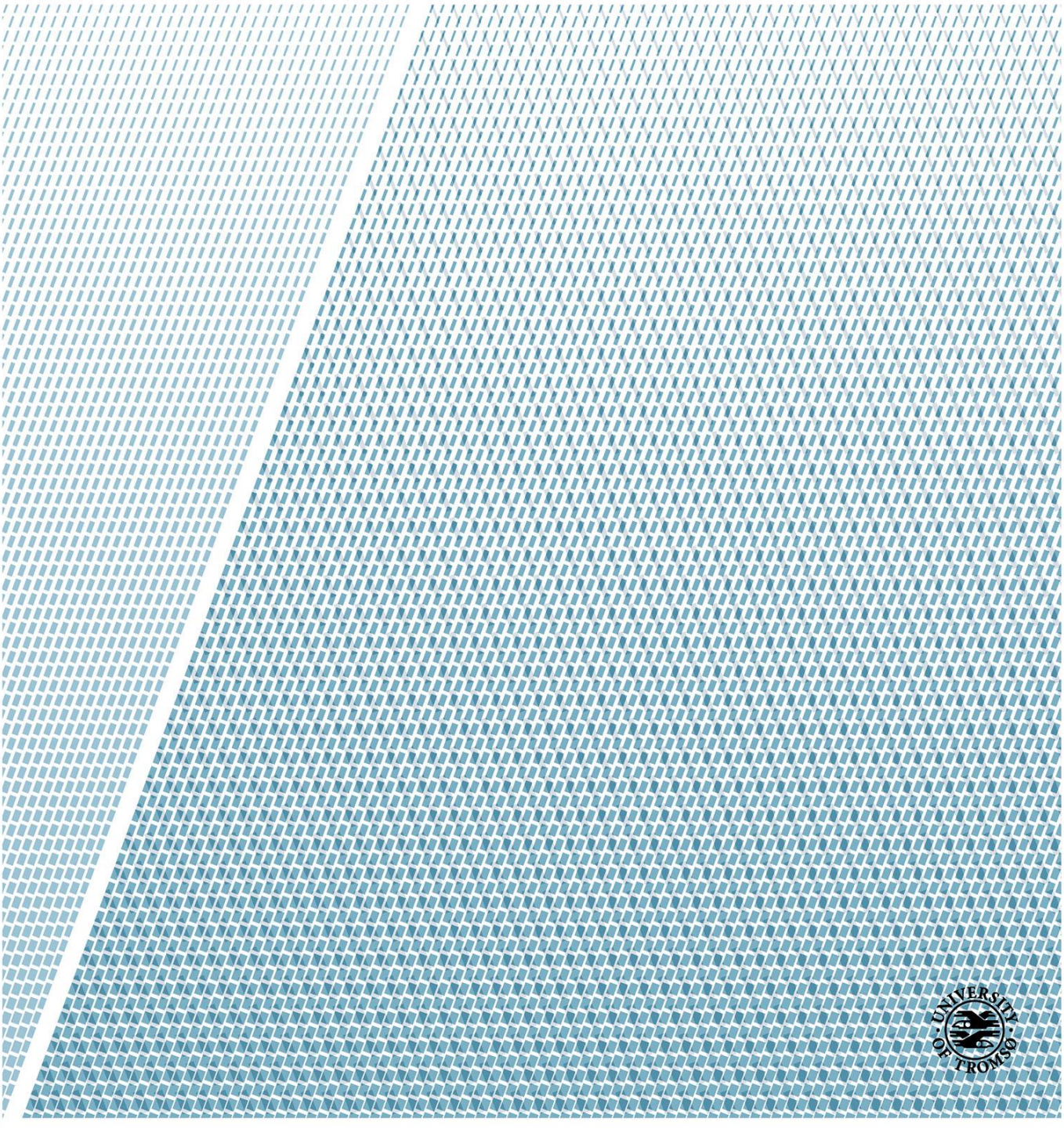
UiT
THE ARCTIC UNIVERSITY OF NORWAY

Institutt for Matematikk og Statistikk.

Overlevelsesanalyse med tidsavhengige forklaringsvariabler med bruk av Cox proporsjonal hasard regresjonsmodell.

Irmelin Kr. Nilsen.

Mastergradsoppgave i Industriell Matematikk, 30stp, STA-3921 Desember 2015.



Forord.

Høsten 2014 tok jeg et 10 studiepoengs fag kalt Individual Special Curriculum, STA-3810, ved universitetet i Tromsø, UiT. Overlevelsesanalyse var en del av dette faget og jeg syntes dette var noe som virket veldig interessant. Etter forespørsel fra amanuensis Edvin Bredrup, ved UiT,om hva jeg ønsket at min mastergradsoppgave skulle omhandle nevnte jeg at jeg syntes at overlevelsesanalyse kunne virke som et interessant tema for min mastergradsoppgave. Edvin satte med i kontakt med førtseamanuensis Elinor Ytterstad, ved insitutte for matematikk og statistikk ved UiT, som var ansvarlig for overlevelsesanlysedelen i faget STA-3810. Elinor ble dermed min veileder både for prosjektoppgaven min og for mastergradsoppgaven min, som begge omhandlet overlevelsesanalyse.
Når det kommer til selve mastergardsoppgaven og arbeidet med den vil jeg først få takke Edvin Bredrup som hjalp med å komme i kontakt med Elinor Ytterstad. Jeg ønsker å takke veilederen min Elinor Ytterstad for veiledning og hjelp. Jeg ønsker også å takke forskningsprosjektet "Kvinner og kreft" (The Norwegian Women and Cancer post genome cohort study - NOWAC) ved Eiliv Lund og Tonje Braaten, Institutt for samfunnsmedisin, UiT Norges arktiske universitet, for utlån av datamateriale som denne mastergradsoppgaven er bygget på. Jeg ønsker å takke familie og venner, for all støtte igjennom denne perioden. Spesielt vil jeg takke mine foreldre, Kati og Svein, som har lært meg en viktig ting her i livet og det er at ingen ting er umulig, men mat man må jobbe for å oppnå det. Jeg vil takke verden beste tante, Junni, som har hjulpet meg utrolig mye, alle burde hadde ei tante som deg. Og jeg ønsker å takke kjæresten min Anders som lyser opp en, noe gang, grå hverdag.

Sammendrag.

I denne mastergradsoppgaven ble det brukt overlevelsesanalyse med tidsavhengige variabler til å estimere en Cox proporsjonal hasard regresjonsmodell. Variablene stammet fra forskningsprosjektet "Kvinner og kreft" (The Norwegian Women and Cancer post genome cohort study - NOWAC) ved universitetet i Tromsø og besto av 57 561 kvinner. Modellen ble brukes til lage en modell av hasard raten til kvinnene og å finne hvilken av forklaringsvariabelen som hadde en effekt på overlevelsesevnen til kvinner.

Innhold

1 Innledning.

Overlevelsesanalyse kan dateres helt tilbake til 1600-tallet da den første livs tabellen ble publisert av den Engelske statistikeren John Graunt, (Liu 2012). Til å begynne med ble overlevelsesanalyse brukt til å analysere dødelighet, men i det senere år har bruken av overlevelsesanalyse økt kraftig noe som kan skyldes datamaskinenes utvikling. Etterhvert som bruken av overlevelsesanalyse økte, økte også utviklingen av nye statistiske metoder. I 1972 ble Cox proporsjonal hasard regresjonsmodell publisert, siden den gang har denne modellen blitt den mest brukt regresjonsmodellen i overlevelsesanalyse. Cox regresjonsmodellen har også hatt stor betydning for utviklingen av andre statistiske metoder som brukes i overlevelsesanalyse.
I denne mastergradsoppgaven skal det brukes overlevelsesanalyse med tidsavhengige variabler til å se på overlevelsesevnen til kvinner. Det skal lages en Cox proporsjonal hasard regresjonsmodell med tidsavhengige variabler som kan brukes til å estimere hasard raten for hver kvinne og se hvilke forklaringsvariabler som har en effekt på overlevelsesevnen til kvinnene.
Datamateriale som blir brukt i denne oppgaven inneholder informasjon om 57 561 kvinner. Informasjonen er hentet inn ved at disse kvinnene i løpet av en tidsperiode på 13 år har svart på 3 spørreundersøkelser som har tatt for seg alt fra kosthold, trening, utdannelse, bosted og prevensjon. Svarene fra disse spørreundersøkelsen er så lagt inn i et datasett. Spørreundersøkelsen ble besvart av kvinner fra 34 år og oppover.
I denne oppgaven skal det kun sees på om alkoholinntak, røykevaner, alder, utdannelse, antall fødsler, egen helse, fysiskaktivitet og BMI har noen å si for dødeligheten til kvinnene. Årsaken til dødsfallene er ikke tatt med i denne oppgaven.
Til å behandle og analysere dette datamaterialet vil det statistiske dataprogrammet R bli brukt sammen med dens innebygde Cox proporsjonale hasard regresjonsmodell som kan behandle tidsavhengige variabler.
Før analysen og resultatene av analysene gjennomgås vil det først bli gitt en forklaring av hva overlevelsesanalyse er, ulike funksjoner som kan brukes i overlevelsesanalyse, en forklaring av datamaterialet og hvordan feil og mangler er blitt håndtert.

2 Teori

2.1 Hva er overlevelsesanalyse.

Overlevelsesanalyse er et fagfelt innenfor statistikk som studerer tiden, T, inntil en hendelse inntreffer, hvor T er en tilfeldig variabel og $T \geq 0$. I overlevelsesanalyse studerer man ofte tiden inntil et individ dør, tiden inntil at et individ blir syk eller tiden inntil individet får en organtransplantasjon etc. Man kan også studer tiden inntil et jordskjelv inntreffer, elektrisk komponent bryter sammen, fødsel inntreffer eller at aksjekursen avtar. Når slike hendelser inntreffer kalles ofte deres analyse noe annet enn overlevelsesanalyse, for eksempel er pålitelighetsanalyse mye brukt i ingeniørfag og analyserer tiden inntil en elektrisk komponent bryter sammen, mens varighetsanalyse ofte brukes i økonomi til å analysere tiden inntil en aksjekurs begynner å avta.

2.2 Behandling av manglende observasjoner.

I enkelte tilfeller vet man ikke det nøyaktige tidspunktet når hendelsen inntraff, eller kanskje noen individer fortsatt er i live etter at studiet ble avsluttet og dermed er deres overlevelsestid ukjent, i slike tilfeller blir individene sensurert. Sensurering deles ofte inn i ulike typer.
Venstre sensurering inntreffer når hendelsen allerede har inntruffet hos individet før studiet har startet.
Høyre sensurering er når hendelsen ikke har inntruffet hos individet i løpet av forsøket.
Intervall sensurering er når man kun vet at hendelsen har inntruffet hos individet i løpet av et tidsintervall og ikke ved et nøyaktig tidspunkt.
Type I sensurering inntreffer når studiet blir avsluttet ved et bestemt tidspunkt og de gjenværende individene, hvor hendelsen ikke ennå har inntruffet, vil da blir høyre sensurert.
Type II sensurering inntreffer når man velger å stoppe studiet etter at hendelsen har inntruffet hos et bestemt antall individer, og de gjenværende individene vil da blir høyre sensurert.

2.3 Parametrisk, ikke-parametrisk og semi-parametrisk overlevelsesanalyse.

I overlevelsesanalyse skiller man ofte mellom parametrisk, ikke-parametrisk og semi-parametrisk overlevelsesanalyse.
I parametrisk overlevelsesanalyse antar man at dataene kommer fra en kjent sannsynlighetsfordeling med kjente parametere. Hvis antagelsene om sannsynlighetfordelingen er riktig vil parametrisk overlevelsesanalyse gi mer presise og nøyaktige estimater enn ikke-parametrisk overlevelsesanalyse.
Ikke-parametrisk overlevelsesanalyse er statistikk som ikke baserer seg på kjente sannsynlighetsfordelingen.
Semi-parametrisk overlevelsesanalyse er statistikk som blander parametrisk og ikke-parametrisk overlevelsesanalyse, det vil si at en semi-parametrisk modell består både av parametriske komponenter og ikke-parametriske komponenter.

2.3.1 Parametrisk overlevlesesanalyse.

Parametrisk overlevelsesfunskjon.

Overlevelsesfunksjonen, $S(t)$, sier noe om hvor stor sannsynlighetene er for at et individ skal overleve utover en gitt tid, t, og er definert som

$$S(t) = P(T > t) \tag{1}$$
$$= 1 - P(T \leq t) \tag{2}$$

$$S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(x) \qquad T \text{ er kontinuerlig} \tag{3}$$
$$S(t) = P(T > t) = \sum_{t_j > t} p(t_j) \qquad T \text{ er diskret} \tag{4}$$

Hvor $f(x)$ er den ikke negative tetthetsfunksjon i det kontinuerlige tilfellet, $f(x) \geq 0$, og $\int_{-\infty}^{\infty} f(x)dx = 1$, og $p(t_j)$ er tetthetsfunksjon i det diskrete tilfellet, $p(t_j) \geq 0$ og $\sum_{t_j \in A} p(t_j) = 1$.

Noen av egenskapene til overlevelsesfunksjonen er at den er monoton avtagende, det vil si $S(t_2) \leq S(t_1)$ for $t_2 > t_1$. Ved starten av et studie er alle individene i livet, det betyr at ved tiden $t = 0$ så er sannsynligheten for å overleve lik 1. Etter hvert som tiden T går mot uendelig vil sannsynligheten for å overleve avta og til slutt bli 0, dermed er $S(0) = 1$ og $S(\infty) = 0$.

Parametrisk hasard rate.

Hasard raten, $h(t)$, sier noe om risikoen for at et individ, med en alder t, vil oppleve hendelsen i løpet av kort tid, kalles ofte for risikofunksjonen.
Selve definisjonen av hasard raten er den betingede sannsynligheten for at hendelsen vil inntreffe i løpet av tidsintervallet $[t, t + \Delta t)$ gitt at den ikke har inntruffet allerede, dividert med lengden av intervallet.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{5}$$

Ved bruk av overlevelsesfunksjonen, $S(t)$, og sannsynlighetstetthetsfunksjonen, $f(t)$ i det kontinuerlige tilfellet og $p(t_j)$ i det diskrete tilfellet, kan hasard raten defineres som

$$h(t) = \frac{-d(ln[S(t)])}{dt} = \frac{f(t)}{S(t)} \qquad T \text{ er kontinuerlig} \tag{6}$$

$$h(t_j) = P(T = t_j | T \leq t_j) = \frac{p(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})} \qquad T \text{ er disrkret} \tag{7}$$

Hasard rate funksjonen er ikke en sannsynlighet, noe som betyr at den kan ha verdier som er større enn 1, men desto større hasard rate desto større risiko er det for at hendelsen skal inntreffe. Hasard raten er også ikke-negativ, slik at at $h(t) \geq 0$.
I motsetning til overlevelsesfunksjonen, som ser på sannsynligheten for at en hendelse ikke skal inntreffe, ser hasard raten på risikoen for at en hendelse skal inntreffe.

Formen på hasard rate funksjonen vil ha forskjellig form utifra hva den modellerer. Hasard rater som modellerer naturlig aldring eller slitasje vil ha en økende hasard rate. Dette skyldes at risikoen for at en hendelse skal inntreffe øker med alderen eller etter hvert som en gjenstand brukes. Hasard rater som er avtagende er ofte typisk for elektroniske apparater med defekte deler og hos organtransplantasjon hos mennesker, hvor risikoen for komplikasjoner er størst rett etter transplantasjonen. Denne type hasard rate en noe uvanlig.
En badekar lignende hasard rate kan man ofte se i studier hvor man følger individer fra fødselen til død. Dette kommer av at risikoen for å dø av naturlige årsaker er størst i begynnelsen og ved slutten av livet, men stabiliserer seg i midten av livet.
En hasard rater som har en konkav form brukes ofte til å modellere hasard raten hos pasienter som har hatt en vellykket operasjon, men hvor risikoen for en infeksjon eller andre komplikasjoner er størst rett etter operasjonen for så å avta.

Kumulativ hasard raten.

Den kumulative hasard raten sier noe om hvor mange ganger man vil forvente at en hendelse inntreffer i løpet av en tidsperiode.
Den kumulative hasard raten er definert som

$$H(t) = \int_0^t h(u)du = -ln[S(t)] \qquad T \text{ er kontinuerlig} \tag{8}$$

$$H(t) = \sum_{t_j \leq t} h(t_j) \qquad T \text{ er diskret} \tag{9}$$

Mean residual life function.

Mean residual life funksjonen er en funksjon for den forventede gjenværende levetiden til et individ med alder t, gitt at hendelsen ikke har inntruffet hos dette individet fra før.
Mean residual life function defineres som

$$mrl(t) = E(T - t|T > t) \tag{10}$$

$$mrl(t) = \frac{\int_t^\infty (x-t)f(x)dx}{S(t)} = \frac{\int_t^\infty S(x)dx}{S(t)} \qquad \text{T er kontinuerlig} \tag{11}$$

$$mrl(t) = \frac{(t_{i+1} - x_i)S(t_i) + \sum_{j \geq i+1}(t_{j+1} - t_j)S(t_j)}{S(t)} \qquad \text{T er diskret} \tag{12}$$

Sannsynlighetsfordelinger.

Ikke alle sannsynlighetsfordelinger beskriver like godt sannsynligheten for å overleve, men noen av fordelingene som ofte brukes i overlevelsesanalyse er eksponentialfordelingen, Weibullfordelingen, gammafordelingen, lognormalfordelingen, normalfordelingen og Gopmertzfordelingen

Weibull fordelingen.
En av de sannsynlighetsfordelingene som er mye brukt i overlevelsesanalyse er Weibull fordelingen. Weibullfordelingen ble for første gang introdusert av den svenske fysikeren Waloddi Weibull i 1939. Eksponentialfordelingen er et spesialtilfelle av Weibullfordelingen, men Weibullfordelingen har ikke den egenskapen at den er minneløs.

Weibullfordelingen har sannsynlighetstetthetsfunksjonen

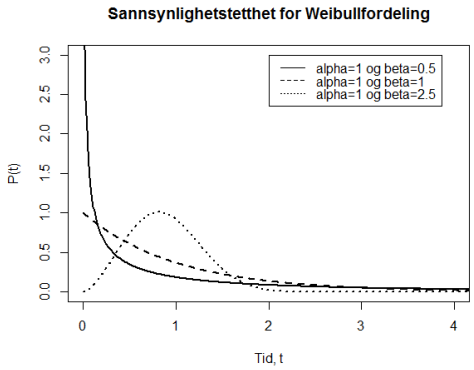
$$f(t) = \alpha\beta t^{\beta-1}e^{-\alpha t^\beta} \qquad 0 < \alpha, \beta \quad 0 \leq t \tag{13}$$

For sannsynlighetstettheten vil grafen til Weibullfordelingen endres etter som β endres. Hvis $\beta = 1$ vil grafen være like grafen til en eksponentialfordeling med samme verdi for α.
Overlevelsesfunksjonen og hasard raten er henholdsvis

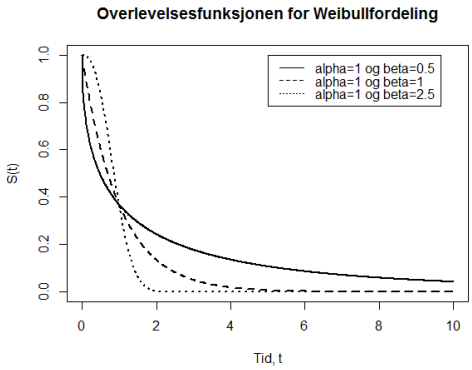
$$S(t) = e^{-\alpha t^\beta} \tag{14}$$

$$h(t) = \alpha\beta t^{\beta-1} \tag{15}$$

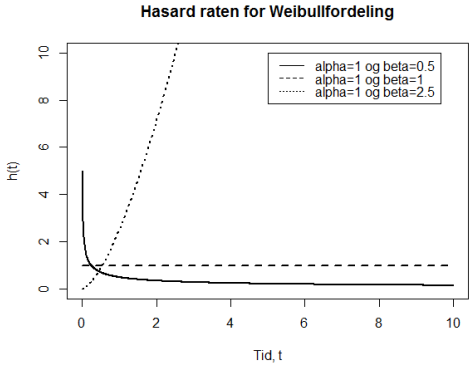
Noen av egenskapene til hasard rate funksjonen til Weibullfordelingen er følgende
1. Hvis $\beta = 1$ vil man få en konstant hasard rate.
2. Hvis $\beta > 1$ så er hasard raten en økende funksjon av tiden t, noe som indikerer at, for eksempel en komponent, slites med tiden.
3. Hvis $\beta < 1$ så vil hasard raten avta med tiden og en komponent vil dermed bli sterkere med tiden.



(a) Sannsynlighetstettheten for en Weibullfordeling med ulike verdier for α og β.



(b) Overlevelsesfunksjonen for en Weibullfordeling med ulike verdier for α og β.



(c) Hasard rate funksjonen for en Weibullfordeling med ulike verdier for α og β.

