
**Matematisk statistik**
**Stockholms universitet**

# En översikt över Cox's proportionella hazard model för inbäddad fall-kontroll studie

Mathias Lindholm

Examensarbete 2004:4

## Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige


## Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2004:4**,
http://www.math.su.se/matstat

# En översikt över Cox's proportionella hazard model för inbäddad fall-kontroll studie

Mathias Lindholm*

januari 2004

**Sammanfattning**

Syftet med det här examensarbetet är att ge en intuitiv känsla för hur Cox's proportionella hazard modell är uppbyggd för vanliga kohortstudiedata, och hur den modellen kan modifieras så att den även kan användas för att analysera data från en inbäddad fall-kontroll-studie (*nested case control study*). Det här görs genom att först gå igenom några vanliga frågor, problem och storheter inom överlevnadsanalys, för att sedan ge en kort sammanfattning av de matematiska begrepp som behövs för att kunna förklara teorin. För den inbäddade fall-kontroll designen kommer det även presentreras en skattning av den absoluta risken. I de avslutande kapitlen kommer vi att analysera en inbäddad fall-kontroll-studie om cervical carcinoma in situ, livmoderhalscancer, som består av 373 fall och kontroller genom att tillämpa de metoder som har diskuterats tidigare i texten. Den analys som utförs kommer däremot inte att vara av den vanliga modellanpassningstypen, utan kommer mer att betona metodernas känslighet för val av vikter som berör hur kontrollerna har dragits ur grundpopulationen. Anledningen till detta är att vi inte har tillgång till de faktiska vikterna. Den korta analys som görs antyder att vikterna har en inverkan på resultaten.

**Abstract**

The aim of this thesis is to give an intuitive feeling for how Cox's proportional hazards model is built up for ordinary cohort study data, and how this model can be modified to work for data from a nested case control design. This is done by first introducing some common survival analytic questions, problems and quantities, and then give a short summary of the mathematics needed to develop the theory. For the nested case control design an estimate of the absolute risk will be presented. In the latter chapters we will have a look at a nested case control study about cervical carcinoma in situ containing information about 373 cases and controls, and use the methods that have been discussed earlier in the text. However, the analysis will not be of the normal model fitting kind, instead it will concentrate on the methods sensitivity to different choices of weights, concerning the sampling procedure of the controls. The reason for this is that the information about the needed weights are not availible. This brief analysis also indicate that the choice of weights is important to the results.

*E-post: m99mli@math.su.se. Handledare: Mikael Andersson.

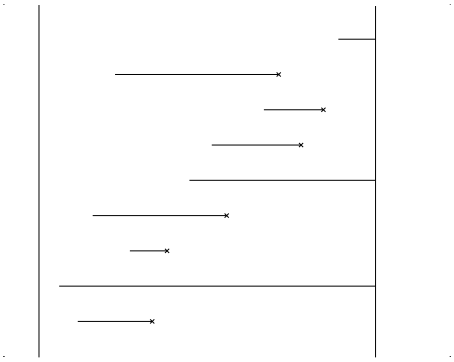
# Innehåll

# 1 Introduktion till överlevnadsanalys

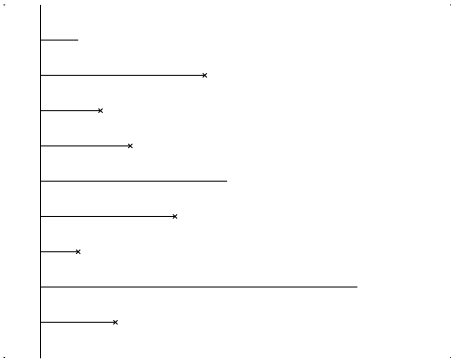
Överlevnadsanlays handlar om att hitta modeller som beskriver tid till det att någon viss händelse inträffar i processer som går från ett tillstånd till ett annat, t.ex. från frisk till sjuk eller levande till död. Sådana här försök brukar läggas upp i form av olika slags studier, och vanliga problem som då uppkommer är bortfall av information p.g.a. censurering och/ eller trunkering som man försöker kompensera för efter bästa möjliga förmåga.

Den vanligaste typen av censurering kallas högercensurering, och innebär att man inte får fullständig information för de individer som inte upplever händelsen som ska observeras före studiens slut. Trunkering innebär att vissa individer inte kommer med i studien t.ex. p.g.a. att de har upplevt en kontrollhändelse före studiens början, så man vet att händelsen har inträffat, men inte när den har inträffat. Den här typen av trunkering kallas vänstertrunkering.

