq 0}$ er en tælleproces, hvis $N(0) = 0$, $N(t) < \infty$, og med sandsynlighed 1; er udfaldsstien af $N(t)$ højre-kontinuert og stykvis konstant med hop af størrelse +1.

Givet en højrecensureret observation, så er processerne $N_i(t) = I[T_i \leq t, \delta_i = 1]$, som er 0 indtil individ i opnår hændelsen og hopper til 1, tælleprocesser.

Tælleprocesserne giver information om hvornår hændelser sker. Udover disse informationer kan yderligere oplysninger have om individerne til tid t . For højrecensureret data så inkluderer denne information til tid t kendskab til hvem der er censureret inden tid t , og hvem der opnår hændelsen før eller til tid t . Denne information kan indeholde faste kovariater eller tidsafhængige kovariater. Informationen om hvad der er sket for individet op til tid t kaldes historien eller filtrationen af tælleprocessen til tid t og noteres $\mathbf{F}_t$. Der gælder at $\mathbf{F}_s \subset \mathbf{F}_t$ for $s \leq t$. Det vil sige, at mængden af information øges over tid. Hver filtration, $\mathbf{F}_t$, er en σ -algebra som er inkluderet i $\mathbf{F}$.

I tilfælde af højrecensureret data, så indeholder filtrationen til tid t information om parret (T_i, δ_i) forudsat at $T_i \leq t$, og at $T_i > t$ for de individer som stadig er med i studiet til tid t . Filtrationen til tiden umiddelbart før t skrives $\mathbf{F}_{t-}$.

For en given tælleproces defineres $dN(t)$ til at være forskellen i processen over et kort interval $[t, t + dt)$. Det vil sige, at $dN(t) = N[(t + dt)^-] - N(t^-)$, hvor t^- er tiden lige før t . For højrecensureret data er $dN(t) = 1$, hvis hændelsen sker til tid t , og ellers 0. Defineres processen $Y(t)$ til at være antallet af individer med studietid $T_i \geq t$:

$$\begin{aligned}\mathbb{E}[dN(t)|\mathbf{F}_{t-}] &= \mathbb{E}[\text{Antal observationer med } t \leq X_i \leq t + dt, C_i > t + dt_i | \mathbf{F}_{t-}] \\ &= Y(t)h(t)dt,\end{aligned}$$

da sandsynligheden for at have en hændelse til tid t , givet filtrationen lige før t er $h(t)dt$ for $T_i \geq t$. Den stokastisk proces $Y(t) = \sum_{i=1}^n Y_i(t)$ giver antallet af individer i risiko til tid t .

Processen $\lambda(t) = Y(t)h(t)$ kaldes *intensitetsprocessen* af en tælleproces. $\lambda(t)$ er i sig selv en stokastisk proces der afhænger af information indeholdt i filtrationen, $\mathbf{F}_t$, gennem $Y(t)$.

Den kumulative intensitetsproces er defineret som $\Lambda(t) = \int_0^t \lambda(s)ds$, for $t \geq 0$. $\Lambda(t)$ har egenskaben at $\mathbb{E}[N(t)|\mathbf{F}_{t-}] = \mathbb{E}[\Lambda(t)|\mathbf{F}_{t-}] = \Lambda(t)$. Den sidste lighed følger, da filtrationen lige før t er kendt, så værdien af $Y(t)$ er fast og dermed er $\Lambda(t)$ er ikke-stokastisk.

Den stokastiske proces $M(t) = N(t) - \Lambda(t)$ kaldes en *tælleproces martingale*. Denne proces har egenskaben, at forskellen i et kort interval af processen har en forventet værdi som er nul, givet fortiden $\mathbf{F}_{t-}$:

$$\begin{aligned}\mathbb{E}[dM(t)|\mathbf{F}_{t-}] &= \mathbb{E}[dN(t) - d\Lambda(t)|\mathbf{F}_{t-}] \\ &= \mathbb{E}[dN(t)|\mathbf{F}_{t-}] - \mathbb{E}[d\Lambda(t)|\mathbf{F}_{t-}] \\ &= Y(t)h(t)dt - Y(t)h(t)dt \\ &= 0,\end{aligned}\tag{A.1}$$

hvor tredje lighed følger, da $\lambda(t)$ har en fast værdi givet $\mathbf{F}_{t-}$.

Definition A.1.2 (Martingale)

En stokastisk proces $M = \{M(t)\}_{t \geq 0}$ er en martingale med hensyn til $\mathbf{F} = \{\mathbf{F}_t\}_{t \geq 0}$ hvis:

- $\mathbb{E}[M(t)|\mathbf{F}_s] = M(s)$ for alle $0 \leq s \leq t$.
- Hvis $M(t)$ bestemmes af $\mathbf{F}_t$, det vil sige at $M(t)$ er $\mathbf{F}_t$ -målelig, så siges M at være adapteret af $\mathbf{F}$.

For at se at denne definition er det samme som $\mathbb{E}[dM(t)|\mathbf{F}_{t-}] = 0$ for alle t . Bemærk, at hvis $\mathbb{E}[dM(t)|\mathbf{F}_{t-}] = 0$ så:

$$\begin{aligned}\mathbb{E}[M(t)|\mathbf{F}_s] - M(s) &= \mathbb{E}[M(t) - M(s)|\mathbf{F}_s] \\ &= \mathbb{E}\left[\int_s^t dM(u) \middle| \mathbf{F}_s\right] \\ &= \int_s^t \mathbb{E}[\mathbb{E}[dM(u)|\mathbf{F}_{u-}] | \mathbf{F}_s] \\ &= 0,\end{aligned}$$

dermed er tælleproces martingalen en martingale.