Figur 1: Sannsynlighetstettheten, overlevelsesfunksjonen og hasard rate funksjonen for en Weibullfordeling med ulike verdier for α og β.

2.4 Ikke- parametrisk overlevelsesanalyse.

Kaplan-Meier estimator.

Kaplan-Meier estimatoren, også kalt product limit estimator, er en ikke parametrisk estimator for overlevelsesfunksjonen, betegnes $\hat{S}(t)$.

Gitt at der totalt er D antall hendelser som inntreffer i løpet av bestemt tid slik at $t_1 < t_2 < ... < t_D$ og $i = 1, 2, ..., D$, da er Kaplan-Meier estimatoren gitt ved

$$\hat{S}(t) = \begin{cases} 1 & \text{hvis t}\leq t_1 \\ \prod_{t_1 \leq t}[1 - \frac{d_i}{y_i}] & \text{hvis } t_1 \leq t. \end{cases}$$

Hvor d_i er antallet hvor hendelsen har inntruffet hos, for eksempel antall døde ved tidspunktet t_i, og y_i er antallet hvor hendelsen enda ikke har inntruffet.
Kaplan-Meier estimatoren tar også hensyn til at observasjonene kan være sensurerte, da vil y_i være antallet hvor hendelsen enda ikke har inntruffet minus antallet som har blitt sensurert.

Siden Kapalan-Meier estimatoren er en trappefunksjon vil plotting av estimatoren føre til et trappetrinn lignende utseende av dens graf. Hvert trappetrinn representerer en eller flere individer hvor hendelsen har inntruffet ved tidspunktet t, desto større trappetrinn desto flere hendelser ar inntruffet. Sensurering av individer representeres som et kors på grafen. Det er et kors for hvert individ ved tidspunktet t.

Nelson-Aalen estimator.

Nelson-Aalen estimatoren brukes ofte for å estimere den kumulative hasard raten, $\tilde{H}$.

Primært sett har denne estimatoren to bruksområder:
1. Den kan brukes til å tilpasse parametriske modeller til dataene. Plotter man Nelson-Aalen estimatoren på et spesielt type papir sammen med en parametrisk modell vil grafen til dataene være tilnærmet lineære hvis denne parametriske modellene passer til dataene.
2. Den kan også brukes til å gi et grovt estimat av hasard raten $h(t)$, som vises som skråningen av grafen i Nelson-Aalen plottet.

Neslon-Aalen estimatoren er gitt ved

$$\tilde{H}(t) = \begin{cases} 0 & \text{hvis t}\leq t_1 \\ \sum_{t_1 \leq t} \frac{d_i}{y_i} & \text{hvis } t_1 \leq t \end{cases}$$

Hvor d_i er antallet hvor hendelsen har inntruffet ved tiden t_i og y_i er antallet hvor hendelsen enda ikke har inntruffet.

Plottet av Nelson-Aalen estimatoren representere den kumulative hasard raten, mens skråningen på grafen representere hasard raten, det betyr at desto brattere graf desto større er risikoen for at en hendelse skal inntreffe.

Nelson-Aalen estimatoren er ofte en bedre estimator for små prøve utvalg enn Kaplan-Meier estimatoren.

2.5 Semi-parametrisk overlevelsesanalyse.

2.5.1 Modellere hasard rate funksjonen.

Ofte ønsker man å finne hvordan overlevelse påvirkes av en eller flere forklaringsvariabler og hvor stor risikoen er for at den bestemte hendelsen skal inntreffe. En måte man kan gjøre dette på er å modellere hasard rate funksjonen ved hjelp av Cox proporsjonal hasard regresjonsmodellen.
Cox modellen er semi-parametrisk modell fordi den trenger ingen antagelse om formen til fordelingen av overlevelsestiden, men den trenger en antagelse om hasard raten.

2.5.2 Cox proporsjonal hasard modell.

Man kan bruke Cox proporsjonal hasard regresjonsmodell til å modellere risikoen for at en hendelse skal inntreffe hos et individ, i tillegg kan man sammenligne risikoen for at hendelse skal inntreffer for ulike grupper av individer, kalles relativ risiko eller hasard ratio, for eksempel kan man finnes risikoen for at hendelsen skal inntreffe hos menn i forhold til kvinner, eller for en gruppe som er blitt behandlet med en type medisin sammenlignet med en placebo gruppe. Fordi alle individene ikke er like, forskjellig alder, individet har hatt kreft eller at de har ulik BMI, som kan påvirke om hendelsen inntreffer eller ikke, utstyrer man hvert individ med et sett, eller en vektor, bestående av p antall forklaringsvariabler, kalles også risiko faktorer eller kovariate variabler, betegnes som $\mathbf{X} = [x_1, x_2, ..., x_p]^T$, som forklarer disse karakteristiske forskjellen mellom individene. Disse variablene kan være uavhengige av hverandre, men en eller flere variabler kan også være avhengige, interaksjoner, mer om dette senere. I overlevelsesanalyse ser man også på om noen av disse forklaringsvariablene har en innvirkning på overlevelsesevnen til individene, kanskje vil det å røyke gjøre at individet lever kortere enn et individ som aldri har røkt.

Forklaringsvariabler kan deles inn i ulike typer variabler. For eksempel kan en eller flere forklaringsvariabelen være kodet som dummy variabler. En dummy variable tar vanligvis to verdier, ofte er disse verdiene 0 og 1. For eksempel hvis forklaringsvariabelen x_1 står for hvilket kjønn individet er så kan den være 0 for mann og 1 for kvinne.
Forklaringsvariabelen kan også være en kategorisk variable det vil si at variabelen tar kun et begrenset antall verdier. For eksempel hvis x_2 var en variabel som rangerte hvor lykkelig individet var på en skala fra 1 til 10. Så ville denne variabelen være en kategorisk variable bestående av kategoriene $1, 2, 3, ..., 10$, eller hvis x_3 var en variabel for sivilstatusen til individet kunne den bestå av kategorien singel, gift, samboer, enkemann og enke.
En eller flere forklaringsvariabelen kan også være en kontinuerlig variabel, det vil si at variabelen kan være hvilken som helst verdi, for eksempel hvis x_4 var variabelen for høyden til individet kan den ta hvilken som helst verdi, 164,6 cm, 153 cm, 190 cm o.s.v.

Forklaringsvariablene kan tidsuavhengig, det vil si at verdien til variabelen ikke forandrer seg med tiden, men er konstant, da bruker man en Cox proporsjonal hasard regresjonsmodell som er tidsuavhengig. Hvis verdien på forklaringsvariablene forandrer seg med tiden bruker man en Cox proporsjonal hasard regresjonsmodell som er avhengige av tiden, denne modellen kalles ofte utvidet Cox regresjonsmodell. Cox proporsjonal hasard regresjonsmodellen kan også bestå av både tidsuavhengige forklaringsvariabler og tidsavhengige forklaringsvariabler.

2.5.3 Tidsuavhengig Cox regresjonsmodell.

Cox proporsjonal hasard regresjonsmodell ble for første gang introdusert av Sir David Roxbee Cox i 1972. En tidsuavhengig Cox proporsjonal hasard regresjonsmodell uttrykker hasard raten ved tiden t for et individ med et gitt sett av forklaringsvariabler som er uavhengig av tiden t. Hvis man har n antall individer hvor, $j = 1, 2, ..., n$ så er Cox regresjonsmodellen for hasard raten til individ nummer j definert som

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{k=1}^{p} \beta_k x_k} \tag{16}$$

$h(t|\mathbf{X})$ er hasard raten ved tiden t gitt forklaringsvariablene til individ j $\mathbf{X} = [x_1, x_2, ..., x_p]^T$ og, $h_0(t)$ er en funksjons av tiden t og kalles grunnlinje hasard funksjon, baseline function på engelsk. $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_p]^T$ er parametervektoren bestående av regresjonskoeffisienter. Desto større verdien for regresjonskoeffisienten til forklaringsvariabelen er desto større blir hasard raten og dermed øker risikoen for at hendelsen inntreffer.

Noen av egenskapene til grunnlinje funksjonen er at $h_0(t)$ kan ikke være negativ, $h_0(t) \geq 0$, og når $h(t, \mathbf{X}) = h_0(t)e^0 = h_0(t)$ representerer dette start funksjonen til hasard raten. $h_0(t)$ er en uspesifisert funksjon.

Den relative risikoen til en gruppe av individer, hvor alle har en vektor bestående av samme forklaringsvariabler, $\mathbf{X}^*$, sammenlignet med de individene som ikke er en del av denne gruppen, men hvor alle individene har en vektor bestående av de samme forklaringsvariablene, $\mathbf{X}$, blir da

$$Relativ\ \ risiko = \frac{h(t|\mathbf{X}^*)}{h(t|\mathbf{X})} = \frac{h_0(t)e^{\sum_{k=1}^{p} \beta_k x_k^*}}{h_0(t)e^{\sum_{k=1}^{p} \beta_k x_k}} \tag{17}$$

$$= e^{\sum_{k=1}^{p1} \beta_k (x_k^* - x_k)} \tag{18}$$

Hvis den relative risikoen er 1.5 betyr det at risikoen for at hendelsen skal inntreffe er 1.5 ganger større hos gruppen av individer med vektor av forklaringsvariabler $\mathbf{X}^*$ sammenlignet med de individene med vektor av forklaringsvariabler $\mathbf{X}$.

Estimering av koeffisientene $\boldsymbol{\beta}$.

Fordi Cox regresjonsmodellen kun ser på sannsynligheten til de individene hvor hendelsen faktisk inntreffer, ikke de som sensureres, må man bruke partiell likelihood for å kunne estimere koeffisientene β. Estimeringen gjøres på følgende måte.
Gitt at man har et datasett bestående n antall individer. j representerer her individ nummeret det vil si $j = 1, 2, ..., n$. Hvert individ j har en vektor bestående av p antall forklaringsvariabler, $\mathbf{X}_j = [x_{j1}, x_{j2}, ..., x_{jp}]^T$. Hver vektor av forklaringsvariabler, $\mathbf{X}_j$ har uavhengig hendelsetid med sensureringstid.
I løpet av studiet inntreffer D antall hendelser, disse hendelsene inntreffer ikke på samme tidspunkt og de er ordnet slik at $t_1 < t_2 < ... < t_D$ for $i = 1, 2, ..., D$. For et individ hvor hendelsen

inntreffer ved tiden t_i og som har en vektor av forklaringsvariabler bestående av p antall forklaringsvariabler hvor $k = 1, 2, ..., p$, kan man da finne koeffisientene ved bruk av partiell likelihooden som baserer seg på hasard raten og som blir

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{\sum_{k=1}^{p} \beta_k x_{(i)k}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k x_{jk}}} \tag{19}$$

$$log(L(\beta)) = \sum_{i=1}^{D} \left[\sum_{k=1}^{p} \beta_k x_{(i)k} - log\left(\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k x_k} \right) \right] \tag{20}$$

$R(t_i)$ er risikosettet ved tiden t_i og består av alle individer hvor hendelsen ikke har inntruffet ved tidspunktet like før t_i.
I partiell likelihood uttrykket avhenger telleren kun av de individene hvor hendelsen har inntruffet, mens nevneren avhenger av de individene hvor hendelsen enda ikke har inntruffet.

Videre bruker man så maximum likelihood metoden på log partiell likelihooden, $log(L(\beta))$, og løser med hensyn på β.

$$\hat{\beta}_k = \frac{\partial log L(\beta)}{\partial \beta_k} = 0 \qquad \text{for } k = 1, 2, ..., p \tag{21}$$

De estimerte koeffisienten, $\hat{\beta}_k$, er asymptotisk normal fordelt, asymptotisk forventningsrett og har forventning lik β_k.
Utregningen av koeffisientene kan være kompliserte men man kan bruke statistiske dataprogrammer, som for eksempel R og SAS, til å regne ut disse.

Metoder for å behandle flere hendelser som inntreffer på samme tidspunkt.

Når flere hendelser inntreffer på samme tidspunkt er det vanskelig å rangere de etter hvilken som inntraff først, men det finnes ulike metoder som kan brukes for å beregne partiell likelihooden for hendelser som inntreffer på samme tidspunkt. Breslow metoden, Efron metoden og diskret tid metoden er noen måter å behandle hendelser som inntreffer på samme tidspunkt.
Felles for alle metodene er at gitt at man har $t_1 < t_2 < ... < t_D$ som er de D antall ordnede tidene for når hendelsen inntreffer, $i = 1, 2, ..., D$. Så lar man d_i være antall hendelser som inntreffer på samme tidspunkt t_i. $\mathbb{D}_i$ er alle de individene hvor hendelsen har inntruffet ved tiden t_i og $\mathbf{s}_i$ er summen av alle vektorene bestående av forklaringsvariabler, $\mathbf{X}_j$ for de individene hvor hendelsen har inntruffet ved tiden t_i, det vil si $\mathbf{s}_i = \sum_{j \in \mathbb{D}_i} \mathbf{X}_j$. Vi innfører også et risikosett, R_i som består av alle de individene hvor hendelsen enda ikke har inntruffet, men som har en risiko for at den kan inntreffe like før tiden t_i.

Breslow metoden.
I Breslow metoden ser man på alle de d_i antall hendelsen som inntreffer på tidspunktet t_i som om de var forskjellige. Man regner man ut hvert av disse hendelsen likelihood og multipliserer så disse likelihoodene sammen ved tiden t_i. Den approksimerte partiell likelihood funksjonen blir dermed

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \mathbf{s}_i}}{\left[\sum_{j \in R_i} e^{\boldsymbol{\beta}^T \mathbf{X}_j} \right]^{d_i}} \tag{22}$$

Her er $\mathbf{X}_j$ vektoren bestående av p antall forklaringsvariabler for individ j, det vil si $\mathbf{X}_j = [x_{j1}, x_{j2}, ..., x_{jp}]^T$. $\boldsymbol{\beta}^T$ er parametervektoren bestående av p antall koeffisienter, altså $\boldsymbol{\beta}^T = [\beta_1, \beta_2, ..., \beta_p]^T$. Og $\boldsymbol{\beta}^T \mathbf{X}_j = \sum_{k=1}^{p} \beta_k x_{jk}$.

Denne metoden er lettere og bergene enn noen av de andre metodene, men den fungerer best på få hendelser som inntreffer på samme tidspunkt.

Efron metoden.
Den approksimerte partiell likelihood funksjonen for Efron metoden er

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \mathbf{s}_i}}{\prod_{j=1}^{d_i} \left[\sum_{k \in R_i} e^{\boldsymbol{\beta}^T \mathbf{X}_k} - \frac{j-1}{d_i} \sum_{k \in \mathbb{D}_i} e^{\boldsymbol{\beta}^T \mathbf{X}_k} \right]} \tag{23}$$

Efron metoden er også ganske enkel å bergene. Når de er få hendelser som inntreffer på samme tidspunkter gir Efron metoden og Berslow metoden nesten samme partiell likelihood, men når det er flere hendelser som inntreffer på samme tidspunkt er Efron metoden bedre en Breslow metoden, men denne metoden er ikke å foretrekke for ganske mange hendelser som inntreffer på samme tidspunkt.

Diskret metode.
Diskret metode baserer seg på at tiden som hendelsene inntreffer samtidig på er diskret slik at man har en diskret tidsmodell for hasard rate modellen, altså hvis man lar $h(t|\mathbf{X})$ betegne den betingede sannsynligheten for at hendelsen skal inntreffe i løpet av tidsintervallet $(t, t+1)$ gitt at den enda ikke hadde inntruffet like før starten av intervall, og hvis man antar at

$$\frac{h(t|\mathbf{X})}{1 - h(t|\mathbf{x})} = \frac{h_0(t)}{1 - h_0(t)} e^{\sum_{k=1}^{p} \beta_k x_k} \tag{24}$$

Da er partiell likelihooden for en diskret metode gitt ved

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \mathbf{s}_i}}{\sum_{q \in Q_i} e^{\boldsymbol{\beta}^T \mathbf{s}_q^*}} \tag{25}$$

Her betegner d_i alle de individene hvor hendelsen inntreffer på samme tidspunkt, t_i. q er et undersett av d_i som kan velges fra risikosette R_i ved tiden t_i , det vil si $q = [q_1, q_2, ..., q_{d_i}]$, og Q_i er alle mulige kombinasjoner av q med størrelse d_i. Og $\mathbf{s}_q^* = \sum_{j=1}^{d_j} \mathbf{X}_{qj}$

Hvis antall hendelser som inntreffer på samme tidspunkt er stort vill telleren til diskret metoden bli vanskelig å beregne.

Felles for disse tre metodene er at de kun gir en approksimasjon til partiell likelihooden. De brukes kun når flere hendelser inntreffer på samme tidspunkt ellers brukes vanlig partiell likelihood. Utregningen av partiell likelihooden når flere hendelser inntreffer på samme tidspunkt kan bregnes ved bruk av R og SAS. De fleste statistiske dataprogrammer bruker Breslow metoden som standard metode for hendelser som inntreffer på samme tidspunkt, men R bruker Efron metoden.

Estimering av overlevelsesfunksjonen, $S(t)$

Ut i fra den tidsuavhengige Cox proporsjonal hasard regresjonsmodellen kan man også estimere overlevelsesfunksjonen.
Cox proporsjonale hasard regresjonsmodellen er gitt ved

$$h(t|\mathbf{X}) = h_0(t)e^{\sum_{k=1}^{p}\beta_k x_k} \tag{26}$$

Ved bruk av denne ligningen, (26), kan man finne den kumulative hasard funksjonen, $H(t|\mathbf{X})$

$$\begin{aligned} H(t|\mathbf{X}) &= \int_0^t h(u|\mathbf{X})du && (27)\\ &= \int_0^t h_0(u)e^{\sum_{k=1}^{p}\beta_k x_k}du && (28)\\ &= H_0(t)e^{\sum_{k=1}^{p}\beta_k x_k} && (29) \end{aligned}$$

Fordi overlevelsesfunksjonen, $S(t|\mathbf{X}) = P(T > t|\mathbf{X})$, er relatert til den kumulative hasard funksjonen, $S(t|\mathbf{X}) = e^{-H(t|\mathbf{X})}$, kan man dermed finne overlevelses funksjonen, og får da

$$S(t|\mathbf{X}) = e^{-H(t|\mathbf{X})} = e^{-H_0(t)e^{\sum_{k=1}^{p}\beta_k x_k}} \tag{30}$$

Videre har man at

$$\begin{aligned} H_0(t) = \int_0^t h_0(u)du = -ln(S_0(t)) && (31)\\ S_0(t) = e^{-H_0(t)} && (32) \end{aligned}$$

Slik at den estimerte overlevelsesfunksjonen til Cox proporsjonal hasard regresjonsmodell da blir

$$S(t|\mathbf{X}) = S_0(t)^{e^{\sum_{k=1}^{p}\beta_k x_k}} \tag{33}$$

$S_0(t)$ kalles grunnlinje overlevelsesfunksjonen.

Proporsjonalitetsantagelse.

For en tidsuavhengig Cox proporsjonal hasard regresjonsmodell er det kun grunnlinje funksjonen som avhenger av tid, til gjengjeld er grunnlinje funksjonen ikke avhengig av forklaringsvariablene $\mathbf{X}$. Uttrykket $e^{\beta T\mathbf{X}}$ er uavhengig av tiden t, det betyr at verdien til forklaringsvariabelen ikke vil forandre seg med tiden, men de vil være konstant over tid, dermed er Cox modellen proporsjonal og derav navnet Cox proporsjonal hasard regresjonsmodell.

Proporsjonaliteten til en tidsuavhengig Cox hasard regresjonsmodell kan bevise ved at man har gitt to individer med henholdsvis følgende vektorer av forklaringsvariabler $\mathbf{X}^*$ og $\mathbf{X}$. Individenes forhold mellom hasard raten blir da

$$\frac{h(t|\mathbf{X}^*)}{h(t|\mathbf{X})} = \frac{h_0(t)e^{\sum_{k=1}^{p}\beta_k x_k^*}}{h_0(t)e^{\sum_{k=1}^{p}\beta_k x_k}} = e^{\sum_{k=1}^{p}\beta_k(x_k^*-x_k)} = \theta \tag{34}$$

$$\tag{35}$$

$$h(t|\mathbf{X}^*) = \theta h(t|\mathbf{X}) \tag{36}$$

Hvis antagelsen for proporsjonalitet ikke er tilfredsstilt er det to måter man kan håndtere dette på, enten bruke tidsavhengige variabler for å danne en utvidet Cox modell, mer om dette senere, eller man kan bruke en stratifisert Cox modell.

Stratifisert Cox regresjonsmodell.