En ytterligare effekt som ger ofullständig information om en individ kallas competing risk och är då en individ som är med i studien faller ur studien av annan orsak än den händelse som ska observeras t.ex. pga att individen dör eller flyttar så att uppföljning blir omöjlig. Figur 1 ger en schematisk bild av hur en studie kan tänkas se ut i tiden rent kalendermässigt, medan figur 2 visar tiderna till händelse för samma studie.



Figur 1: Observationer som slutar med ett kryss motsvarar exakta tider, medan observationer som når den högra kanten är högercensurerade.



Figur 2: Tider till händelse för studien i figur 1. Observationer som slutar med kryss är exakta och de övriga är högercensurerade

Intressanta frågor att försöka svara på är t.ex. hur stor sannolikhet har en individ att uppleva händelsen som observeras efter tiden $t$? eller, hur stor sannolikhet är det att en individ som inte har upplevt händelsen vid tiden $t$ kommer att ha upplevt händelsen vid $t + dt$?

Den första sannolikheten kan skrivas som $P(T > t)$ och kallas $T$:s överlevnadsfunktion och betecknas $S(t)$ där $T$ är någon stokastisk variabel som beskriver tiden tills en individ upplever händelsen som observeras. Då ser man att $S(t) = 1 - F(t)$, där $F(t)$ är $T$:s fördelningsfunktion. Om $T$ har en täthet, $f$, så kan man få följande samband

$$\begin{aligned} S(t) &= \int_t^\infty f(u)du \\ \Rightarrow f(t) &= -\frac{dS(t)}{dt} \end{aligned}$$

Den andra sannolikheten, normerad på lämpligt sätt, kan skrivas som

$$\lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt}$$

och kallas $T$:s hazard- eller riskfunktion och betecknas $h(t)$. Hazardfunktionen kallas ibland för intensitetsfunktionen och brukar då betecknas med $\lambda(t)$. Om $T$ har en täthet kan man skriva $h(t) = f(t)/S(t) = -d \log S(t)/dt$ där log är den naturliga logaritmen. En annan funktion som är av intresse är den kumulativa hazardfunktionen som betecknas $H(t)$, och som fås av

$$\begin{aligned} H(t) &= \int_0^t h(u)du = -\log S(t) \\ \Rightarrow S(t) &= \exp\{-H(t)\} \end{aligned}$$

På samma sätt som för hazardfunktionen kallas den kumulativa hazardfunktionen ibland även för den kumulativa intensitetsfunktionen.

För att skatta dessa funktioner finns det flera olika sätt att gå till väga på beroende på hur många restriktioner och antaganden som man är villig eller har möjlighet att göra. Dels finns det parametriska och icke-parametriska metoder, men den metod som kommer att användas lite längre fram i texten kallas semi-parametrisk Cox-regression, och är en multiplikativ intensitetsmodell. Det semi-parametriska ligger i att modellen har en icke-parametrisk del, en gemensam grundhazard, även kallad grundrisk eller baselinerisk, och en parametrisk del, där man gör ett parametriskt antagande om bakgrundsfaktorerna i modellen. En av fördelarna med Cox:s modell är just att den tar hänsyn till individuell bakgrundsinformation. Definitionen av en multiplikativ intensitetsmodell återkommer vi till lite längre fram i texten.

# 2 Upplägg

Tanken är nu att försöka ge en intuitiv förståelse för hur teorin till Cox-regressionen är uppbyggd för fallet med högercensurerade data i en kohortstudie, och sedan beskriva hur man utvidgar den modellen till en inbäddad fall-kontroll (*nested case control*) studie samt införa några tester som sedan ska tillämpas på en riktig studie.

För mer ingående detaljer i teorin än de som kommer att ges i den här uppsatsen hänvisas till följande böcker och artiklar:

Teorin som presenteras fram till kapitlet 'Inbäddad fall-kontroll studie' bygger först och främst på Fleming och Harrington [5], Andersen et al. [1] samt Klein och Moeschberger [8]. Efterföljande kapitel bygger på artiklarna Borgan och Langholz [2], Borgan et al. [3] samt Langholz och Borgan [9].

# 3 Beteckningar, definitioner och lite grundläggande teori

Den viktigaste byggstenen för den här teorin är martingaler. Men för att komma igång behöver vi först lite nödvändiga beteckningar och definitioner.

Antag att det finns något sannolikhetsrum $(\Omega, \mathcal{F}, P)$ där $\Omega$ är ett utfallsrum, $\mathcal{F}$ är en $\sigma$-algebra som innehåller hela $\Omega$ samt delmängder av $\Omega$ och $P$ är sannolikhetsmåttet som är definierat på delmängderna i $\mathcal{F}$.