En tælleproces martingale, $M(t) = N(t) - \Lambda(t)$ består af to dele; $N(t)$, som er en ikke-aftagende step funktion, og $\Lambda(t)$, en glat proces, som er prædikterbar, da dens værdi til tid t er fast givet tiden før t . $\Lambda(t)$ kaldes *kompensatoren* for tælleprocessen. Martingalen kan ses som støj, som opstår når den glatte kompensator fratrækkes tælleprocessen.

En prædikterbar variationsproces af $M(t)$ noteres $\langle M \rangle(t)$. Denne egenskab er defineret som kompensatoren for processen $M^2(t)$, da $M^2(t) - \langle M \rangle(t)$ er en martingale. $M(t)$ afspejler støj-
en, som er tilbage efter at kompensatoren, $\langle M \rangle(t)$, er fratrasket, så har $M^2(t)$ tendens til at
stige med tiden. $\langle M \rangle(t)$ er den systematiske del af denne stigning.

Proposition A.1.3 (Prædikterbar variationsproces)

For en martingale $M(t)$, så er

$$\text{Var}[dM(t)|\mathbf{F}_{t-}] = d\langle M \rangle(t).$$

Bevis

Variansen udregnes til at være:

$$\begin{aligned} \text{Var}[dM(t)|\mathbf{F}_{t-}] &= \mathbb{E}[(dM(t))^2|\mathbf{F}_{t-}] - (\mathbb{E}[dM(t)|\mathbf{F}_{t-}])^2 \\ &= \mathbb{E}[(M(t+dt)^-)^2 + M(t^-)^2 - 2M(t+dt)^-M(t^-)|\mathbf{F}_{t-}] - 0 \\ &= \mathbb{E}[(M(t+dt)^-)^2 - M(t^-)^2|\mathbf{F}_{t-}] \\ &= \mathbb{E}[dM^2(t)|\mathbf{F}_{t-}] = d\langle M \rangle(t). \end{aligned} \tag{A.2}$$

■

Lad $K(t)$ være en prædikterbar proces. Det vil sige at $K(t)$ er en stokastisk proces, hvis værdier er kendte givet filtrationen $\mathbf{F}_{t-}$. Et eksempel på sådan en proces er $Y(t)$. På intervallet 0 til t er det stokastiske integral af en sådan proces, med hensyn til en martingale, givet ved $\int_0^t K(u)dM(u)$. Et stokastisk integral har den egenskab, at den selv er en martingale, som funktion af t , og dens prædikterbare variationsproces kan findes fra den prædikterbare variationsproces af den oprindelige martingale:

$$\left\langle \int_0^t K(u)dM(u) \right\rangle = \int_0^t K(u)^2 d\langle M \rangle(u). \tag{A.3}$$

Den central grænseværdi sætningen for martingales siger, at en sekvens af martingaler $M_n = \{M_n(t)\}_{t \geq 0}$ konvergerer mod en Gaussisk proces. Sætningen er som følger:

Sætning A.1.4 (Central grænseværdisætning for martingaler)

Lad M_n være en martingale som opfylder:

-

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} [(M_k - M_{k-1})^2 | \mathbf{F}_{t-}] \rightarrow 1 \text{ når } n \rightarrow \infty,$$

- for alle $\varepsilon > 0$,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} [(M_k - M_{k-1})^2; |M_k - M_{k-1}| > \varepsilon \sqrt{n}] \rightarrow 0 \text{ når } n \rightarrow \infty,$$

så konvergerer $\frac{M_n}{\sqrt{n}}$ i fordeling mod $N(0, 1)$ når $n \rightarrow \infty$.

Sætningen vises ikke.

Dette afsnit er baseret på [3, s.8-9, 43].

LASSO er en metode som kombinerer mindste-kvadraters tab med en begrænsning på den absolute værdi af summen af koefficienterne i modellen. Givet en samling af N prädiktor-respons par, $\{(x_i, y_i)\}_{i=1}^N$ så finder LASSO løsningen $(\hat{\beta}_0, \hat{\beta})$ findes til optimeringsproblemet:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2N} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^m x_{ij} \beta_j)^2 \right\},$$

underlagt begrænsningen $\sum_{j=1}^m |\beta_j| \leq t$. Denne begrænsning skrives ofte som $\|\beta\|_1 \leq t$. Dette LASSO problem kan omskrives på Lagragian form som:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

for et $\lambda \geq 0$, og $\|\cdot\|_2$ er den Euklidiske norm for vektorer. λ er en tunings parameter, som kontrollere styrken af straffen. Hvis λ er tilstrækkelig stor, så tvinger den koefficienterne for nogle forklarende variable til at være præcis nul, hvilket reducerer dimensionen. Jo større λ , jo færre forklarende variable i modellen. Dette er en af fordelene ved LASSO, og grunden til at LASSO bruges til variabel selektion.

For Cox modeller haves data på formen (t_i, δ_i, z_i) , hvor t_i er hændelsestidspunktet, δ_i er indikator for hændelse, og z_i er kovariat. For Cox modeller er LASSO løsningen givet ved:

$$\min_{\beta} \left\{ - \sum_{\{i|\delta_i=1\}} \log \left(\frac{\exp(\beta^T z_i)}{\sum_{j \in R_i} \exp(\beta^T z_j)} \right) + \lambda \|\beta\|_1 \right\},$$

hvor det kan ses at første led er logaritmen til den partielle likelihood, hvilket svarer til den betingede sandsynlighed i risikomængden af de observerede hændelser.