Hvis man har l antall variabler som ikke tilfredsstiller antagelsen om proporsjonalitet, det vil si $z_1, z_2, ..., z_l$, og n variabler som tilfredsstiller antagelsen om proporsjonalitet, $x_1, x_2, ..., x_n$.
Så definerer man en nye variable z^* med l^* kategorier. For eksempel hvis en av z variablene representerte alderne til et individ kan man dele den inn i de individene som er yngre enn 45 år og de individene som er til og med 45 år og eldre, og hvis en av de andre z-variablene representerer røykestatus (røyker, har aldri røkt, tidligere røyker) kan man kombinere disse to z variablene inn i ulike kategorier. Hvis man representerer dette i en tabell får man følgende

	Røyker	Har aldri røkt	Tidligere røyker
< 45	1	2	3
≥ 45	4	5	6

Tabell 1: Stratifisert variable z^* bestående av 6 strata.

z^* består nå av disse 6 kategoriene, $l^* = 1, 2, ..., 6$, det vil si at når $z^* = 1$ betyr det at personen er yngre enn 45 år og røyker og når $z^* = 5$ betyr det at personen er eldre 44 år og har aldri røkt. Man sier at z^* er stratifisert og l^* er strataene.

Cox modellen vil se ut som følger

$$h_g(t|\mathbf{X}) = h_{0g}(t)e^{\sum_{k=1}^{n}\beta_k x_k} \tag{37}$$

hvor $g = 1, 2, ..., l^*$ og $h_{0g}(t)$ er grunnlinje hasard funksjonen for den l^*-te strataen for den stratifiserte variabelen z^*.
β koeffisientene er de samme for hver strata, men grunnlinje hasard funksjonen blir forskjellig for hver strata. z^* er innlemmet i grunnlinje hasard funksjonen.

$$h_{01}(t) = S_1(t) \tag{38}$$
$$h_{02}(t) = S_2(t) \tag{39}$$
$$\vdots \tag{40}$$
$$h_{0l}(t) = S_l(t) \tag{41}$$

2.5.4 Tidsavhengig Cox regresjonsmodell.

Tidsavhengig Cox proporsjonal hasard regresjonsmodell, også kalt utvidet Cox regresjonsmodell, består av forklaringsvariabler som avhenger av tiden t, det vil si at verdiene til disse variablene vil endre seg med tiden. Eksempler på tidsavhengige variabler kan være BMI, sivilstatus, blodtrykk og medisindose.

Tidsavhengige variabler skiller man ofte inn to grupper, eksterne variabler og interne variabler. For interne variabler er forandringen av verdien til disse variablene avhengig av egenskapene til individet. Eksempler på slike variabler kan være røykestatusen ved et bestemt tidspunkt, BMI ved et bestemt tidspunkt, eller blodtrykket ved et bestemt tidspunkt. Verdien til interne variabler kan kun måles så lenge individet er i livet.
Verdien til eksterne variabler avhenger av hvordan omgivelsene rundt individet forandrer seg, og disse variablene påvirker ofte flere individer samtidig. Noen ganger vil verdien til denne variabelen kunne forutsies slik at individet ikke trenger å være i livet for å kunne vite verdien til denne variabelen. Eksempler på slike variabler er alder hvis fødselsdato er oppgitt, forurensning ved et bestemt området ved et bestemt tidspunkt eller dosen medisin som skal gis til individet på et bestemt tidspunkt.

Cox proporsjonal hasard regresjonsmodell med tidsavhengige variabler er gitt ved

$$h(t, \mathbf{X}(t)) = h_0(t) e^{\sum_{k=1}^{p} \gamma_k x_k(t)} \tag{42}$$

γ_k er koeffisientene til de tidsavhengige forklaringsvariablene.
En Cox proporsjonal hasard regresjonsmodell bestående av både tidsuavhengige og tidsavhengige variabler er gitt ved

$$h(t, \mathbf{X}(t)) = h_0(t) e^{\sum_{k=1}^{p1} \beta_k x_k + \sum_{l=1}^{p2} \gamma_l x_l(t)} \tag{43}$$

hvor β er koeffisientene for forklaringsvariablene som er uavhengig av tid og γ er koeffisienten for forklaringsvariablene som er avhengig av tid.

Tidsavhengige variabler kan også oppstå når koeffisientene γ_k er avhengig av tid og forklaringsvariabelen er uavhengig av tid. Disse koeffisienten kalles da tidsvarierende koeffisienter og betegnes $\gamma_k(t)$.
Hvis $\gamma_k(t)$ er en lineær funksjon av tiden, $\gamma_k t$, så kan den uttrykkes som en tidsavhengig forklaringsvariabel fordi

$$\gamma_k(t) x_k = \gamma_k t x_k = \gamma_k x_k(t) \tag{44}$$

Den relative risikoen med tidsavhengige forklaringsvariabler vil forandre seg med tiden. Slik at den relative risikoen ved tidspunktet t mellom to grupper av individer eller mellom to individer med vektor av forklaringsvariabler $\mathbf{X}^*(t)$ og $\mathbf{X}(t)$ er dermed

$$Realtiv\ risiko\ \ ved\ \ t = \frac{h(t|\mathbf{X}^*(t)}{h(t|\mathbf{X}(t)} = \frac{h_0(t)e^{\sum_{k=1}^{p}\gamma_k x_k^*(t)}}{h_0(t)e^{\sum_{k=1}^{p}\gamma_k x_k(t)}} \tag{45}$$

$$= e^{\sum_{k=1}^{p}\gamma_k(x_k^*(t)-x_k(t))} \tag{46}$$

Estimering av koeffisientene $\boldsymbol{\gamma}$.

Estimeringen av γ koeffisientene gjøres på samme måte som for tidsuavhengige forklaringsvariabler. Gitt at man har et datasett bestående av n antall individer slik at $j = 1, 2, ..., n$. Hvert individ j har en vektor bestående av p antall forklaringsvariabler som er avhengig av tiden, t, $\mathbf{X}_j(t) = [x_{j1}(t), x_{j2}(t), ..., x_{jp}(t)]^T$. Hver vektor av tidsavhengige forklaringsvariabler, $\mathbf{X}_j(t)$ har uavhengig hendelsetid med sensureringstid.
Hvis hendelsene som inntreffer ikke inntreffer på samme tidspunkt og det er D antall hendelser som inntreffer i løpet av studiet slik at $t_1 < t_2 < ... < t_D$ for $i = 1, 2, ..., D$, da er partiell likelihooden, for γ koeffisientene for et individ j med p antall tidsavhengige forklaringsvariabler, hvor $k = 1, 2, ..., p$, og hvor hendelsen inntreffer ved tiden t_i, gitt ved

$$L(\gamma) = \prod_{i=1}^{D} \frac{e^{\sum_{k=1}^{p}\gamma_k x_{(i)k}(t_i)}}{\sum_{j\in R(t_i)} e^{\sum_{k=1}^{p}\gamma_k x_{jk}(t_i)}} \tag{47}$$

$$log(L(\gamma)) = \sum_{i=1}^{D}\left[\sum_{k=1}^{p}\gamma_k x_{(i)k}(t_i) - log\left(\sum_{j\in R_{t_i}} e^{\sum_{k=1}^{p}\gamma_k x_{jk}(t_i)}\right)\right] \tag{48}$$

$\mathbf{X}_{(i)}(t_i)$ representerer her vektoren av tidsavhengige forklaringsvariabler til individ i hvor hendelsen inntraff ved tidspunktet t_i. Risikosettet, $R(t_i)$ ved tidspunktet t_i består av alle individene hvor hendelsen enda ikke har inntruffet like før tidspunktet t_i.

Metoder for å behandle flere hendelser som inntreffer på samme tidspunkt for tidsavhengige forklaringsvariabler.

Hvis flere hendelser inntreffer på samme tidspunkt kan man bruke samme metoder som man bruker for tidsuavhengige forklaringsvariabler til å approksimere partiell likelihooden, det vil si Breslow metoden, Efron metoden eller diskret metoden.

Breslow metoden for tidsavhengige forklaringsvariabler.
Den approksimerte partiell likelihooden for tidsavhengige forklaringsvariabler ved bruk av Breslow metoden er

$$L(\gamma) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\gamma}^T \mathbf{s}_i(t)}}{\left[\sum_{j\in R_i} e^{\boldsymbol{\gamma}^T \mathbf{X}_j(t)}\right]^{d_i}} \tag{49}$$

Hvor $\mathbf{s}_i(t) = \sum_{j \in \mathbb{D}_i} \mathbf{X}_j(t)$.

Efron metoden for tidsavhengige forklaringsvariabler.
Den approksimerte partiell likelihooden for tidsavhengige forklaringsvariabler ved bruk av Efron metoden er gitt ved

$$L(\gamma) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\gamma}^T \mathbf{s}_i(t)}}{\prod_{j=1}^{d_i} \left[\sum_{k \in R_i} e^{\boldsymbol{\gamma}^T \mathbf{X}_k(t)} - \frac{j-1}{d_i} \sum_{k \in \mathbb{D}_i} e^{\boldsymbol{\gamma}^T \mathbf{X}_k(t)} \right]} \tag{50}$$

Diskret metode.
Den approksimerte partiell likelihooden for tidsavhengige forklaringsvariabler ved bruk av diskret metode er gitt ved

$$L(\gamma) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\gamma}^T \mathbf{s}_i(t)}}{\sum_{q \in Q_i} e^{\boldsymbol{\gamma}^T \mathbf{s}_q^*(t)}} \tag{51}$$

Hvor $\mathbf{s}_q^* = \sum_{j=1}^{d_i} \mathbf{X}_{qj}(t)$.

Koeffisientene kan også regnes ut ved hjelp av dataprogrammene R og SAS.

Tidsavhenging Cox modell og estimering av overlevelsesfunksjonen $S(t)$.

Den kumulative hasard funksjonen vil ha følgende uttrykk når forklaringsvariablene er tidsavhengig

$$H(t|\mathbf{X}(t)) = \int_0^t h(u|\mathbf{X}(u))du \tag{52}$$

$$= \int_0^t h_0(u) e^{\sum_{k=1}^{p} \gamma_k x_k(u)} du \tag{53}$$

$$\tag{54}$$

Siden forklaringsvariablene er avhengig av tid blir integralet av Cox regresjonsmodellen litt mer komplisert å regne ut, den estimerte overlevelsesfunksjon til en utvidet Cox regresjonsmodell blir dermed

$$S(t|\mathbf{X}(t)) = e^{-H(t|\mathbf{X}(t)} \tag{55}$$

$$= e^{-\int_0^t h_0(u) e^{\sum_{k=1}^{p} \gamma_k x_k(u)} du} \tag{56}$$

Proporsjonalantagelsen og tidsavhengighet.

Når forklaringsvariablene avhenger av tid vil antagelsen om proporsjonalitet ikke holde for en utvidet Cox regresjonsmodell.
Dette kan bevises gjennom at man har to individer, hvor hvert individ har henholdsvis en vektor av tidsavhengige forklaringsvariabler $\mathbf{X}^*(t)$ og $\mathbf{X}(t)$. Den relative risikoen blir da

$$\frac{h(t|\mathbf{X}^*(t))}{h(t,\mathbf{X}(t))} = \frac{h_0(t)e^{\sum_{k=1}^{p}\gamma_k x_k^*(t)}}{h_0(t)\sum_{k=1}^{p}\gamma_k x_k(t)} \tag{57}$$

$$= e^{\sum_{k=1}^{p}\gamma_k(x_k^*(t)-x_k(t))} \tag{58}$$

Altså vil ikke forholdet være konstant, men variere med tiden og antagelsen om proporsjonalitet holder ikke.

2.5.5 Hvorfor Cox regresjonsmodell blir mye brukt i overlevelsesanalyse.

Det at Cox proporsjonal hasard regresjonsmodellen har mange og unike egenskaper gjør at den i dag er mye brukt i overlevelsesanalyse.
For eksempel inneholder Cox regresjonsmodell et eksponential ledd noe som fører til at kravet om at hasard raten ikke kan være negativt oppfylles.
En Cox regresjonsmodell gir en god estimering av regresjonskoeffisientene selv om den kun er en semi-parametrisk modell, i tillegg vil resultatene fra Cox regresjonsmodellen tilnærme seg den rette parametriske modellen slik antagelsen om en parametrisk modell ikke blir feil.
Hasard funksjon $h(t, \mathbf{X})$ og overlevelses funksjon $S(t, \mathbf{X})$ kan estimeres fra en Cox regresjonsmodell uten at man kjenner til grunnlinje hasard funksjonen.
Dette er er bare for å neven noen av dens egenskaper.

2.5.6 Hvordan finne den beste Cox proporsjonale hasard regresjonsmodellen.

2.5.7 Test av signifikans.

For og teste om en eller flere av variablene har en effekt på overlevelsesevnen brukes det tre tester Wald test, Score test og loglikelihood rate test. Disse tre testene tester alle den samme hypotesen,

$$H_0 : \gamma_1 = \gamma_2 = ... = \gamma_p = 0 \qquad H_1 : \gamma_1 = \gamma_2 = ... = \gamma_p \neq 0 \tag{59}$$

$$H_0 : \boldsymbol{\gamma} = 0 \qquad H_1 : \boldsymbol{\gamma} \neq 0 \tag{60}$$

H_0 hypotesen forkastes hvis p-verdien til disse tre statistikkene er større enn 5% og man sier da at variabelen har en effekt på overlevelsesevnen.

Wald test.
Wald test statistikken er gitt ved

$$z_W^2 = \hat{\boldsymbol{\gamma}}^T I(\hat{\boldsymbol{\gamma}}) \hat{\boldsymbol{\gamma}} \tag{61}$$

$\hat{\boldsymbol{\gamma}}$ er her de estimerte Cox regresjonskoeffisientene, $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_p]^T$. $I(\hat{\boldsymbol{\gamma}})$ er den p×p informasjonsmatrisen til $\hat{\boldsymbol{\gamma}}$, se appendiks 4.1 for formel.
z_W^2 er kji-kvadratfordelt med p frihetsgrader hvis H_0 er sann for store utvalg.

Likelihood rate testen.
Likelihood rate testen er gitt ved

$$z_{LR}^2 = 2[logL(\hat{\boldsymbol{\gamma}}) - logL(0))] \tag{62}$$

z^2_{LR} er kji-kvadratfordelt med p frihetsgrader hvis H_0 er sann for store utvalg. $logL(\hat{\boldsymbol{\gamma}})$ er her gitt ved uttrykket som i (48).

Score test.
Score test statistikken er gitt ved

$$z^2_{SC} = U(0)^T I()\hat{\boldsymbol{\gamma}})^{-1} U(0) \tag{63}$$

z^2_{SC} er kji-kvadratfordelt med p frihetsgrader hvis H_0 er sann for store utvalg.

Alle de tre testene kan gjøres i det statistiske dataprogrammet R.

2.5.8 Konfidensintervall.

Et konfidensintervall, KI, består av en rekke verdier, som er beregnet ut fra et utvalg observasjoner, og som man med en bestemt prosentandel kan være sikker på at inneholder den sanne verdien til en parameter. Det betyr at hvis estimeringen av parameteren ble gjentatt flere ganger så ville en bestemt prosentandel av de beregnede KI inneholde den sanne verdien av parameteren. Man foretrekker ofte korte intervall med en høy prosentandel.

Alle verdiene i KI er mulige verdier som den estimerte parameteren kan ha, alle verdier som er utenfor KI blir forkastet.
KI kan brukes til å si noe om signifikansen til en estimert parameter. Hvis man regner ut et 95% KI for en parameter og intervallet ikke inneholder verdien 0, så indikerer dette at at parameteren er signifikant forskjellig fra 0. Ved hypotesetesting betyr signifikant forskjellig fra 0 det samme som at H_0 hypotesen kan forkastes og den alternative hypotesen, H_1 kan godtas.

Et 95% KI til den relative risikoen for en kategorien av en variable sammenlignet med referansekategorien er gitt ved

$$KI = \left[e^{\hat{\gamma}_k - 1.96 \cdot se(\hat{\gamma}_k)}, e^{\hat{\gamma}_k + 1.96 \cdot se(\hat{\gamma}_k)} \right] \tag{64}$$

KI kan gi en indikasjon på om den estimerte relative risikoen til en kategori av en variable sammenlignet med referansekategorien er til å stole på eller ikke. Et vidt konfidensintervall er et tegn på at den estimerte relative risikoen ikke er til å stole på.
Hvis KI til den relative risikoen inneholder verdien 1 betyr det at variabelen ikke er signifikant.

2.5.9 Interaksjonsledd.

Cox regresjonsmodellen kan inneholde interaksjoner mellom variablene, det vil si at en variabel kan påvirke resultatene for en annen variabel. For eksempel kan det finnes en interaksjon mellom det å røyke og helsen til et individ eller kjønn og helse.

En måte man kan teste om Cox regresjonsmodellen inneholder interaksjoner er å teste om den estimerte γ koeffisienten, til produktet av de to variablene, er signifikant ved bruk av de ulike signifikans testene, man kan også bruke KI for å se om KI for den estimerte relative risikoen inneholder verdien 1. Hvis den gjør det er det et tegn på at interaksjonsleddet ikke er signifikant.
Ved å dele inn noen av variabelen i kategorier, stratifisere, for så og plotte deres estimerte overlevelseskruve for en av de gjenværende forklaringsvariablene kan man finne ut om modellen skal inneholde interaskjonsledd grafisk. Hvis det er typiske mønster som går igjen i plottet for en av kategoriene kan det tyde på at det er en interaksjon mellom denne kategorien og forklaringsvariabelen.

2.5.10 Proporsjonalitets antagelsen.

Man må teste om Cox regresjonsmodellen tilfredsstiller antagelsen om proporsjonalitet, dette kan gjøres grafisk, ved bruk av statistiske tester og ved bruk av residualer.
Grafisk kan man plotte overlevelsesfunksjonen for hver kategori av en kategorisk variable. Hvis grafene er parallelle er antagelsen om proporsjonalitet tilfredsstil, hvis grafene krysser hverandre er antagelsen om proporsjonalitet ikke tilfredsstilt.
Ved bruk av statistiske tester vil en p-verdi som er større en et gitt signifikansnivå indikere at antagelsen om proporsjonalitet er tilfredsstilt.
Også residualer kan brukes til å kontrollere antagelsen om proporsjonalitet. Schoenfeld residualer og martingale residualer kan brukes til dette.

Hvis antagelsen om proporsjonalitet ikke holder kan man enten bruke stratifisering eller danne en interaksjon mellom tid og de variabelen hvor antagelsen om proporsjonalitet ikke holder.

2.5.11 Residualer.

Til å se på Cox regresjonsmodellens tilpassing til dataene kan man plotte residualene til den aktuelle modellen.
Et residual er feilen i tilpassingen av en modell. I lineære modeller, $y_i = \beta_0 + \beta_1 x_i$, bergenes residualene ut ved å finne forskjellen mellom den faktiske modellen y_i og den estimerte modellen $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$

hvor e_i er den $i - te$ residualen. Man ønsker at verdien til residualen skal være minst mulig slik at den estimerte modellen er nærmest lik den sann modellen.
Cox-Snell residualer, Martingale residualer og deviance residualer er noen residualer som kan brukes til å analysere tilpassingen av en Cox regresjonsmodell til datamaterialet.

Cox-Snell residualer.

Til å se på modellens overordnede tilpasningen til dataene brukes Cox-Snell residualer. Hvis Cox regresjonsmodellen har blitt tilpasset korrekt til dataene, det vil si at de estimerte $\hat{\gamma}_k$ er tilnærmet lik de faktiske γ_k, vil den kummulative hasard funksjonen, $H(t|\mathbf{X}(t))$ med tidsavhengige variabler, ha en eksponentialfordeling med en hasard rate lik 1.
Cox-Snell residualer for en utvidet Cox regresjonsmodell med j antall individer, $j = 1, 2, ..., n$, er gitt ved

$$r_j = \hat{H}_{0j}(T_j)e^{\sum_{k=1}^{p}\hat{\gamma}_k x_{jk}(T_j)}$$

$\hat{H}_0$ er her den estimerte Breslow estimatoren av grunnlinje hasard raten.
Ved å plotte den kumulative hasard raten til Cox-Snell residualene mot Cox-Snell residualene vil man kunne få en rett linje med stigningstall 1, linjen vil ha en vinkel på 45°, gjennom origo hvis Cox regresjonsmodellen er godt tilpasset dataene.
Hvis noen av de j individene er sensurert så vil tiden t_j bli sensurert.