Teorin som presenteras framöver kommer att till största delen handla om stokastiska processer och främst sådana som gör diskreta hopp i kontinuerlig tid. Formellt är en sådan här stokastisk process $X(t)$ en samling stokastiska variabler indexerade i tiden enligt $t \in [0, \infty)$.

Om man betraktar utfallet för $X(t)$ som en funktion $X(t, \omega)$ för $\omega \in \Omega$, så kan man se på $X(t, \omega)$ som en funktion av $t$ för fixt $\omega$. Den här funktionen brukar kallas för processen $X$:s realisering eller väg. Det här begreppet är viktigt när man talar om vilka egenskaper som en stokastisk process har i form av begränsningar och kontinuitet. Om man ser till en stokastisk integral $\int_0^t X(u)dY(u)$ så integrerar man processen $X$ med avseende på processen $Y$:s väg.

**Definition**: Filtration, historia, kontinuitet
En filtration $\{\mathcal{F}_t : t \geq 0\}$ är en växande familj av högerkontinuerliga sub-$\sigma$-algebror. Att filtrationen är högerkontinuerlig innebär att $\mathcal{F}_s = \cap_{t>s}\mathcal{F}_t$ för alla s. Filtrationer brukar även kallas

historier till stokastiska processer, vilket kommer av att $\mathcal{F}_t$ innehåller informationen om någon viss process $X$ upp till tiden $t$ samt eventuell annan tidsberoende information och ibland även mängden av nollhändelser. Om filtrationen innehåller mängden av nollhändelser sägs filtrationen vara komplett. Den minsta sub-$\sigma$-algebran som innehåller processen $X$:s historia upp till tiden t är den som är genererad av processen $X$. Den sub-$\sigma$-algebran definieras enligt $\mathcal{F}_t = \sigma\{X(s) : 0 \leq s \leq t\}$. Längre fram i texten kommer beteckningen $\mathcal{F}_{t-}$ användas, vilken syftar på historien för någon process i intervallet $[0, t)$.

**Definition**: Räkneprocess
En räkneprocess, i kontinuerlig tid, $N = \{N(t) : t \geq 0\}$ definieras av att $N(0) = 0$, $N(t) < \infty$, $N(t)$ antar bara heltalsvärden, är högerkontinuerlig och icke-avtagande. Ur detta följer direkt att $N(t) - N(s)$ är antalet steg som processen tagit i tidsintervallet $(s, t]$.

Antag att man har ett datamaterial med $m$ st individer. Låt $N_i(t)$ vara en $0/1$ räkneprocess som kontrollerar om individ $i$ har upplevt händelsen som ska observeras vid tiden $t$, dvs. $N_i(t) = I(T_i \leq t)$ där $T_i$ är den stokastiska variabeln som beskriver tiden till händelse för individ $i$. Alla individer förutsätts även ha unika hopptider, dvs. inga två individer hoppar samtidigt. Antag att datamaterialet i studien har högercensurerade observationer. Om man antar att högercensureringen är icke-informativ, vilket innebär att censureringstiden är oberoende av tiden till händelsen och att de censurerade tiderna inte tillför någon ytterligare information, så kan man definiera om räkneprocessen $N_i(t)$ såhär:

Låt $U_i$ vara den stokastiska variabel som motsvarar tiden för när studien avslutas för individ $i$, som är oberoende av $T_i$. För att en individ ska ha upplevt händelsen vid tiden $t$ så måste, som tidigare, $T_i \leq t$, men nu har man även kravet att $T_i \leq U_i$. Den räkneprocess som kontrollerar om individ $i$ har upplevt händelsen vid tiden $t$ definieras alltså av $N_i(t) = I(X_i \leq t, \delta_i = 1)$, där $X_i = \min(T_i, U_i)$ och $\delta_i = 1$ om $T_i \leq U_i$. Nu kan man sätta upp processen $\mathbf{N} = (N_1, \ldots, N_m)'$ som alltså innehåller informationen om alla $m$ st individerna.

Sätt nu $N(t) = \sum_{i=1}^{m} N_i(t)$, som då också blir en räkneprocess, men som endast gör hopp av längden $+1$, men som istället räknar det totala antalet individer som har upplevt händelsen vid tiden $t$ i försöket, dvs. hur många hopp som processen $\mathbf{N}$ har gjort vid tiden $t$.