Martingale residualer.
Martingale residualene prøver å bestemme den funksjonelle formen til en forklaringsvariabel for å kunne forklare dens effekt på overlevelsesevenen gjennom Cox regresjonsmodellen. Dens funksjonelle form kan for eksempel være $log(x)$, $xlog(x)$, x^2 eller en diskret versjon av forklaringsvariabelen. Martingale residualene inneholder ingen sensureringer.
Martingale residualene for en utvidet Cox modell er gitt ved

$$\hat{M}_j = N_j(\infty) - \int_0^{\infty} Y_i(t)e^{\hat{\boldsymbol{\gamma}}^T \boldsymbol{X}_j(t)}d\hat{H}_0(t) \tag{65}$$

j er antall individer, $j = 1, 2, ..., n$. $N_j(t)$ er en indikator variabel som er 1 hvis individ j har opplevd hendelsen, og 0 hvis individet enda ikke har opplevd hendelsen. $Y_j(t)$ er også en indikator variabel som sier noen om individet er sensurert eller ikke like før hendelsen inntreffer ved tiden t. $\hat{H}_0(t)$ er den kumulative Breslow estimatoren til grunnlinje hasard funksjonen.
Martingale residualen er altså observerte antall hendelser minus forventet antall hendelser.
Martingale residualene har den egenskapen at $\sum_{j=1}^{n}\hat{M}_j = 0$ og hvis Cox regresjonsmodellen er tidsuavhengig så er martingale residualene gitt ved $\hat{M}_j = \delta_j - \hat{H}_0(T_j)e^{\sum_{k=1}^{p}\hat{\beta}_k x_{jk}} = \delta_j - r_j$, hvor r_j er Cox-Snell residualene og kan tolkes som forventet antall hendelser som inntreffer og δ_j indikerer om individet er sensurert eller ikke. For store utvalg vil $\hat{M}_j$-ene være ukorrolerte utvalg fra en fordeling med forventning lik 0 og deres plott vil ikke være særlig informative. Martingale residualen er ikke symmetrisk rundt 0, men har domen rundt $(-\infty, 1)$.
Plotter man martingale residualene mot en av variabelen i modellen kan man finne ut om variabelens funksjonelle form er den rette. Hvis plottet er ikke-lineært tyder dette på at variabelens funksjonelle form er feil, man kan da prøve å erstatte den med en log transformasjon, ta kvadratroten til denne variabelen eller opphøye den i andre eller polynom av høyere grad.

Deviance residualer.
For at martingale residualen skal kunne være symmetrisk rundt 0, hvis Cox regresjonsmodellen er korrekt, kan man bruke transformasjon, resultatet blir det man kaller deviance residualer.
Deviance residualer er gitt ved

$$D_j = sign[\hat{M}_j]\sqrt{-2[\hat{M}_j + \delta_j log(\delta_j - \hat{M}_j)]} \tag{66}$$

$\hat{M}_j$ er martingale residualet for individ j, δ_j er 1 hvis individet har blitt sensurert og 0 hvis det ikke har blitt sensurert.
Observasjoner som har en relativ stor deviance residual verdi kalles en utenforligger. Disse utenforliggerne er de som ikke har blitt tilpasset godt nok av Cox regresjonsmodellen. Har man residualer som er svært negative er det et tegn på at Cox regresjonsmodellen overestimere sannsynligheten for at hendelsen skal inntreffe, er residualene svært positive er det et tegn på at Cox regresjonsmodellen underestimerer sannsynligheten for at hendelsen skal inntreffe.
Deviance residualen kan man enten plotte mot tiden, t eller mot antall observasjoner.

2.5.12 AIC.

Akaike information Criterion, AIC, er en metode som prøver å finne en modell som passer bra til datamaterialet, men som samtidig består av få parametere. AIC er gitt ved

$$AIC = -2logL + kp \tag{67}$$

p er her antall regresjonskoeffisienter i modellen. L er maksimum verdien til likelihoodfunskjonen til modellen, og k er en forhåndsbestemt konstant, ofte er denne satt til å være 2.
Man sammenligner ofte to eller feler modeller og desto lavere AIC verdi desto bedre er modellen. Ved å starte med en tom Cox modell for så å føye til en og en variabel til modellen vil AIC verdien avta, men ved et punkt vil den begynne å øke, dette er et tegn på at den tilføyde variabelen er unødvendig. Man kan også bruke en full Cox modell, det vil si en modell som består av alle de opprinnelige variablene, for så og ta bort en og en variabel for å se hvilken av de gjenværende variabelen som gir minst AIC verdi. Ved et punkt vil AIC verdien begynne og øke, noe som er et tegn på at man har tatt bort en variabel for mye og man står igjen med en endelig modell.

Alle disse metodene kan brukes til å konstantere om man har funnet den Cox regresjonsmodell som tilpasser dataene godt.

3 Analyse og Resultater.

3.1 Hva går oppgaven ut på.

Som nevnt tidligere skal det i denne oppgaven brukes overlevelsesanalyse til å lage en Cox proporsjonal regresjonsmodell bestående av tidsavhengige forklaringsvariabler for å se hvilken av disse som har en effekt på overlevelsen til kvinner.
Datamaterialet som her er blitt brukt er hentet i fra forskningsprosjektet "Kvinner og kreft" (The Norwegain Women and Cancer post genome cohort study) ved universitetet i Tromsø og består av forklaringsvariabler som representerer svar fra tre spørreundersøkelser som ble sendt ut til ulike kvinner.
Den første spørreundersøkelsen ble besvart i mai 1991 av 57 561 kvinner. Den andre spørreundersøkelsen ble sendt ut i 1998 til de samme kvinne som besvarte spørreundersøkelse nummer 1 og den siste spørreundersøkelsen ble sendt ut i 2004 til de samme kvinnen som besvarte spørreundersøkelse nummer 2. Spørreundersøkelsene inneholdt blant annet spørsmål om hvor mange år kvinnen hadde gått på skole, hvor gammel var hun da hun fikk sin første menstruasjon, om kvinnen gikk regelmessig til undersøkelse av brystene sin til mammografi, hvor ofte hun spiste ris og spagetti og hun ble blant annet bedt om å rangere sin egen fysiske aktivitet på en skala fra 1-10.
I denne oppgaven ble ikke alle forklaringsvariablene fra det originale datamaterialet brukt, men kun et utvalg. Hendelsen som ble analysert var tiden inntil døden inntraff, men dødsårsaken til kvinnen ble ikke tatt med. Starten på studiet ble satt til 15.5.1991 og ble avsluttet 31.12.2013.
Analysen ble gjort ved hjelp av ulike funksjoner i det statistiske dataprogrammet R, se appendiks.

3.2 Behandling av datamateriale.

Før selve analysen kunne begynne måtte det originale datamaterialet endres på. De utvalgte forklaringsvariablene som ble brukt finnes i en tabell oversikt i appendiks, i tillegg til en oversikt over nye variabler som ble satt inn.
Datasetet som ble brukt i den tidsavhengige Cox proporsjonale hasard regresjonsmodellen besto av følgende forklaringsvariabler
SKOLE som besto av to kategoriser, *T.O.M ungdomskole* og *Høyere utdanning. T.O.M ungdomskole* inneholdt alle som svarte at de hadde totalt gått på skole i 10 år eller færre, mens *Høyere utdanning* var alle de som hadde svart at de tilsammen hadde gått mer enn 10 år.
Egenhelse var en variabelen som inneholder svarene på spørsmålet hvor kvinnene ble bedt om å rangere deres egen oppfatning av helsen på en skala fra 0 til 3 der 0 er meget god og 3 er meget dårlig. Denne skalaen ble reversert slik at 0 ble meget dårlig, 1 ble dårlig, 2 ble god og 3 ble meget god.
Forklaringsvariabelen ***Aktivitet*** ble kodet til å være en kategorisk variabel. Denne variabelen fortalte noe om den fysiske aktiviteten til kvinnen. Den opprinnelige variabelen inneholdt tall far 1 til 10 hvor 1 var svært lite aktiv og 10 var svært aktiv, jeg omkodet denne variabelen til å inneholde tre kategorier 1 ble svært aktiv, det vil si de som var i kategorien 1-3 i den opprinnelige variabelen, 2 ble moderat aktiv, de som var i kategorien 4-7 i den opprinnelige variabelen, og 3 ble svært lite aktiv, de som var i kategorien 8-10 i den opprinnelige variabelen.
I spørreundersøkelsen ble det spurt om kvinnen noen gang hadde røykt og i så fall hvor mange sigaretter hun røykte daglig, dette ble brukt til å lage en kategorisk variabel kalt ***SmokingHabit***

som besto av følgende kategorier *NONsmoker*, *EXsmoker*, *max14Daily* og *15plussDaily*.
Det gjennomsnittlige alkoholinntaket til kvinnene finnes i den kategoriske variabelen kalt ***Alkoforbruk*** og besto av følgende kategorier, *Avhold*, *Lite_alkoholinntak*, *Moderat_alkoholinntak* og *Stort_alkoholinntak*. For å rangere alkoholmengden ble det brukt informasjon fra *drink-less.com*.
Vekten og høyden til kvinnen er brukt til å lage en forklaringsvariabel kalt ***BMI***. BMI (Body Mass Index), også kalt KMI (Kropps Masse Indeks) på norsk, ble regnet ut på følgende måte

$$BMI = \frac{Vekt}{Høyde^2} = \frac{Kg}{m^2}$$

Forklaringsvariabelen ***BMI*** ble delt inn i tre kategorier, *Undervektig*, *Normal* og *Overvektig* ved bruk av informasjon fra *helsedirektoratet.no*. Disse kategoriene ble senere endret til å være *Fedme*, *Normal/overvektig* og *Undervektig* da det nesten ikke var noe forskjell på sannsynligheten for å overleve for *Normal* og *Overvektig*
Variabelen ***Alder*** inneholdt informasjon om hvor gammel kvinne var når hun besvarte spørreundersøkelsen. Denne forklaringsvariabelen var en kontinuerlig variabel.
Forklaringsvariabelen ***Barn*** fortalte noe om hvor mange barn kvinne hadde født, her ble kvinnen bedt om også å ta med de barna som var dødfødt og som hadde død ved et sener tidspunkt i livet. Denne variabelen var også en kontinuerlig variabel, men jeg valgte å lage en kategorisk forklaringsvariabel kalt ***Barn.fac*** som besto av to kategorier nemlig de som ikke hadde født barn *Ikke_født* og de som hadde født barn *Har født*.
For å kunne analysere tiden inntil kvinnene døde måtte man også ha en forklaringsvariabel som sier noen om tiden inntil døden inntraff. Med tidsavhengige variabler, som jeg hadde her, måtte jeg ha to forklaringsvariabler som sa noe om tiden, tidsvariabelen ***Start*** og en tidsvariabel som ble kalt ***Slutt***, som begge var kontinuerlige. ***Start*** variabelen inneholder tiden på når kvinnen besvarte den første spørreundersøkelsen, dette er satt til å være tiden $t = 0$, tiden fra den første spørreundersøkelsen til hun besvarte spørreundersøkelse nummer 2, og tiden fra hun besvarte spørreundersøkelse nummer 3. Tiden er i antall dager fra studiet startet det vil si fra første gang kvinnen besvarte spørreundersøkelse nummer 1. Variabelen ***Slutt*** fortalte ved hvilken tid hendelsen inntraffer eller ved hvilken tid kvinnen ble sensurert.
I tillegg til at man måtte ha en variabel som sa noe om tiden så måtte man også ha en variabel som sa noe om når hendelsen inntreffer det vil si når døden inntreffer. En variabel kalt ***Status*** sier noe om døden har inntruffet 1, 0 hvis hendelsen ikke har inntruffet.

3.3 Mangler og feil i datamaterialet.

Datamaterialet inneholdt en del feil og mangler og det ble gjort et forsøk på å rette opp så mange feil som mulig.
Et lite antall kvinner hadde blitt sensurert før studiet startet, det betyr at disse kvinnen fikk negativ start tid. Disse kvinnen valgte jeg å ikke ta med i analysen slik at totalt var det 57 555 kvinner som hadde besvart spørreundersøkelse nummer 1.
Kvinner som hadde svart på spørreundersøkelse 1, men i ikke 2 og 3 valgte jeg å sensurer. Slutt tiden deres ble satt til å være en dag før den første kvinnen svarte på neste spørreundersøkelse.
En kvinne hadde blitt sensurert like før hun skulle ha besvart spørreundersøkelse nummer 3, men kvinnen hadde faktisk besvart spørreskjemaet etter at hun hadde blitt sensurert. Denne kvinne ble satt til å overleve utover studiets slutt og dermed ble hun høyre sensurert.

En kvinne dødd samme dagen hun hadde besvart spørreundersøkelsen. Satte at slutt tiden til denne kvinnen ble dagen etter at hun besvarte undersøkelsen.
Noen kvinner manglet dato for når de hadde besvart spørreundersøkelsen. Satte inn en dato basert på hvilken dato hvor flest kvinner hadde besvart denne undersøkelsen.
Kvinner som har valgt å ikke svare på enkelte spørsmål i spørreundersøkelsen lot jeg være med, R vil automatisk fjerne disse ved utregning av koeffisientene til den utvidede Cox regresjonsmodellen.
Noen hadde svart at de har født færre barn i spørreundersøkelse nummer 2 eller nummer 3 enn i spørreundersøkelse nummer 1. Endret svaret deres i disse spørreundersøkelsen til å ha født like mange barn som i spørreundersøkelse nummer 1.
Noen hadde også svart at de hadde røkt i spørreundersøkelse nummer 1, men i spørreundersøkelse 2 og 3 hadde de svart at de aldri hadde røkt. Endret disse svarene til at de hadde røkt.
En kvinne hadde besvart at hun var avholdskvinne, men at hun drakk gjennomsnittlig 1.52 g daglig.
Når det kom til spørsmålet hvor mange års skolegang hun hadde i alt, og her skulle også antall år på folkehøyskole og ungdomskole tas med, var det noen som har svart 1, 2 og 3 år. Disse satte jeg i kategorien *T.O.M ungdomskole*.

3.4 Valg av tidsskala.

Det ble brukt en tidsavhengig Cox regresjonsmodell til å modelere hasard raten til kvinnene, finne den relativ risiko og finne hvilken av forklaringsvariablene som hadde en effekt på overlevelsesevnen til kvinnene, dette fordi forklaringsvariablene var avhengige av tiden. For eksempel ville noen av kvinnene svarer at de hadde født flere barn i spørreundersøkelse nummer 2 enn i nummer 1, eller at hun hadde lavere BMI i spørreundersøkelse nummer 3 enn i spørreundersøkelse nummer 1.
Både kvinnens alder og hvilken dato hun hadde besvart spørreundersøkelsen ble oppgitt i datamaterialet dette kunne brukes til å fastslå tiden. Jeg valgte å måle tiden i antall dager fra kvinne for første gang svarte på spørreundersøkelse nummer 1, slik at tidspunktet ble $t = 0$ for forklaringsvariablene ved spørreundersøkelse nummer 1. Tiden til forklaringsvariablene fra spørreundersøkelse nummer 2 ble antall dager siden hun besvarte spørreundersøkelse nummer 1 til hun besvarte spørreundersøkelse nummer 2, og tiden for forklaringsvariabelen for spørreundersøkelse nummer 3 ble tiden fra hun besvarte spørreundersøkelse nummer 1 til hun besvarte spørreundersøkelse nummer 3.
For de kvinnen som døde eller ble sensurert i løpet av studiet ble tiden målt i antall dager fra hun besvarte spørreundersøkelse nummer 1 til hun døde eller ble sensurert.
Slutt tiden til studiet ble satt til å være 31.12.2013. For de som ble høyre sensurert, ble slutt tiden satt til være tiden i antall dager fra kvinnen besvarte spørreundersøkelse nummer 1 til studiet ble avsluttet.

3.5 Gjennomsnittlig levetid.

Gjennomsnittlig levetid for kvinner, de sensurert kvinne har ikke blitt tatt med, i løpet av tiden studiet pågikk var 56 år.
Av de totalt 57 555 kvinne som deltok i studiet var det 2 198 kvinner som døde i løpet av den tiden studiet pågikk, det utgjør 3.82% av det totale antallet.

42 828 kvinner ble sensurert i løpet av studiet, mens 12 529 ble høyre sensurert, det vil si de var enda i livet når studiet ble avslutte, dette utgjorde henholdsvis 74.4% og 21.8% av det totale antallet.

Spørreundersøkelse	**1**	**2**	**3**
Antall besvarte spørreundersøkelser	57 555	46 960	35 805
Antall kvinner som døde	503	721	974
Antall kvinner som ble sensurert i løpet av spørreundersøkelsen	3 920	4 297	22 302
Antall kvinner som ble sensurert fordi de ikke svarte på neste spørreundersøkelse	6 172	6 137	

Tabell 2: Tabell oversikt over antall kvinner som har svart på spørreundersøkelsene, antall kvinner som har død og antall kvinner som har blitt sensurert.

Spørreundersøkelse	**1**	**2**	**3**
Yngste kvinne	34	41	47
Eldste kvinne	49	55	63
Gjennomsnittsalder for de som har svart	41	48	54
Gjennomsnittsalder for de som døde	43	50	56

Tabell 3: Tabell oversikt over alderen på den yngste og eldste kvinnen som har svart på spørreundersøkelsene, gjennomsnittsalderen på de kvinnene som har svart og gjennomsnittsalderen på de kvinnen som har død.

Spørreundersøkelse	**1**	**2**	**3**
BMI			
Fete ($BMI > 35$)	434	849	879
Normal/overvektig ($18.5 < BMI \leq 35$)	53 968	44 317	32 979
Undervektig ($BMI \leq 18.5$)	1 991	722	443
Ikke svart	1 162	1 078	1 505
SKOLE			
T.O.M ungdomskole	21 211		
Høyere utdanning	35 474		
Ikke svart	870		
Alkoforbruk			
Avhold (0 gram alko. pr dag)	5 832	3 933	2 666
Lite alkoholinntak (mer enn 0 og mindre enn 14 gram alko. per dag)	38 781	36 485	28 214
Moderat alkoholinntak(mindre enn 24 og større enn 14 gram alko. per dag)	1 269	1 073	700
Stort alkoholinntak (mer enn 24 gram alko. pr dag)	519	95	74
Ikke svart	11 154	5 374	4 151
Smokinghabit			
NONsmoker	18 912	16 003	10 400
EXsmoker	22 316	13 018	14 060
max14Daily	10 846	8 907	5 910
15plussDaily	5 297	4 552	2 392
Ikke svart	184	4 480	3 043
Egenhelse			
Meget god	19 489	14 753	11 601
God	31 777	26 636	20 344
Dårlig	3 370	2 743	2 240
Meget dårlig	296	148	127
Ikke svart	2 623	2 680	1 493
Aktivitet			
Svært aktiv	9 122	5 698	6 637
Moderat aktiv	35 525	32 938	24 027
Svært lite aktiv	6 961	5 427	3 419
Ikke svart	5 947	2 897	1 722
Fødsler			
Har født barn	51 859	42 939	31 463
Har ikke født barn	5 696	4 021	481
Har ikke svart	0	0	3 861

Tabell 4: Tabelloversikt over hvor mange som har svart de ulike svaralternativene for variablene som ble brukt i denne analysen.

3.6 Utførelse av analyse.

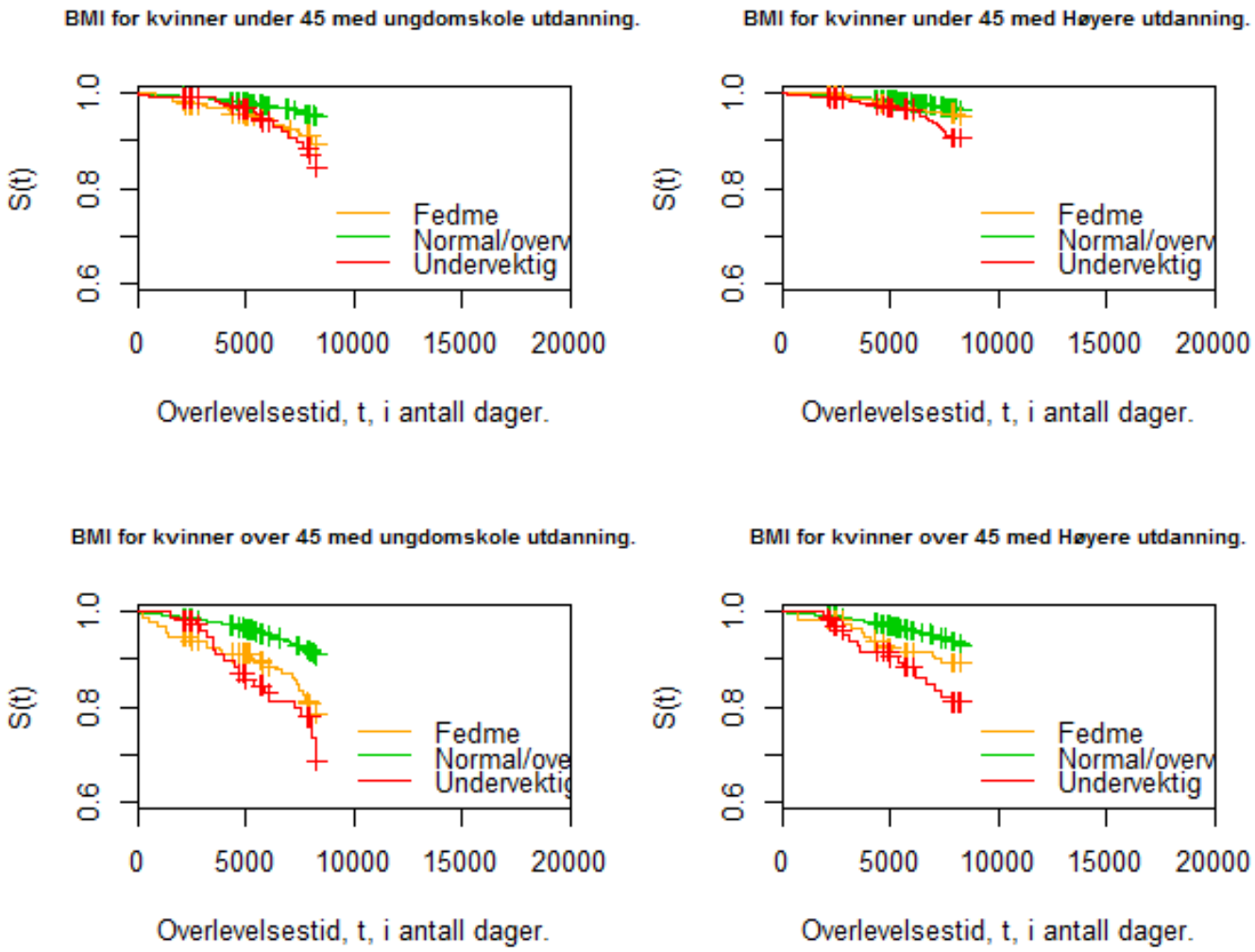
Etter at datamaterialet hadde blitt kodet om og en rekke feil og mangler hadde blitt forsøkt å rettet på kunne det nå utføre en overlevelsesanalyse ved bruk av utvidet Cox regresjonsmodell i statistikk programmet R.
Den utvidede Cox regresjonsmodellen ville da være på formen

$$h(t|\mathbf{X}(t)) = h_0(t)e^{\sum_{k=1}^{p}\gamma_k x_k(t)}$$

hvor man måtte ta høyde for at modellen også kunne inneholde interaksjonsledd.

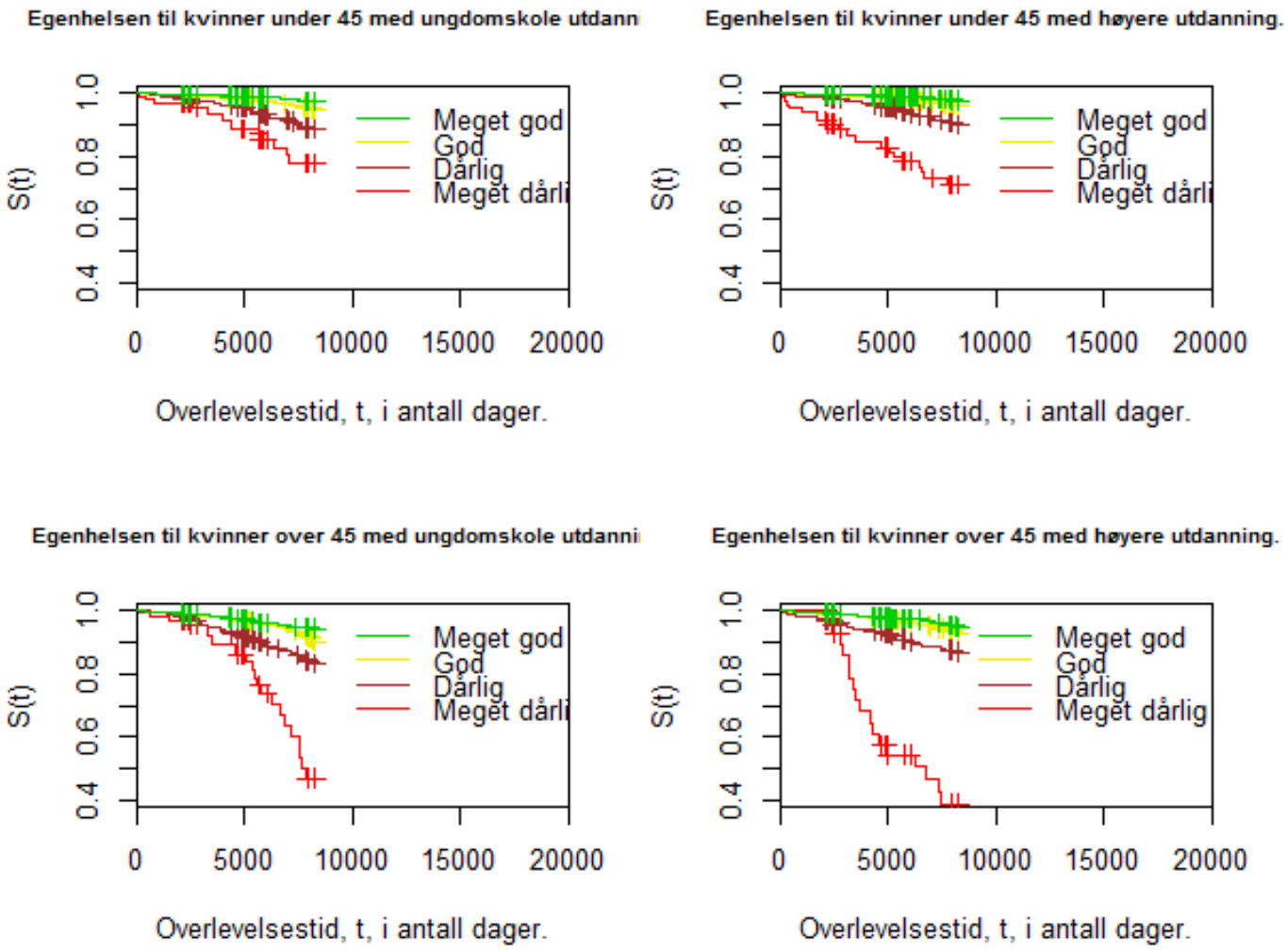
Interaksjonsledd.

Før man kunne sette opp en utvidet Cox regresjonsmodellen måtte man kontrollere om det fantes interaksjoner mellom noen av forklaringsvariablene. Det kunne kanskje være naturlig og tro at det vil være en interaksjon mellom ***BMI***, ***Egenhelse***, ***Aktivitet***, og ***SmokingHabit***, en interaksjon mellom ***Aktivitet***, ***Egenhelse*** og ***Alder***, og at det kunne være en interaksjon mellom det og ha født barn og antall år utdannelse.
Til å finne ut om modellen inneholdt interaksjonsledd valgte jeg å dele datamaterialet inn i fire grupper etter alder og utdannelse ved spørreundersøkelse nummer 1. De fire gruppene ble da de som var under 45 år ved spørreundersøkelse 1 og som hadde *T.O.M ungdomskole*, de som var under 45 år ved spørreundersøkelse 1 og som hadde *Høyere utdanning*, de som var over 45 år ved spørreundersøkelse 1 og som hadde *T.O.M ungdomskole* og de som var over 45 år ved spørreundersøkelse 1 og hadde *Høyere utdanning*. Dette var for å se om det kunne være en interaksjon mellom alder eller utdannelse og noen av de andre forklaringsvariabelen. I R valgte jeg å bruke Kaplan-Meier plott av overlevelsesfunksjonene av alle forklaringsvariablene for disse fire gruppene. Jeg tar ikke med alle plottene her, men et utvalg for vise hvordan man kan se om det er en interaksjon eller ikke.



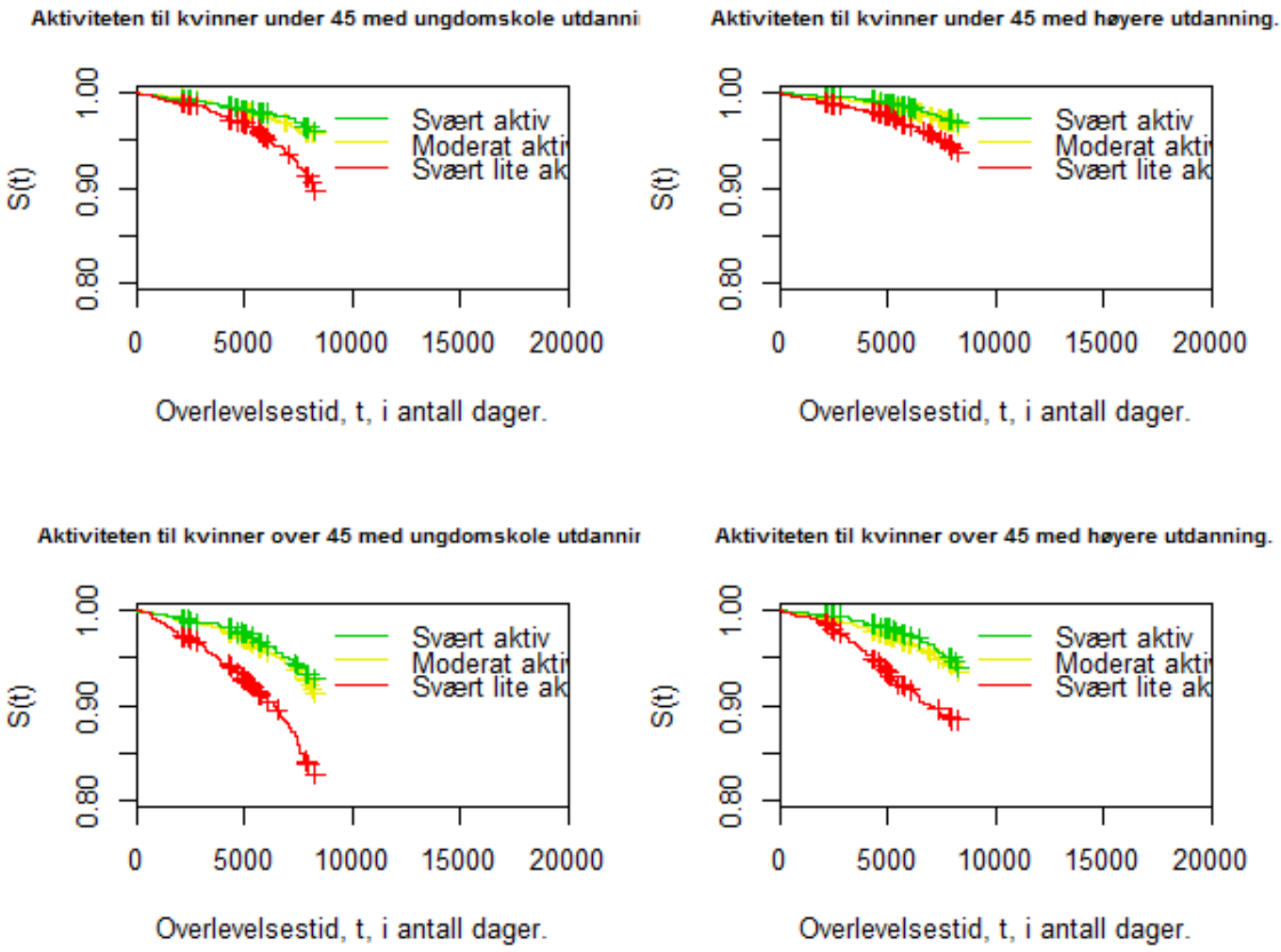
Figur 2: Kaplan-Meier plott av overlevelsesfunksjonen til de fire gruppene og ***BMI***.

Ut i fra plottet, figur 2, kunne man se at det var kvinner som var i kategorien undervektig som hadde minst sannsynlighet for å overleve utover tiden t og de som var i kategorien *Normal/overvektig* har størst sannsynlighet for å overleve utover tiden t.
Den opprinnelige variabelen ***BMI*** besto av kategoriene *Overvektig*, *Normal* og *Undervektig*, men fordi Kaplan-Meier plottet viste at det ikke var noen stor forskjell på det og være i kategorien normal og overvektig ble disse kategoriene endret på. Variabelen ***BMI*** består nå av følgende kategorier *Undervektig* de som har en BMI under 18.5, *Normal/overvektig* de som har en BMI mellom 18.5 og 35 og *Fedme* de som har en BMI større enn 35.
Siden plottene for de som var over 45 år hadde litt større spredning mellom kategoriene enn de som var under 45 år kan det tyde på at det var en interaksjon mellom alder og BMI. For å bekrefte dette måtte man ta med dette interaksjonsleddet i Cox regresjonsmodellen og se om den hadde en signifikant p-verdi.



Figur 3: Kaplan-Meier plott av overlevelsesfunksjonen til de fire gruppene og ***Egenhelsen***.

Kaplan-Meier plottet, figur 3, viste at det var de som var i kategorien meget dårlig egenhelse som hadde minst sannsynlighet for å overleve utover tiden t. De kvinnen som var i kategorien *meget god* og *god* egnehelse hadde nesten samme sannsynlighet for å overleve utover tiden t, dette var et tegn på at disse to kategorien egentlig kunne slås sammen. Når det kommer til interaksjoner viste plottet at det kunne være en interaksjon mellom ***Alder*** og ***Egenhelse***. De over 45 år hadde litt større spredning mellom overlevelsesfunksjonen enn de som var under 45 år.

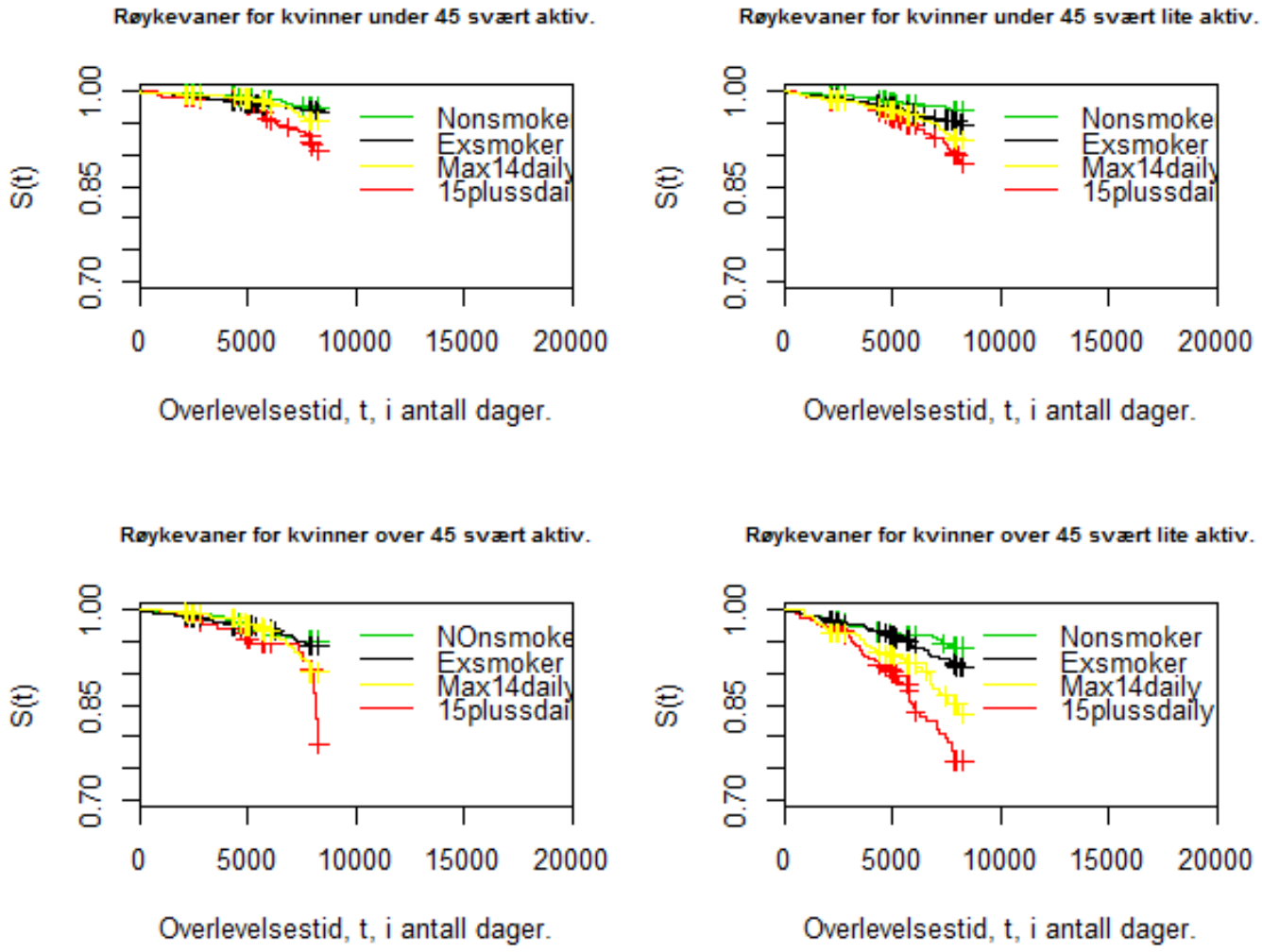


Figur 4: Kaplan-Meier plott av overlevelsesfunksjonen til de fire gruppene og ***Aktivitet***

Fra dette plottet var det kvinner som falt inn under kategorien svært lite aktiv som hadde minst sannsynlighet for å overleve utover tiden t. Kategoriene moderat aktiv og svært aktiv hadde nesten lik sannsynlighet noe som betyr at man kunne egentlig slå sammen disse to kategoriene. Det var ingen tydelige mønster som skilte seg ut, altså det var ikke større spredning blant overlevelsesfunksjonene for de med høyere utdanning enn de med ungdomskole utdanning. Det var heller ikke større spredning for de som var under 45 år enn de som var over 45 år, altså var det ingen interaksjoner mellom ***Alder*** og ***Aktivitet*** og ***SKOLE*** og ***Aktivitet***.

Man kan også gruppere inn etter de som hadde født barn og var under 45 år ved spørreundersøkelse 1, de som ikke hadde født barn og var under 45 år ved spørreundersøkelse 1, de som hadde født barn og var over 45 år ved spørreundersøkelse 1 og de som ikke hadde født barn og var over 45 år ved spørreundersøkelse 1, eller gruppere inn i de som var svært aktive og hadde ungdomskole utdanning ved spørreundersøkelse 1, de som var svært lite aktiv og hadde ungdomskoleutdanning, de som var svært aktiv og hadde høyere ved spørreundersøkelse nummer 1 og de som var svært lite aktiv og hadde høyere utdanning ved spørreundersøkelse 1.

Jeg delt også inn fire grupper bestående av de som var svært aktiv og under 45 år ved spørreundersøkelse 1, de som var svært lite aktiv og under 45 år ved spørreundersøkelse 1, de som var svært aktiv og over 45 år ved spørreundersøkelse 1 og de som var svært lite aktiv og over 45 år ved spørreundersøkelse 1.



Figur 5: Kaplan-Meier plott av overlevsesfunksjonen til de fire nye gruppene og ***SmokingHabit***

Plottet, figur 5, viste at de kvinnen som røykte mer enn 15 sigaretterdaglig hadde minst sannsynlighet for å overleve utover tiden t. Plottene viste ingen tegn til interaksjoner mellom ***Aktivitet*** og ***SmokingHbait***, og ***Alder*** og ***SmokingHabit***.

Oppbygging av utvidet Cox regresjonsmodell.