En sådan här räkneprocess får några trevliga egenskaper. Om man ser till processen $N$ i ett litet tidsintervall $(t, t + dt]$, så kommer $dN(t) = N((t + dt)-) - N(t-)$ att vara approximativt en $0/1$ variabel. Det leder i sin tur till att $E(dN(t) \mid \mathcal{F}_{t-}) = P(dN(t) = 1 \mid \mathcal{F}_{t-})$. Väntevärdet för räkneprocessen $N$ i ett litet tidsintervall kommer alltså att motsvara sannolikheten för att processen tar ett steg i det tidsintervallet givet historien upp till $t-$.

För att nu återkomma till varför hazardfunktionen även brukar kallas för intensitetsfunktionen så ska vi titta på följande samband. Låt $N_i(t)$ vara en $0/1$ räkneprocess med icke-informativ högercensurering definierad enligt ovan, och anta att det existerar en täthet till $T_i$. Då gäller följande:

$$\begin{aligned}
E(N_i(t)) &= P(X_i \leq t, \delta_i = 1) \\
&= P(T_i \leq t, T_i \leq U_i) \\
&= \left\{ T_i \text{ antas ha en täthet} \right\} \\
&= \int_0^t P(U_i \geq u) f(u) du \\
&= \int_0^t P(U_i \geq u) S(u) \frac{f(u)}{S(u)} du \\
&= \left\{ h(u) = \lambda(u) = f(u)/S(u) \right\} \\
&= \int_0^t P(U_i \geq u) P(T_i > u) \lambda(u) du
\end{aligned}$$

$$\begin{aligned} &= \Big\{ T_i \text{ och } U_i \text{ oberoende } \Big\} \\ &= \int_0^t P(U_i \geq u, T_i \geq u)\lambda(u)du \\ &= E\Big( \int_0^t I(T_i \geq u, U_i \geq u)\lambda(u)du \Big) = E(A_i(t)) \end{aligned}$$

där $A_i(t) = \int_0^t I(T_i \geq u, U_i \geq u)\lambda(u)du$ och $A_i(t)$ kallas för $N_i$:s intensitetsprocess där $\lambda(u)$ är intensitetsfunktionen.

Ur detta följer att

$$E(N(t)) = E\Big( \sum_{i=1}^m N_i(t) \Big) = \sum_{i=1}^m E(N_i(t)) = \sum_{i=1}^m E(A_i(t)) = E(A(t))$$

där $A(t) = \sum_{i=1}^m A_i(t)$.

**Definition**: Martingaler, submartingaler
Om $E(\mid M(t) \mid) < \infty$ för alla $t \in [0, \infty)$ samt att
$E(M(t) \mid \mathcal{F}_s) = M(s)$ för $s \leq t$ så kallas den stokastiska processen $M(t)$ för martingal.
$E(M(t) \mid \mathcal{F}_s) \geq M(s)$ för $s \leq t$ så kallas den stokastiska processen $M(t)$ för submartingal.

Låt $s \leq t$ och $N(t)$ vara en räkneprocess definierad enligt ovan. Givet historien upp till tiden $s$ genererad av $\mathbf{N}$, så vet man vilka processer som har hoppat. Låt $N_i(t)$ beteckna den processen som gör ett hopp vid tiden $t_i$. Då gäller att

$$\begin{aligned} E(N(t) \mid \mathcal{F}_s) &= E\Big( \sum_{t_i \leq t} N_i(t) \mid \mathcal{F}_s \Big) \\ &= E\Big( \sum_{t_i \leq s} N_i(t) + \sum_{s < t_i \leq t} N_i(t) \mid \mathcal{F}_s \Big) \\ &= N(s) + E\Big( \sum_{s < t_i \leq t} N_i(t) \mid \mathcal{F}_s \Big) \\ \Rightarrow E(N(t) \mid \mathcal{F}_s) &\geq N(s) \end{aligned}$$

så $N(t)$ är alltså en submartingal. Ett annat sätt att komma fram till samma resultat är att betrakta de enskilda $N_i(t)$ som alla är icke-avtagande, ur vilket det följer att $N(t)$ är submartingal.

Eftersom vi ovan har visat att $E(N(t)) = E(A(t))$, samt visat att $N(t)$ är en submartingal så kommer följande resultat kanske att kännas lite mer rimligt.

Enligt Doob-Meyers dekompositionssats så existerar det till varje submartingal $N$ en unik högerkontinuerlig och bestämbar kompensator $A$ sådan att $M = N - A$ är en martingal, vilket leder till att $A$ är intensitetsprocessen till $N$. Ur detta följer att även alla $m$ st $M_i = N_i - A_i$ kommer att vara martingaler och att man kan sätta upp den simultana processen $\mathbf{M} = \mathbf{N} - \mathbf{A}$.