For å lage en utvidet Cox regresjonsmodell for de tidsavhengige variablene kan man bruke en funksjon kalt *coxph()* som ligger i R pakken *survival*. Denne funksjonen regner ut den relative risikoen til den spesifikke kategorien sammenlignet med en referansekategori, referansekategorien velger man selv hva skal være. I R blir den kategoriske variabelen omkodet til være en dummy variabel hvor referansekategorien blir satt til og være 0, mens den andre kategorien blir satt til å være 1. For eksempel

$$SKOLE = x_1 = \begin{cases} 1 & \text{T.O.M ungdosmskole} \\ 0 & \text{Høyere utdanning.} \end{cases}$$

Referansekategoriens hasard raten blir da $e^{\gamma} = e^0 = 1$. Den relative risikoen for en kategorien i forhold til referanse kategorien blir da

$$\frac{h(t|x_1(t))}{h(t|x_1(t))} = \frac{e^{\hat{\gamma}_1 \cdot 1}}{e^{\hat{\gamma}_1 \cdot 0}} = \frac{e^{\hat{\gamma}_1}}{e^0} = e^{\hat{\gamma}_1}$$

Og den relative risikoen for referansekategorien blir

$$\frac{h(t|x_1(t))}{h(t|x_1(t))} = \frac{e^{\hat{\gamma}_1 \cdot 0}}{e^{\hat{\gamma}_1 \cdot 1}} = \frac{e^0}{e^{\hat{\gamma}_1}} = e^{-\hat{\gamma}_1}$$

De ulike referansekategoriene ble satt til å være
Normal/overvektig for ***BMI***
NONsmoker for ***SmokingHabit***
Avhold for ***Alkoforbruk***
Har født for***Barn.fac***
2, god for ***Egehelse***
2, moderat aktiv for ***Aktivitet***
Høyere utdanning for ***SKOLE***

coxph() funksjonen i R gav følgende utskrift.

```
Call:
coxph(formula = Surv(Start, Slutt, Status) ~ BMI + SmokingHabit +
    Barn.fac + Alkoforbruk + Egenhelse + Aktivitet + Alder +
    SKOLE, data = data)

  n= 97343, number of events= 1413
   (42977 observations deleted due to missingness)

                                   coef   exp(coef)  se(coef)   z   Pr(>|z|)
BMIFedme                       0.254981  1.290438  0.155940  1.635 0.102023
BMIUndervektig                 0.704532  2.022900  0.133896  5.262 1.43e-07 ***
SmokingHabit15plussDaily       0.954099  2.596330  0.086862 10.984  < 2e-16 ***
SmokingHabitEXsmoker           0.249304  1.283132  0.073319  3.400 0.000673 ***
SmokingHabitmax14Daily         0.615845  1.851220  0.079654  7.731 1.07e-14 ***
Barn.facIkke_født              0.543815  1.722567  0.095855  5.673 1.40e-08 ***
AlkoforbrukLite_alkoholinntak -0.163821  0.848894  0.088970 -1.841 0.065575 .
AlkoforbrukModerat_alkoholinntak-0.169732  0.843891  0.179017 -0.948 0.343062
AlkoforbrukStort_alkoholinntak 0.433957  1.543352  0.274181  1.583 0.113481
Egenhelse0                     1.687214  5.404402  0.178412  9.457  < 2e-16 ***
Egenhelse1                     0.712381  2.038840  0.079637  8.945  < 2e-16 ***
Egenhelse3                    -0.412877  0.661743  0.068769 -6.004 1.93e-09 ***
Aktivitet1                     0.008826  1.008865  0.078217  0.113 0.910156
Aktivitet3                     0.262152  1.299724  0.073559  3.564 0.000366 ***
Alder                          0.069018  1.071456  0.006106 11.304  < 2e-16 ***
SKOLET.O.M ungdomskole         0.088530  1.092567  0.056212  1.575 0.115275
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                  exp(coef) exp(-coef) lower .95 upper .95
BMIFedme                             1.2904     0.7749    0.9506    1.7518
BMIUndervektig                       2.0229     0.4943    1.5560    2.6299
SmokingHabit15plussDaily             2.5963     0.3852    2.1899    3.0782
SmokingHabitEXsmoker                 1.2831     0.7793    1.1114    1.4814
SmokingHabitmax14Daily               1.8512     0.5402    1.5836    2.1640
Barn.facIkke_født                    1.7226     0.5805    1.4275    2.0786
AlkoforbrukLite_alkoholinntak        0.8489     1.1780    0.7131    1.0106
AlkoforbrukModerat_alkoholinntak     0.8439     1.1850    0.5942    1.1986
AlkoforbrukStort_alkoholinntak       1.5434     0.6479    0.9017    2.6415
Egenhelse0                           5.4044     0.1850    3.8096    7.6668
Egenhelse1                           2.0388     0.4905    1.7442    2.3833
Egenhelse3                           0.6617     1.5112    0.5783    0.7572
Aktivitet1                           1.0089     0.9912    0.8655    1.1760
Aktivitet3                           1.2997     0.7694    1.1252    1.5013
Alder                                1.0715     0.9333    1.0587    1.0844
SKOLET.O.M ungdomskole               1.0926     0.9153    0.9786    1.2198

Concordance= 0.679  (se = 0.008 )
Rsquare= 0.007   (max possible= 0.258 )
Likelihood ratio test= 684  on 16 df,   p=0
Wald test            = 811.8  on 16 df,   p=0
Score (logrank) test = 935.9  on 16 df,   p=0
```


Figur 6: Utskrift fra coxph() funskjonen i R.

I utskriften fra R, i øverste halvdel, er første kolonne navnet på variabelen og kategorien. Kolonne nummer to viser hva den estimerte $\hat{\gamma}$ koeffisienten er. For eksempel var $\hat{\gamma}$ koeffisienten for *SmokingHabit* og kategorien *15plussDaily* lik $\hat{\gamma}_3 = 0.954099$ og for *Alder* er den $\hat{\gamma}_{15} = 0.069018$. Kolonne nummer 3 viser den relative risikoen sammenlignet med referanse kategorien, desto større relativ risiko desto større er risikoen for at hendelsen skal inntreffe. Kolonne nummer 4 viser stander feilen til $\hat{\gamma}_k$ koeffisienten, mens kolonne nummer 5 er Z Wald test statistikken, beregnes her ved $\frac{\hat{\gamma}_k}{se(\hat{\gamma}_k)}$, og kolonne nummer 6 er dens p-verdi.
Nederste halvdel sammenlignes referansekategorien mot de resterende kategoriene i variablene. For eksempel var koeffisienten for kategorien *BMIFedme* lik $\hat{\gamma}_k = 0.24981$. Den relative risikoen for kategorien *Normal/overvektig* sammenlignet med *BMIFedme* ble da $e^{-ga\hat{m}ma_k} = e^{-0.254981} = 0.7749$. Kan dermed konkludere med at det var mindre risiko for at hendelsen skulle inntreffe hvis man var i kategorien *Normal/undervektig*, 0.7749, enn hvis man var i kategorien *Fedme*, 1.29043. Også nedre og øvre 95% konfidensintervall for relative risikoen til de ulik variablene og kategoriene av variablene er oppgitt i den nedre halvdelen. Det nedre konfidensintervallet er gitt ved $e^{\hat{\gamma}_k - 1.96 \cdot se(\hat{\gamma}_k)}$, og det øvre konfidensintervallet er gitt ved $e^{\hat{\gamma}_k + 1.96 \cdot se(\hat{\gamma}_k)}$.

Fra utskriften kan man se at kategorien *Fedme* fra variabelen ***BMI*** ikke var signifikant, p-verdi=0.10 som var større enn et signifikansnivå på 0.05. Kategorien *1*, svært aktiv fra variabelen ***Aktivitet*** og kategorien *T.O.M ungdomskole* fra variabelen ***SKOLE*** og alle kategoriene i variabelen ***Alkoforbruk*** var heller ikke signifikante, da disse har p-verdien som var større enn signifikansnivået på 0.05. For de variabelen som kun hadde enkelte kategorier som ikke var signifikant valgte jeg å slå disse kategoriene sammen noen av de andre kategoriene. Kategorien *Fedme* ble slått sammen med med kategorien *Normal/overvektig* og kategorien *1* svært aktiv ble slått sammen med kategorien *2* moderat aktiv. Siden variabelen***Alkoforbruk*** og ***SKOLE*** ikke var signifikante ble disse variabelen ikke tatt med i den endelige utvidede Cox regresjonsmodellen, men før man kunne gjør det måtte man kontrollere om det var noen interaksjoner mellom noen av variablene.
I tillegg til å bruke plott for å se om det var noen interaksjoner valgte jeg også å bruke *coxph()* funksjonen i R til å se om noen av interaksjonene som kanskje var til stede på plottene var signifikante. For sikkerhetskyld prøvde jeg også med alle mulige interaksjoner av variablene og kom fram til at de eneste interaksjonen som var signifikant var interaksjonen mellom ***BMI*** og ***Alder*** og mellom ***BMI*** og ***Aktivitet***.
Når interaksjonsleddene hadde blitt funnet kunne man ta bort de variablene som ikke var signifikante, disse hadde ingen effekt på overlevelsesevnen. ***SKOLE*** og ***Alkoforbruk*** var ikke signifikant, de har ingen effekt på overlevelsesevnen til kvinnene og ble derfor ikke tatt med i den endelige modellen. Utskriften viste også at det var to av kategoriene til to variabler som ikke var signifikant, *Fedme* og *1* svært aktiv. Det ble så laget en ny utvidet Cox regresjonsmodell som inneholdt de signifikante interaksjonsleddene, hvor de ikke signifikante variablene ikke var tatt med og hvor to av variabelen hadde fått slått sammen noen av kategoriene sine.

Den endelige utvidede Cox regresjonsmodellen hadde følgende utskrift for variabelen i R.

```
Call:
coxph(formula = Surv(Start, Slutt, Status) ~ BMI + SmokingHabit +
    Barn.fac + Egenhelse + Alder + Aktivitet + BMI:Alder + BMI:Aktivitet,
    data = data)

  n= 114619, number of events= 1689
   (25701 observations deleted due to missingness)

                             coef exp(coef)  se(coef)      z Pr(>|z|)
BMIUndervektig            -1.824336  0.161325  0.863625 -2.112  0.03465 *
SmokingHabitEXsmoker       0.190058  1.209319  0.065571  2.899  0.00375 **
SmokingHabit15plussDaily   0.934876  2.546899  0.076036 12.295  < 2e-16 ***
SmokingHabitmax14Daily     0.554906  1.741776  0.070621  7.857 3.89e-15 ***
Barn.facIkke_født          0.497470  1.644556  0.088439  5.625 1.85e-08 ***
Egenhelse0                 1.742519  5.711713  0.151331 11.515  < 2e-16 ***
Egenhelse1                 0.711010  2.036046  0.071389  9.960  < 2e-16 ***
Egenhelse3                -0.395507  0.673338  0.062463 -6.332 2.42e-10 ***
Alder                      0.066850  1.069135  0.005588 11.964  < 2e-16 ***
Aktivitet3                 0.302811  1.353658  0.066253  4.571 4.86e-06 ***
BMIUndervektig:Alder       0.045183  1.046219  0.016964  2.664  0.00773 **
BMIUndervektig:Aktivitet3  0.728565  2.072106  0.250774  2.905  0.00367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                          exp(coef) exp(-coef) lower .95 upper .95
BMIUndervektig               0.1613     6.1987   0.02969    0.8766
SmokingHabitEXsmoker         1.2093     0.8269   1.06347    1.3752
SmokingHabit15plussDaily     2.5469     0.3926   2.19427    2.9562
SmokingHabitmax14Daily       1.7418     0.5741   1.51663    2.0003
Barn.facIkke_født            1.6446     0.6081   1.38283    1.9558
Egenhelse0                   5.7117     0.1751   4.24574    7.6839
Egenhelse1                   2.0360     0.4911   1.77019    2.3418
Egenhelse3                   0.6733     1.4851   0.59575    0.7610
Alder                        1.0691     0.9353   1.05749    1.0809
Aktivitet3                   1.3537     0.7387   1.18882    1.5414
BMIUndervektig:Alder         1.0462     0.9558   1.01201    1.0816
BMIUndervektig:Aktivitet3    2.0721     0.4826   1.26751    3.3874

Concordance= 0.675  (se = 0.007 )
Rsquare= 0.007   (max possible= 0.265 )
Likelihood ratio test= 825.7  on 12 df,   p=0
Wald test            = 1039  on 12 df,   p=0
Score (logrank) test = 1258  on 12 df,   p=0
```

Figur 7: R utskrift av endelig utvidet Cox regresjonsmodell.

Utskriften fra R, figur 7, viste at alle forklaringsvariablene var signifikante da alle hadde en p-verdi under signifikansnivået 0.05.
Ut i fra KI til forklaringsvariablene kunne man se at de fleste hadde korte KI, noe som betydde at estimering av relativ risiko var til å stole på. Forklaringsvariabelen *Egenhelse0* og interaksjosnleddet *BMIUndervektig:Aktivitet3* hadde litt større KI en de andre forklaringsvariablene.
Alle KI inneholdt ikke verdien 1, noe som betydde at alle variablene hadde en effekt på overlevelsesevnen til kvinnene.
Modellen som helhet var også signifikant da alle de tre testene, likelihood rate testen, Wald testen og Score testen, hadde en p-verdi som var mindre enn signifikansnivået på 0.05.

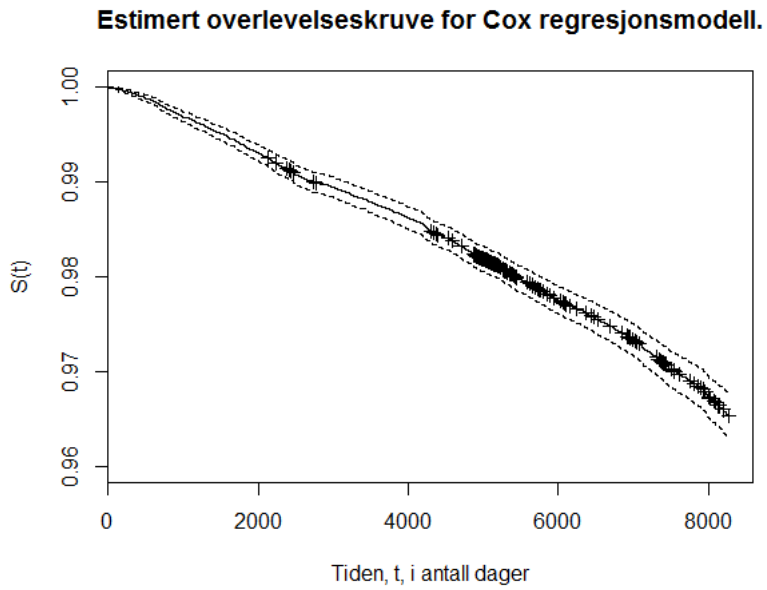
Siden forklaringsvariablene var avhengig av tid, betydde dette at antagelsen om proporsjonalitet ikke holdt, dette kunne bekreftes ved brukt av *cox.zph()* funksjonen på den utvidede Cox regresjonsmodellen, figur 8.

	rho	chisq	p
BMIUndervektig	0.0154	0.388	0.533237
SmokingHabit15plussDaily	0.0884	13.387	0.000253
SmokingHabitEXsmoker	-0.0131	0.291	0.589449
SmokingHabitmax14Daily	0.0521	4.630	0.031410
Barn.facIkke_født	-0.0123	0.255	0.613348
Egenhelse0	-0.0132	0.285	0.593174
Egenhelse1	-0.0643	6.910	0.008572
Egenhelse3	-0.0276	1.280	0.257827
Alder	0.0261	1.185	0.276407
Aktivitet3	-0.0214	0.750	0.386504
BMIUndervektig:Alder	-0.0170	0.466	0.495045
BMIUndervektig:Aktivitet3	0.0196	0.642	0.422835
GLOBAL	NA	33.102	0.000933

Figur 8: Test av proporsjonalitets antagelse for en tidsavhengig Cox hasard regresjonsmodell.

Testen viste at antagelsen om proporsjonalitet ikke holdt for variabelen ***SmokingHabit*** med kategorien *15plussDaily* og kategorien *max14Daily*, antagelsen holdt heller ikke for variabelen ***Egenhelse*** med kategorien *1*, dårlig, da deres p-verdi var mindre enn signifikansnivået på 0.05.
I tillegg viste testen at antagelsen om proporsjonalitet ikke holdt for hele modellen, men dette resultatet var som forventet da forklaringsvariabelen var tidsavhengig.

Den estimerte overlevelseskurven til den utvidede Cox regresjonsmodellen var



Figur 9: Estimert overlevelseskurve for den endelige utvidede Cox regresjonsmodellen med konfidensbånd.

3.7 Resultater.

Ut i fra analysen ble Cox hasard regresjonsmodellen med tidsavhengige forklaringsvariabler lik

$$h(t|\mathbf{X}(t) = h_0(t)exp[\hat{\gamma}_1 x_1(t) + \hat{\gamma}_2 x_2(t) + \hat{\gamma}_3 x_3(t) + \hat{\gamma}_4 x_4(t) + \hat{\gamma}_5 x_5(t) + \hat{\gamma}_6 x_6(t) + \hat{\gamma}_7 x_7(t) \quad (68)$$

$$+\hat{\gamma}_8 x_8(t) + \hat{\gamma}_9 x_9(t)\hat{\gamma}_{10} x_{10} + \hat{\gamma}_{11} x_1(t) x_9(t) + \hat{\gamma}_{12} x_1(t) x_{10}(t)] \quad (69)$$

$$h(t|\mathbf{X}(t)) = h_0(t) e^{\sum_{k=1}^{10} \hat{\gamma}_k x_k(t) + \hat{\gamma}_{11} x_1(t) x_9(t) + \hat{\gamma}_{12} x_1(t) x_{10}(t)} \quad (70)$$

Hvis man lot j være en bestemt kvinne slik at $j = 1, 2, ..., 57555$, så ble hasard raten til kvinne j, gitt at hun hadde en bestemt vektor bestående av tidsavhengige forklaringsvariabler, $\mathbf{X}_j(t) = [x_{j1}(t), x_{j2}(t), ..., x_{j10}(t)]^T$, lik

$$h(t|\mathbf{X}_j(t)) = h_0(t) e^{\sum_{k=1}^{10} \hat{\gamma}_k x_{jk}(t) + \hat{\gamma}_{11} x_{j1}(t) x_{j9}(t) + \hat{\gamma}_{12} x_{j1}(t) x_{j10}(t)} \quad (71)$$

Den relative risikoen ved et fast tidspunkt, t, for en kvinne med en vektor av forklaringsvariabler $\mathbf{X}^*(t)$ sammenlignet med en kvinne med en vektor av forklaringsvariabler $\mathbf{X}(t)$ ble da

$$\frac{h(t|\mathbf{X}^*(t))}{h(t|\mathbf{X}(t))} = \frac{h_0(t) e^{\sum_{k=1}^{10} \hat{\gamma}_k x_k^*(t) + \hat{\gamma}_{11} x_1^*(t) x_9^*(t) + \hat{\gamma}_{12} x_1^*(t) x_{10}^*(t)}}{h_0(t) e^{\sum_{k=1}^{10} \hat{\gamma}_k x_k(t) + \hat{\gamma}_{11} x_1(t) x_9(t) + \hat{\gamma}_{12} x_1(t) x_{10}(t)}} \quad (72)$$

$$= e^{\sum_{k=1}^{10} \hat{\gamma}_k x_k^*(t) + \hat{\gamma}_{11} x_1^*(t) x_9^*(t) + \hat{\gamma}_{12} x_1^*(t) x_{10}^*(t) - \hat{\gamma}_k x_k(t) + \hat{\gamma}_{11} x_1(t) x_9(t) + \hat{\gamma}_{12} x_1(t) x_{10}(t)} \quad (73)$$

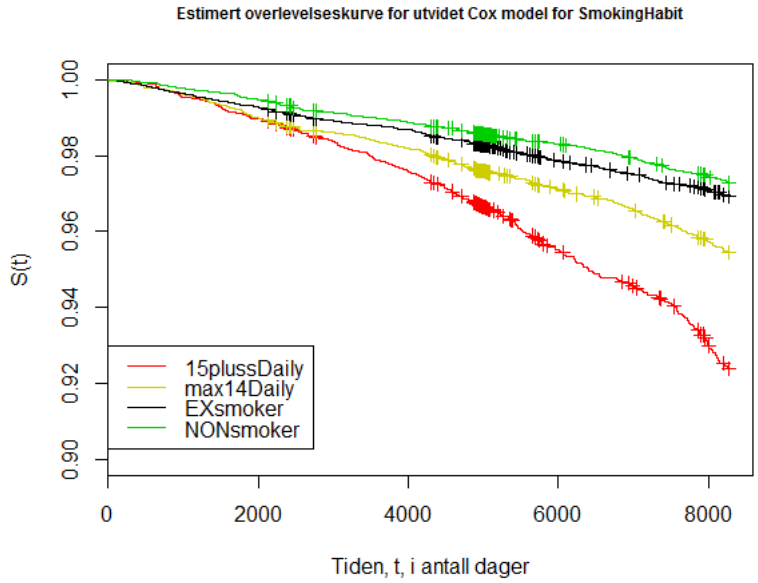
Ut i fra figur 7 har jeg oppsummert de ulike koeffisientene, de relativ risikoene, KI og p-verdiene for den tilpassede utvidede Cox regresjonsmodellen i følgende tabell. (Ref betegner referanse kategorien til variabelen.)

Variable	$x_k(t)$	$\hat{\gamma}_k$	$e^{\hat{\gamma}_k}$	KI	P-verdi
BMI, Undervektig	$x_1(t) = 1$	$\hat{\gamma}_1 = -1.824336$	0.161325	$[0.030, 0.877]$	0.035
Ref: BMI,øvrige	$x_1(t) = 0$	$\hat{\gamma}_1 = -1.824336$	6.19867773		
SmokingHabit, 15plussDaily	$x_2(t) = 1$	$\hat{\gamma}_2 = 0.934876$	2.546899	$[2.194, 2.965]$	$< 2 \cdot 10^{-16}$
Ref: SmokingHabit,NONsmoker	$x_2(t) = 0$	$\hat{\gamma}_2 = 0.934876$	0.3926		
SmokingHabit, max14Daily	$x_3(t) = 1$	$\hat{\gamma}_3 = 0.554906$	1.741776	$[1.517, 2.000]$	$3.89 \cdot 10^{-15}$
Ref: SmokingHabit,NONsmoker	$x_3(t) = 0$	$\hat{\gamma}_3 = 0.554906$	0.5741		
Smokinghabit, EXsmoker	$x_4(t) = 1$	$\hat{\gamma}_4 = 0.190058$	1.209319	$[1.0634, 1.375]$	0.004
Ref: SmokingHabit,NONsmoker	$x_4(t) = 0$	$\hat{\gamma}_4 = 0.190058$	0.8269		
Barn.fac, Ikke_født	$x_5(t) = 1$	$\hat{\gamma}_5 = 0.497470$	1.644556	$[1.383, 1.956]$	$1.85 \cdot 10^{-8}$
Ref: Barn.fac, Har født	$x_5(t) = 0$	$\hat{\gamma}_5 = 0.497470$	0.6081		
Egenhelse, 0	$x_6(t) = 1$	$\hat{\gamma}_6 = 1.742519$	5.711713	$[4.246, 7.684]$	$< 2 \cdot 10^{-16}$
Ref: Egenhelse, 2	$x_6(t) = 0$	$\hat{\gamma}_6 = 1.742519$	0.1751		
Egenhelse, 1	$x_7(t) = 1$	$\hat{\gamma}_7 = 0.711010$	2.036046	$[1.77, 2.342]$	$< 2 \cdot 10^{-16}$
Ref: Egenhelse, 2	$x_7(t) = 0$	$\hat{\gamma}_7 = 0.711010$	0.4911		
Egenhelse, 3	$x_8(t) = 1$	$\hat{\gamma}_8 = -0.395507$	0.67338	$[0.596, 0.761]$	$2.42 \cdot 10^{-10}$
Ref: Egenhelse, 2	$x_8(t) = 0$	$\hat{\gamma}_8 = -0.395507$	1.4851		
Alder	$x_9(t)$	$\hat{\gamma}_9 = 0.066850$	1.069135	$[1.057, 1.081]$	$< 2 \cdot 10^{-16}$
Aktivitet, 3	$x_{10}(t) = 1$	$\hat{\gamma}_{10} = 0.302811$	1.353658	$[1.189, 1.541]$	$4.86 \cdot 10^{-6}$
Ref: Aktivitet, øvrige	$x_{10}(t) = 0$	$\hat{\gamma}_{10} = 0.302811$	0.7387		
BMI, Undervektig:Alder	$x_1(t)x_9(t)$	$\hat{\gamma}_{11} = 0.045183$	1.046219	$[1.012, 1.082]$	0.007
BMI, Undervektig:Aktivitet, 3	$x_1(t)x_{10}(t)$	$\hat{\gamma}_{12} = 0.728565$	2.072106	$[1.268, 3.387]$	0.004

Tabell 5: Resultater fra den tilpassede Cox hasard regresjonsmodellen. (Ref står for referansekategorien)

Ved bruk at figur 7 og tabell 5 kunne man se at kvinner som røykte mer enn 15 sigaretter daglig hadde større risiko for å dø enn kvinner som aldri hadde røkt. Faktisk var risikoen for å dø for en kvinne som røykte mer enn 15 sigaretter daglig 2.54 ganger høyere enn for kvinner som aldri hadde røkt. Kvinner som røykte mindre enn 14 sigaretter daglig og kvinner som hadde røkt tidligere hadde også større risiko for å dø sammenlignet med en kvinne som aldri hadde røkt.

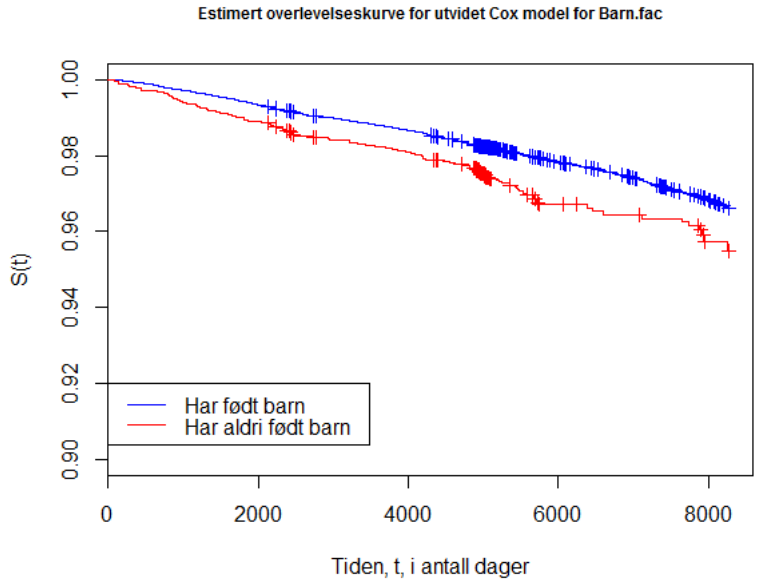
Ved å plotte den estimerte overlevelseskurven basert på røykevanene, se figur 10, viste det seg at kvinner som røykte mer enn 15 sigaretter daglig hadde minst sannsynlighet for å overleve. Kvinner som aldri hadde røkt hadde størst sannsynlighet for å overleve etterfulgt av kvinner som hadde røkt tidligere.



Figur 10: Estimert overlevelseskruve for tilpasset Cox regresjonsmodell basert på røykevaner.

Kvinner som aldri hadde født barn hadde 1.64 ganger større risiko for å dø enn kvinner som hadde født barn.

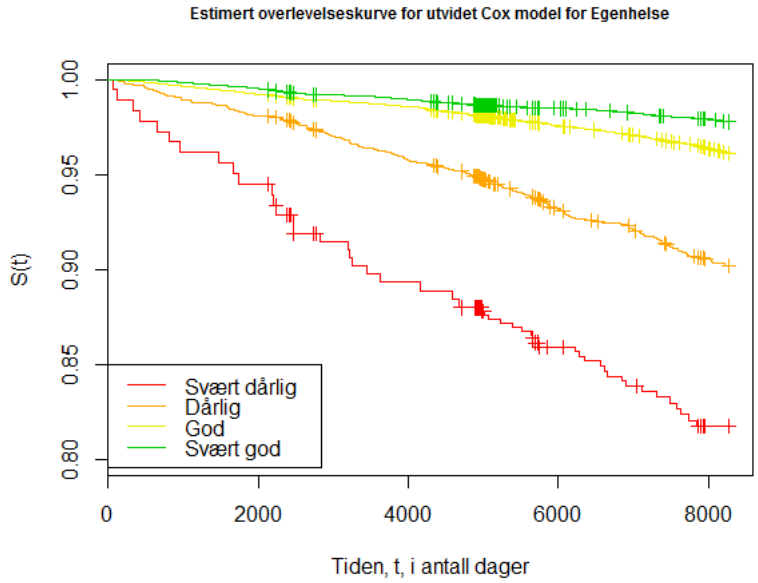
Den estimerte overlevelseskurven til den tilpassede Cox regresjonsmodellen basert på om kvinnen hadde født barn eller ikke, figur 11, viste at kvinner som hadde født barn hadde større sannsynlighet for å overleve enn kvinner som aldri hadde født barn.



Figur 11: Estimert overlevelseskruve for tilpasset Cox regresjonsmodell for fødsler.

Kvinner som hadde rangert sin egen helse til å være meget dårlig hadde større risiko for å dø enn kvinner som hadde god egen helse, faktisk hadde disse kvinnene hele 5.7 ganger større risiko for å dø enn kvinner som hadde en god egen helse. Kvinner som hadde en dårlig egen helse hadde også større risiko for å dø enn kvinner som hadde en god egen helse. Kvinner som hadde en svært god egen helse hadde mindre risiko for å dø enn kvinner som har en god egen helse, det viste seg nemlig at kvinner som hadde god egen helse hadde 1.5 ganger større risiko for å dø enn kvinner som hadde svært god egen helse, men hvis man sammenlignet de kvinne med svært god egen helse med de kvinne som hadde god egen helse hadde de som hadde svært god egen helse bare 0.8 ganger større risiko for å dø enn kvinner som hadde god egen helse.

Den estimerte overlevelseskurven til den tilpassede Cox regresjonsmodellen basert på egen helse, se figur 12, viste at kvinner som hadde svært god egen helse hadde minst sannsynlighet for å dø. Størst sannsynlighet for å dø hadde kvinner som hadde svært dårlig egen helse.



Figur 12: Estimert overlevelseskruve for tilpasset Cox regresjonsmodell for egenhelse.

Den relative risikoen for en kvinne som var undervektig sammenlignet med en kvinne som ikke var undervektig ble

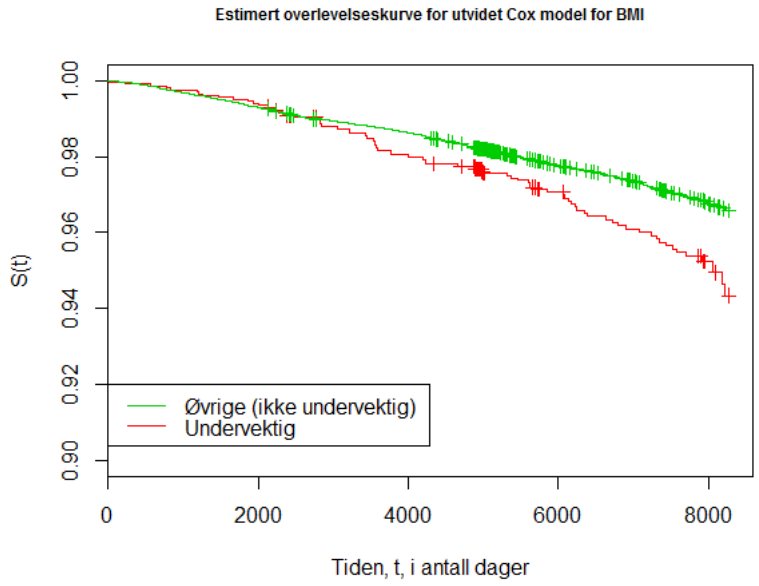
$$BMI = x_1(t) = \begin{cases} 1 & \text{Undervektig} \\ 0 & \text{Annet.} \end{cases}$$

$$\begin{aligned} \frac{h(t|\mathbf{X}^*(t))}{h(t|\mathbf{X}(t))} &= \frac{e^{\hat{\gamma}_1 x_1(t) + \hat{\gamma}_9 x_9(t) + \hat{\gamma}_{11} x_1(t) x_9(t)}}{e^{\hat{\gamma}_1 x_1(t) + \hat{\gamma}_9 x_9(t) + \hat{\gamma}_{11} x_1(t) x_9(t)}} \\ &= \frac{e^{\hat{\gamma}_1 + \hat{\gamma}_9 x_9(t) + \hat{\gamma}_{11} x_9(t)}}{e^{\hat{\gamma}_9 x_9(t)}} \\ &= e^{\hat{\gamma}_1 + \hat{\gamma}_{11} x_9(t)} \\ &= e^{-1.824 + 0.045 x_9(t)} \end{aligned}$$

Den relative risikoen for BMI var dermed avhengig av alderen til kvinnen.
Hvis begge kvinnen var 35 år ble den relative risikoen for å dø for en kvinne som var undervektig sammenlignet med en kvinne som ikke var undervektig lik 0.0566
Altså var risikoen for å dø 0.0566 ganger større for en kvinne som var undervektig, enn for en kvinne som ikke var undervektig.
Den relative risikoen for å dø for en kvinne som var undervektig og i 50 årene sammenlignet med en kvinne som ikke var undervektig og i 50 årene ble 1.531 ganger større enn for en kvinne som ikke var undervektig og 50 år.
Ved å se på den relative risikoen for å dø for en kvinne som ikke var undervektig sammenlignet med en kvinne som var undervektig og hvor begge kvinnen var 35 år fikk man en relativ risiko lik

1.27. For en kvinne som var i 50 årene og som ikke var undervektig var den relative risikoen 0.65 ganger større enn for en kvinne som var undervektig og 50 år.
Dette betyr altså at risikoen for å dø når man var undervektig økte med alderen. Det betyr også at unge undervektige hadde mindre risikoen for å dø enn unge kvinner som ikke var undervektige, men etter som disse blir eldre ble deres risiko for å dø større enn for de kvinne som ikke var undervektig og på samme alder. Ved nærmere ettersyn viste det seg at kvinner under 40 år som var undervektig hadde mindre risiko for å dø enn kvinner som ikke var undervektig og under 40 år, men denne trenden endret seg så snart kvinne ble elder enn 40 år.

Den estimerte overlevelseskurven for den tilpassede Cox regresjonsmodellen basert på BMI, se figur 13, viste at til å begynne hadde de kvinne som var undervektig størst sannsynlighet for å overleve, men etter en tid viste det seg at det var disse kvinne som hadde minst sannsynlighet for å overleve.



Figur 13: Estimert overlevelseskruve for tilpasset Cox regresjonsmodell for BMI.

Den relative risikoen for en kvinne som var undervektig og svært lite aktiv sammenlignet med en kvinne som var undervektig og ikke svært lite aktiv, og hvor begge kvinnen var 35 år ble

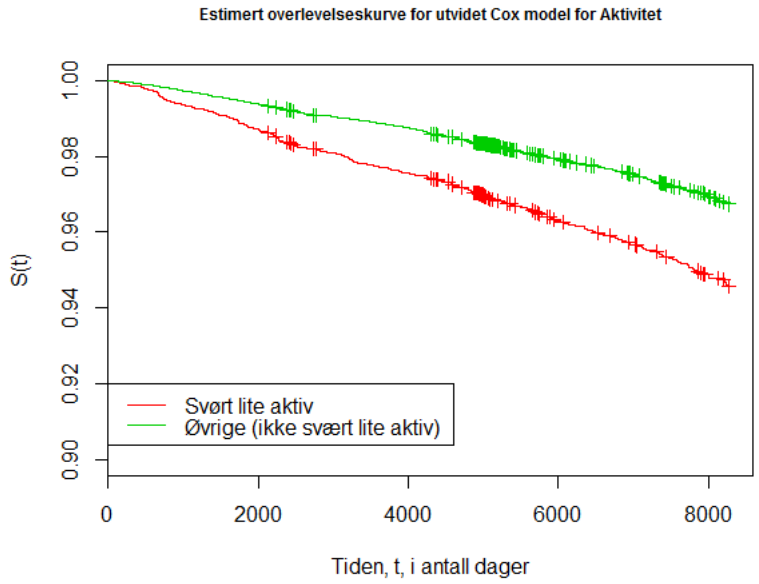
$$BMI = x_1(t) = \begin{cases} 1 & \text{Undervektig} \\ 0 & \text{Annet.} \end{cases}$$

$$Aktivitet = x_{10}(t) = \begin{cases} 1 & \text{3, svært lite aktiv} \\ 0 & \text{Annet.} \end{cases}$$

$$
\begin{aligned}
\frac{h(t|\mathbf{X}^*(t))}{h(t|\mathbf{X}(t))} &= \frac{e^{\hat{\gamma}_1 x_1(t)+\hat{\gamma}_9 x_9(t)+\hat{\gamma}_{10} x_{10}(t)+\hat{\gamma}_{11} x_1(t) x_9(t)+\hat{\gamma}_{12} x_1(t) x_{10}(t)}}{e^{\hat{\gamma}_1 x_1(t)+\hat{\gamma}_9 x_9(t)+\hat{\gamma}_{10} x_{10}(t)+\hat{\gamma}_{11} x_1(t) x_9(t)+\hat{\gamma}_{12} x_1(t) x_{10}(t)}} \\
&= \frac{e^{\hat{\gamma}_1+\hat{\gamma}_9 x_9(t)+\hat{\gamma}_{10}+\hat{\gamma}_{11} x_9(t)+\hat{\gamma}_{12}}}{e^{\hat{\gamma}_1+\hat{\gamma}_9 x_9(t)+\hat{\gamma}_{11} x_9(t)}} \\
&= e^{\hat{\gamma}_{10}+\hat{\gamma}_{12}} \\
&= e^{0.303+0.729} \\
&= 1.23
\end{aligned}
$$

Den relative risikoen for en kvinne som var undervektig og svært lite aktiv sammenlignet med en kvinne som var undervektig og som ikke var svært lite aktiv ble da 1.23 ganger større.
Den relative risikoen for å dø for en kvinne som var undervektig og svært lite aktiv sammenlignet med en kvinne som ikke var undervektig og svært lite aktiv, og hvor begge kvinnen var 35 år er 1.63 ganger større. Altså var det større risiko for at døden skulle inntreffe for en kvinne som var undervektig og svært lite aktiv enn for enn kvinne som ikke var undervektig og svært lite aktiv. Risikoen for å dø for kvinner som var undervektig og svært lite aktiv økte også med alderen.
Den relative risikoen for at døden skulle inntreffe for en kvinne som var undervektig og svært lite aktiv sammenlignet med en kvinne som ikke var undervektig og som ikke var svært lite aktiv, og hvor begge var 35 år, var 2.20 ganger større og denne risikoen økte med alderen.

Den estimerte overlevelseskurven for den tilpassede Cox regresjonsmodellen basert på den fysiske aktiviteten, se figur 14, viste at de kvinnen som var svært lite aktiv hadde minst sannsynlighet for å overleve.



Figur 14: Estimert overlevelseskruve for tilpasset Cox regresjonsmodell for fysisk aktivitet.

Hasard raten til en kvinne som var 35 år, røykte mer en 15 sigaretter daglig, hadde født barn, som ikke var undervektig, som hadde en dårlig egen helse og var svært lite aktiv, ved et gitt tidspunkt t ble

$$\begin{aligned} h(t|\mathbf{X}(t)) &= h_0(t)e^{\hat{\gamma}_2+\hat{\gamma}_6+\hat{\gamma}_9\cdot 35+\hat{\gamma}_{10}} \\ &= h_0(t)e^{0.934876+1.742519+(0.066850\cdot 35)+0.302811} \\ &= h_0(t)\cdot 204.37 \end{aligned}$$

Hasard raten til en kvinne som var 35 år, som aldri hadde røkt, som hadde født barn, som ikke var undervektig, som hadde en svært god egen helse og som ikke var svært aktiv ble

$$\begin{aligned} h(t|\mathbf{X}(t)) = h_0(t)e^{\hat{\gamma}_8+\hat{\gamma}_9\cdot 35} & \\ &= h_0(t)e^{-0.39557+()0.066850\cdot 35} \\ &= h_0(t)\cdot 6.92 \end{aligned}$$

Sammenlignet man kvinne som røykte 15 sigaretter daglig med kvinne som aldri hadde røkt, fikk kvinne som røyker 15 sigaretter daglig en relativ risiko for å dø lik 29.5 ganger større enn for kvinne som aldri hadde røkt, som hadde svært god egen helse og som var svært aktiv.

3.8 Diskusjon.

Den gjennomsnittlige levealderen for kvinner som deltok i spørreundersøkelsene var 56 år, noe som er nokså lavt. Dette skyldes av at studiet ikke følgte alle kvinnen til de døde, kun 2 198 kvinner døde iløpet av studiet or resten ble sensurert.
Når det gjelder feilene i det originale datasette kan disse skyldes at kvinnen kanskje ikke har missforstått spørsmålet. Det kan også være at kvinne har unngått å svare på spørsmål da hun kanskje syntes at noen av spørsmålene ikke passet seg å svare på, for eksempel alkoholinntak. Feilen kan også skyldes av at svarene fra spørreundersøkelsene ikke har blitt lest av riktig.
Å bruke BMI til å kategorisere kvinner etter om de er undervektig, normalvektig eller overvektig kan være noe misvisende, fordi en kvinne som er veltrent vil muligens havne i kategorien overvektig. En bedre måte å måle undervekt, normalvekt og overvekt er å måle fettprosent eller kolesterol, men dette må gjøres av helsepersonell.
Det at kvinnene selv skulle få rangere sin egen helse og sin egen fysiske aktivitet kan føre til at noen kvinner har bedre helse enn de mener, eller at de for eksempel er mindre fysisk aktiv enn det de har gitt uttrykk for i spørreundersøkelsen. Man burde heller har spurt om hvor mye kvinnen trener i løpet av en uke, om hun går, sykler eller kjører til jobb, hvor lenge er det siden hun sist var syk, hvor mange ganger har hun i løpet av det siste året vært syk eller om hun gjorde noe tiltak for å forhindre å bli syk o.s.v.
Årsaken til at kvinner som er undervektige og under 40 år har mindre risiko for å dø enn kvinner som ikke er undervektig og under 40 år er noe vanskelig å fastslå. Noen årsaken til at kvinner er undervektig kan være at hun har en spiseforstyrrelse, at hun er syk, for eksempel med kreft eller at hun har diabetes, eller at hun trener ganske hardt. Kvinner som er undervektig mister ofte menstruasjon, noe som igjen fører til at de ikke kan bli gravide, i tillegg kan de utvikle beinskjørhet. Det er også en stor belastning på hjerte for personer som er undervektig. Det og være i kategorien undervektig kan også føre til at immunsystemet svekkes slik at det er lettere å bli forkjølet.
Det viste seg at kvinner som hadde født barn hadde mindre risiko for å dø enn kvinner som ikke hadde født barn, årsaken til hvorfor kvinner som har født barn lever lengre enn kvinner som ikke ar født barn er noe usikker, men en del studier har vist at kvinner som ammer barn har mindre risiko for å få bryst kreft enn kvinner som ikke har ammet barn.
Det har ikke blitt brukt residual plott til å undersøke om den tilpassede Cox hasard regresjonsmodellen var en god modell for dataene, dette var på grunn av at datasettes var av en slik stor størrelse at det gjorde det vanskelig å tolke residual plottene.

3.9 Konklusjon.

Ut ifra analysen hvor det ble brukt Cox proporsjonal hasard regresjonsmodell har det vist seg at alkoholinntaket og antall år utdannelse ikke hadde noe effekt på overlevelsesevnen til kvinne.
Det vist seg at kvinner som var ung, det vil si kvinner som var under 40 år og undervektig hadde større sjanse for å overleve enn kvinner som var ung og ikke overvektig, men at dersom hun var 40 år eller eldre og undervektig så hadde hun større sjanse for å dø enn kvinner som ikke var undervektig og på samme alder.
Kvinner som hadde født barn hadde mindre sjanse for å dø enn kvinner som aldri hadde født barn.
Hvis kvinnen røyke 15 sigaretter daglig økte dette risikoen for å dø, risikoen økte også for kvin-

ner som var svært lite aktiv og som hadde svært dårlig helse. Over tid vil risikoen for å dø være minst for de kvinnen som aldri hadde røkt, var i god fysisk form, hadde en god helse, ikke var undervektig og som hadde født barn.

4 Appendiks

4.1 Formler.

Score ligningen.

$$U_b(\boldsymbol{\gamma}) = \frac{\partial logL(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_b} \qquad b = 1, ..., p \tag{74}$$

$$= \sum_{i=1}^{D} x_{(i)b}(t) - \sum_{i=1}^{D} \frac{\sum_{j \in R)t_j)} x_{jb}(t) e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}}{\sum_{j \in R(t_j)} e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}} \tag{75}$$

n er her størrelsen på utvalget slik at $j = 1, 2, ..., n$. Hendelsene som inntreffer , inntreffer ikke på samme tidspunkt og består av D forskjellige hendelser. $x_{(i)k}(t)$ er den $k - ende$ forklaringsvariabelen til et individ hvor hendelsen har inntruffet ved tiden t_i. $R(t_i)$ er risikosettet ved tidspunktet t_i og består av alle de individene som enda er med i studiet like før en hendelse inntreffer.

Informasjonsmatrisen.

$$\boldsymbol{I}(\boldsymbol{\gamma}) = [I_{gb}(\boldsymbol{\gamma})]_{p \times p} \tag{76}$$

Element (g, b) er gitt ved

$$I_{gb}(\boldsymbol{\gamma}) = \sum_{i=1}^{D} \frac{\sum_{j \in R(t_i)} x_{jg}(t) x_{jb}(t) e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}} - \sum_{i=1}^{D} \left[\frac{\sum_{j \in R(t_i)} x_{jg}(t) e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}} \right] \left[\frac{\sum_{j \in R(t_i)} x_{jb}(t) e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \gamma_k x_{jk}(t)}} \right] \tag{77}$$

4.1.1 Tabell over variabler.

Variabel navn	Beskrivelse	Enhet	Skjema nr.
AKTIDAG	Hvordan oppfattet kvinnen sin egen fysiske aktivitet.	1,2,...,10 hvor 1 er svært lite og 10 er svært mye	Skjema 1
AKTIDAG.fac	Hvordan oppfattet kvinnen sin egen fysiske aktivitet.	3 = svært lite aktiv, 2 = moderat aktiv og 1 = svær aktiv	Skjema 1
ALKOGR	Gjennomsnittlig inntak av ren alkohol per dag i gram.	Gram alkohol per dag	Skjema 3
ANTBARN	Antall født barn, både dødfødte og barn som har dødd ved en en senere anledning.	Stk	Skjema 1
AVHOLD	Darkk kvinnen i det hele tatt alkohol.	0 = ja og 1 = nei	Skjema 3
DELTA	Har personen dødd eller er hun blitt sensurert.	0 = sensurert og 1 = død	
doddt	Dato for kvinnens bortgang	År-måned-dag	
EGENHELS	Hvordan oppfatter kvinnen sin egen helse.	3 = meget god, 2 = god, 1 = dårlig og 0 = meget dårlig	Skjema 1
emigdt	Dato for kvinnens sensurering	År-måned-dag	
EVERROK	Har kvinnen noen gang røkt.	0 = ja og 1 = nei	Skjema 1
HOYDE	Kvinnens høyde.	Cm	Skjema 1
ROK1	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 10-14 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK2	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 15-19 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK3	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 20-24 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1

ROK4	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 25-29 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK5	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 30-34 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK6	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 35-39 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK7	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 40-44 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
ROK8	Hvis kvinnen noen gang hadde røkt, hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i en alder av 45-49 år	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+stk = 6	Skjema 1
SKOLE	Antall år kvinnen hadde gått på skole inkludert ungdomskole.	År	Skjema 1
SKOLE.fac	Kategorisk variabel som fortalte om kvinnen kun hadde ungdomskole eller om hun også hadde høyere utdanning.	"T.o.m ungdomskole" eller "Høyere utdanning"	Skjema 1
STARTALD	Kvinnens daværende alder.	År	Skjema 1
STARTDAT	Dato for når svarte på spørreundersøkelse nummer 1.	År-Måned-Dag	Skjema 1
svar_skjema	Kategorisk variabel som forteller hvilke spørreundersøkelser kvinnen har latt være å besvare.	Ikke_svart1, Ikke_vart2 og Ikke_svart3	
t1Alkoforbruk.fac	Kategorisk variabel som sier noe om kvinnens alkoholinntak.	Avhold, Lite alkoholinntak, Moderat alkoholinntak og Stort alkoholinntak	Skjema 1
t1BMI	BMI.	Kg per høyde^2	Skjema 1

t1BMI.fac	Kategorisk variabel som sier noe om kvinnens BMI-en.	Undervektig, Normal/overvektig og Fedme	Skjema 1
t1BMIgr	Rangerer BMI-en i intervall.	Intervall [12.6,18.5], (18.5,25], (25,30] og (3,63.4]	Skjema 1
t1SmokingHabit	Kategorisk variabel som foteller om kvinnens røykevaner, denne variablen er laget av Elinor.	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 1
t1SmokingHabit.fac	Kategorsik varaibel som forteller noe om kvinnens røykevaner.	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 1
t2Alkoforbruk.fac	kvinnens alkoholforbruk	Avhold, Lite alkoholinntak, Moderat alkoholinntak og Stort alkoholinntak	Skjema 2
t2BMI	BMI	Kg per høyde^2	Skjema 2
t2BMI.fac	BMI ved ny spørreundersøkelse høsten 1998	Undervektig, Normal/overvektig og Fedme	Skjema 2
t2BMIgr	BMI ved starten av ny spørreundersøkelse høsten 1998	Intervall [12.6,18.5], (18.5,25], (25,30] og (3,63.4]	Skjema 2
t2SmokingHabit	forteller om kvinnen aldri har røkt, er tidligere røker, røyker maks 14 sigaretter om dagen eller om hun røyker mer enn 15 sigaretter om dagen	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 2
t2SmokingHabit.fac	er en rettet versjon av variabelen t2SmokingHabit	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 2
t3Alkoforbruk.fac	kvinnens alkoholforbruk ved ny spørreundersøkelse høsten 2004	Avhold, Lite alkoholinntak, Moderat alkoholinntak og Stort alkoholinntak	Skjema 3
t3BMI	BMI ved ny spørreundersøkelse høsten 2004	Kg per høyde^2	Skjema 3
t3BMI.fac	BMI ved ny spørreundersøkelse ved høsten 2004	Undervektig, Normal/overvektig og Fedme	Skjema 3

t3BMIgr	BMI ved ny spørreundersøkelse høsten 2004	Intervall [12.6,18.5], (18.5,25], (25,30] og (3,63.4]	Skjema 3
t3SmokingHabit	forteller om kvinnen aldri har røkt, er tidligere røker, røyker maks 14 sigaretter om dagen eller om personen røyker mer enn 15 sigaretter om dagen	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 3
t3SmokingHabit.fac	er en rettet versjon av variabelen t3SmokingHabit	NONsmoker, EXsmoker, max14Daily og 15plussDaily	Skjema 3
TID	tiden det tar, i dager, til personen dør eller blir sensurert	Antall dager	
VEKTANA	kvinnens vekt ved start av studie i 1991	Kg	Skjema 1
YAKTIDAG	hvordan oppfattet kvinnen sin egen fysiske aktivitet ved høsten 1998	1,2,...,10 hvor 1 er svært lite og 10 er svært mye	Skjema 2
YALKOGR	hvis nei hvor mye kvinnen drakk i gjennomsnitt per dag i gram rein alkohol	gram alkohol per dag	Skjema 2
yANTBARN	antall fødte barn inntil høsten 1998 (gjelder også dødfødte og barn som døde senere)	Stk	Skjema 2
YAKTIDAG.fac	hvordan oppfattet kvinnen sin egen fysiske aktivitet ved høsten 1998	3 = svært lite aktiv, 2 = moderat aktiv og 1 = svært aktiv	skjema 2
yAVHOLD	var personen avholdskvinne	0 = ja og 1 = nei	Skjema 2
yEGENHELS	hvordan oppfattet kvinnen sin egen helse ved høsten 1998	3 = meget god, 2 = god, 1 = dårlig og 0 = meget dårlig	Skjema 2
yEVERROK	har personen noen gang røkt inntil høsten 1998	0 = ja og 1 = nei	Skjema 2
yHOYDE	kvinnens høyde ved høsten 1998	Cm	Skjema 2
yROYKAR1	hvis ja hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i perioden 1991-1994	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+ stk = 6	Skjema 2

yROYKAR2	hvis ja hvor mange sigaretter røykte kvinnen i gjennomsnitt per dag i perioden 1995-1998	1-4 stk = 1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+ stk = 6	Skjema 2
yROYKNAA	røkte kvinnen daglig i 1998	0 = ja og 1 = nei	Skjema 2
ystartald	kvinnens alder ved ny spørreundersøkelse høsten 1998	År	Skjema 2
ystartdat	dato for når kvinnen entret studiet i 1998	År-Måned-Dag	Skjema 2
yVEKTANA	kvinnens vekt ved ny spørreundersøkelse ved høsten 1998	Kg	Skjema 2
ZAKTIDAG	hvordan oppfattet kvinnen sin egen fysiske aktivitet ved høsten 2004	1,2,...,10 hvor 1 er svært lite og 10 er svært mye	Skjema 3
ZAKTIDAG.fac	hvordan oppfattet kvinnen sin egen fysiske aktivitet ved høsten 2004	3 = svært lite aktiv, 2 = moderat aktiv og 1 = svært aktiv	Skjema 3
ZALKOGR	hvis nei hvor mye kvinnen drakk i gjennomsitt per dag i gram rein alkohol	gram alkohol per dag	Skjema 3
ZANTBARN	antall fødte barn inntil høsten 2004 (gjelder også dødfødte og barn som døde senere)	Stk	Skjema 3
ZAVHOLD	var personen avholdskvinne	0 = ja og = nei	Skjema 3
ZEGENHELS	hvordan oppfatte kvinnen sin egen helse ved høsten 2004	3 = meget god, 2 = god, 1 = dårlig og 0 = meget dårlig	Skjema 3
ZHOYDE	kvinnens høyde ved høsten 2004	Cm	Skjema 3
ZROYKNAA	røkte kvinnen daglig i 2004	0 = ja og 1 = nei	Skjema 3
ZROKSIST5	hvis kvinnen røkte daglig, hvor mange sigaretter røkte hun i gjennomsnitt per dag i de siste fem årene	0 stk = 0, 1-4 stk =1, 5-9 stk = 2, 10-14 stk = 3, 15-19 stk = 4, 20-24 stk = 5 og 25+ stk = 6	Skjema 3
ZROYKSTOP	hvis ZROYKNAA er nei hvor gammel var kvinnen når hun sluttet	År	Skjema 3
ZSIGALDER	hvor gammel var kvinnen da hun tok sin første sigarett	År	Skjema 3

ZSIGROYK	har kvinnen, i løpet av livet, røkt mer enn 100 sigaretter tilsammen	0 = ja og 1 = nei	Skjema 3
zstartald	kvinnens alder ved ny spørreundersøkelse høsten 2004	År	skjema 3
zstartdat	dato for når kvinnen entret studiet i 2004	År-Måned-Dag	Skjema 3
ZVEKTANA	kvinnens vekt ved ny spørreundersøkelse ved høste 2004	Kg	Skjema 3

4.1.2 R funksjoner.

Her er noen av funksjonen i R som ble brukt.

as.factor(x,levels): Denne funksjonen er brukt til å omgjøre en vektor enten bestående av bokstaver eller tall om til en faktor. x er vektoren bestående av bokstaver eller tall og *levels* er navnene på faktor nivåene. Denne har jeg brukt til å lage karakteristiske variabler.

as.Date(x,format): Denne funksjonen omgjør verdiene til dato. $x$er verdien som skal omgjøres til en dato, *format* kan være $\%Y - \%m - \%d$ det vil si År-Måned-Dag, for eksempel 2001-01-15, eller andre typer.

difftime(time1,time2,units=days",...): Denne funksjonen tar to datoer og finner ut hvor mange dager det er i fra den første datoen, *time1* til den andre datoen, *time2*. Denne funksjonen kan også finne antall sekunder, timer og uker det er mellom to gitte datoer. Har brukt denne til å finne ***Start*** og ***Slutt*** variablene for kvinnene.

For følgende funksjoner trenger man å laste ned pakken ***survival*** i R.

Surv(time,time2,event,...): Denne funksjonen lager et såkalt overlevelsesobjekt som kan brukes til å lage utvidet Cox regresjonsmodell. *time* er starte tiden, mens *time2* er slutttiden, eller tiden når hendelsen inntreffer eller tiden for når personen blir sensurert. Tidsuavhengig Cox regresjonsmodell har kun *time*. *Event* er en indikator variabel som er 1 hvis hendelsen har inntruffet og 0 hvis personen enda er i live eller har blitt sensurert.

coxph(formula,ties=c(efron",breslow",exact"),...): Denne funksjonen brukes til å tilpasse en Cox proporsjonal hasard regresjonsmodell til de ulike variablene. *formula* består av et overlevelsesobjekt samt variablene som skal være med i modellen, for eksempel
Surv(time,time2,event) sex+age+BMI,data=tabell *data* er her navnet på matrisen eller tabellen som variabelen er lageret i. *ties* forteller hvilken metode som skal brukes hvis flere hendelser inntreffer på samme tidspunkt. Hvis ingen metode er valgt blir efron metoden brukt til å behandle disse hendelsene.

summary(): Ved å bruke ***summary()*** på en Cox regresjonsmodell vil R vise de forskjellige koeffisienten til variabelen i modellen, de relative risikoene, KI til den relative risikoen, samt statistiske tester og p-verdier for modellen som helhet, men også for hver enkelt variabel.

survfit(formula,...):Denne funksjonen kan enten brukes til å tilpasse en Kaplan-Meier overlevelsesfunksjon til et overlevelsesobjekt av typen***Surv() ...*** eller finne den estimerte overlevlsesfunksjonen til Cox regresjonsmodell av typen ***coxph()***. Ved å bruke funksjonen ***plot()*** kan man plotte den estimerte overlevelsesfunksjonen til modellene.

cox.zph(): Denne funksjonen tester antagelsen om proporsjonalitet ved å bruke ***coxph()*** funksjonen. Hvis p-verdien er større enn et signifikansnivå på 0.05 betyr det at antagelsen om proporsjonalitet holder. Denne funksjonen tester antagelsen for hver enkelt variabel, men også modellen som helhet. Ved å bruke ***plot(cox.zph())*** kan også antagelsen om proporsjonalitet sees grafisk.

Kryssende grafer er et tegn på at antagelsen om proporsjonalitet ikke holder.

residuals(object,type=c("martingale",deviance", score",schoenfeld",...),collapse=False)): Denne funksjonen finner ulike typer residualer til Cox regresjonsmodellen. *object* er her av typen ***coxph()***, mens *collapse* kan summere over radene, det vil si hvis det er flere rader med data for individ med id nummer 1 summerer den sammen radene for individ nummer 1. Ved å bruke *plot(residuals())* kan man plotte de ulike typene av residualer for modellen.

AIC(object): er en funksjon som finner AIC verdien til, for eksempel, Cox regresjonsmodellen. *object* kan for eksempel være av type ***coxph()***

4.1.3 R-koder.

Har kun tatt med en liten del av R-koden min, dette er fordi R-kodene strekker seg ut over flere filer og ville tatt opp ganske mye plass hvis alt ble tatt med i oppgaven.

```
#Leser inn datasettet som består av tidsavhengige variabler.
setwd("C:/Users/Irmelin/Documents/Mastergradsoppgave")
data=read.delim("Analyse.txt",header=T)


#Laster inn pakken "survival" som trengs for coxph() funskjonen og splines som
#trengs for å få Kaplan-Meier estimat og plott.
library(survival)
library(splines)



#Omgjør noen av variablene til faktorer. Av en ukjent grunn ville ikke R
#godta de som faktorer i txt filen.
data$Aktivitet=factor(data$Aktivitet)
data$Egenhelse=factor(data$Egenhelse)


#Endrer referanse variablen.
data$BMI=relevel(data$BMI,ref="Normal/overvektig")

data$SmokingHabit=relevel(data$SmokingHabit,ref="NONsmoker")

data$Alkoforbruk=relevel(data$Alkoforbruk,ref="Avhold")

data$Egenhelse=relevel(data$Egenhelse,ref="2")

data$Aktivitet=relevel(data$Aktivitet,ref="2")

data$Barn.fac=relevel(data$Barn.fac,ref="Har født")



#Koder om varaiblene aktivitet og BMI til å bestå av mindre nivåer.
#BMI faktor med to nivåer, undervektig og annet.
data$BMI=as.character(data$BMI)
data$BMI[which(data$BMI=="Fedme")]="Annet"
data$BMI[which(data$BMI=="Normal/overvektig")]="Annet"

data$BMI=as.factor(data$BMI)
```

```
#Aktivitet faktoren kodes til to nivåer, svært lite aktiv og annet.
data$Aktivitet=as.character(data$Aktivitet)
data$Aktivitet[which(data$Aktivitet=="1")]="2"
data$Aktivitet=as.factor(data$Aktivitet)

data$Egenhelse=as.factor(data$Egenhelse)



#Tilpassing av tidsavhengig Cox hasard regresjonsmodell med interaksjoner og uten
#forkalringsvaraiblene Alkoforburk og SKOLE.
fit=coxph(Surv(Start,Slutt,Status)~BMI+SmokingHabit+Barn.fac+Egenhelse+
                 Alder+Aktivitet+BMI:Alder+BMI:Aktivitet
               ,data=data)
summary(fit)



#AIC verdien til den tilpassede Cox regresjonsmodellen.
AIC(fit)



#Test av antagelsen om proporsjonalitet for den utvidede Cox regresjonsmodellen.
cox.zph(fit)



#Plott av den estimerte overlevelseskruven tid den utvidede Cox regresjonsmodellen.
plot(survfit(fit),ylim=c(0.96,1),
     main="Estimert overlevelseskruve for Cox regresjonsmodell.",
     xlab="Tiden, t, i antall dager", ylab="S(t)")
```

Bibliografi

Aalen, O., Borgan, O. & Gjessing, H. (2008), *Survival and event history analysis: a process point of view*, Springer Science & Business Media.

Allison, P. D. (2010), *Survival analysis using SAS: a practical guide*, Sas Institute.

Balakrishnan, N. (2010), *Methods and applications of statistics in the life and health sciences*, John Wiley & Sons.

Bland, M. (2000), *An introduction to medical statistics*, Oxford University Press.

Box-Steffensmeier, J. M. & Jones, B. S. (2004), *Event history modeling: A guide for social scientists*, Cambridge University Press.

Broström, G. (2012), *Event history analysis with R*, CRC Press.

Cleves, M. (2008), *An introduction to survival analysis using Stata*, Stata Press.

Collett, D. (2015), *Modelling survival data in medical research*, CRC press.

Crawley, M. J. (2012), *The R book*, John Wiley & Sons.

Dalgaard, P. (2008), *Introductory statistics with R*, Springer Science & Business Media.

Fan, J. & Jiang, J. (2009), 'Non-and semi-parametric modeling in survival analysis'.

Fox, J. & Sanford, W. (2011), 'Cox proportional-hazards regression for survival data in r'. **URL:** *http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Cox-Regression.pdf*

Gallin, J. I. & Ognibene, F. P. (2012), *Principles and practice of clinical research*, Academic Press.

Geisser, S. & Johnson, W. O. (2006), *Modes of parametric statistical inference*, Vol. 529, John Wiley & Sons.

Hothorn, T. & Everitt, B. S. (2014), *A handbook of statistical analyses using R*, CRC press.

Klein, J. P. & Moeschberger, M. L. (2003), *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media.

Liu, X. (2012), *Survival analysis: models and applications*, John Wiley & Sons.

Marubini, E. & Valsecchi, M. G. (2004), *Analysing survival data from clinical trials and observational studies*, Vol. 15, John Wiley & Sons.

Muche, R. (2001), 'Applied survival analysis: Regression modeling of time to event data. dw hosmer, jr., s lemeshow. new york: John wiley,', *International Journal of Epidemiology* **30**(2).

Therneau, T. M. & Grambsch, P. M. (2000), *Modeling survival data: extending the Cox model*, Springer Science & Business Media.

Thomas, L. & Reyes, E. M. (n.d.), 'Tutorial: survival estimation for cox regression models with time-varying coefficients using sas and r'.

Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. (1993), *Probability and statistics for engineers and scientists*, Vol. 5, Macmillan New York.