Vad innebär då det här bestämbara? Att processen $A$ är bestämbar syftar i det här avseendet på att $E(A(t) \mid \mathcal{F}_{t-}) = A(t)$, dvs. filtrationen som går upp till $t-$ bestämmer utfallet i $t$. Formellt säger man att $A(t)$ är $\mathcal{F}_{t-}$-mätbar. Om man ser till räkneprocessen $N(t) = \sum_{i=1}^m N_i(t)$, definierad ovan för högercensurerade data, och ser till

$$\begin{aligned} E(dN(t) \mid \mathcal{F}_{t-}) &= E(dA(t) \mid \mathcal{F}_{t-}) \\ &= dA(t) = I(T \geq t, U \geq t)\lambda(t)dt \end{aligned}$$

men $E(dN(t) \mid \mathcal{F}_{t-}) = P(dN(t) = 1 \mid \mathcal{F}_{t-})$ så $dA(t) = P(dN(t) = 1 \mid \mathcal{F}_{t-})$. $dA(t)$ är alltså den betingade sannolikheten att räkneprocessen $N$ hoppar ett steg i intervallet $(t, t + dt]$ givet historien upp till $t-$.

För att gå tillbaka till uppdelningen $M = N - A$, så kan man se på den här processen som en centrerad process med väntevärde 0, och där $A$ kan betraktas som systematisk faktor som $N$ rör sig slumpmässig kring för varje tidpunkt $t$, givet historien upp till $t-$.

# 4 Kohortstudie

En kohortstudie är upplagd så att man tar in individer i ett försök/studie mellan två tidpunkter, start och stopp, enligt något eller några kriterier, för att undersöka något visst samband mellan tid till någon viss händelse som t.ex. att jämföra hur två olika behandlingar påverkar tiden till tillfrisknande. Informationen om individer i studien samlas in prospektivt, vilket innebär att insamlingen av information sker under försökets gång, och när väl försöket är avslutat finns det inget behov av tillbakablickar för att kunna använda den insamlade informationen. Det som kohort syftar på i det här sammanhanget är att man gör gruppjämförelser. Antag att individ $i$ upplever händelsen som man är intresserad av vid tiden $t_i$. Den jämförelse man då gör är den mellan individ $i$ och de övriga individer som är vid risk vid tiden $t_i$. Att en individ är vid risk vid tiden $t_i$ innebär att personen inte har upplevt händelsen vid tiden $t_i-$ eller upplever händelsen vid tiden $t_i$. Riskmängden vid tiden $t$ är mängden av individer som då befinner sig vid risk och betecknas med $\mathcal{R}(t)$.

Notera att individerna som är vid risk vid tiden $t_i$ kommer även att kunna vara individer som senare kommer att högercensureras, men eftersom den enda information som man har till sitt förfogande vid $t_i$ är om individen har upplevt händelsen eller inte, samt eventuellt individuell bakgrundsdata. Antagandet om icke-informativ högercensurering kommer alltså inte leda till att man kastar bort all information om de censurerade individerna, utan snarare informationen som kan tänkas vara kopplad till censureringstidpunkten. Ur figur 2, ovan, kan man se hur riskmängden för de olika fallen i studien som visats i figur 1 kommer att se ut, och att dessa kommer att innehålla individer som kommer att censureras.

## 4.1 Cox modell

För att förstå hur Cox modell fungerar för en kohortstudie så behövs egentligen inte martingaler, utan det räcker med vanliga betingade sannolikheter. Anledningen till att det ändå är martingalsynsättet som presenteras här är för att det ska bli lättare att se hur man utvidgar den här modellen så att den fungerar för en inbäddad fall-kontroll studie.

Cox modell är, som tidigare nämnts, en semi-parametrisk regressionsmodell. Det semi-parametriska består i att modellen har en icke-parametrisk grundrisk samt att man antar ett parametriskt beroende av bakgrundsvariablerna. En av de stora fördelarna med Cox modell jämfört med de rent parametriska eller icke-parametriska modellerna är att den just tar hänsyn till bakgrundsvariabler, även kallade kovariater, som även kan tillåtas vara beroende av tiden.

**Definition**: Cox modell för fixa kovariater

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) \exp\{\boldsymbol{\beta}' \boldsymbol{Z}\}$$

där $\lambda_0(t)$ är grundrisken, baseline hazard, $\mathbf{Z} = (z_1, \ldots, z_p)$ är vektorn med kovariater, och $\boldsymbol{\beta}$ är vektorn med regressionskoefficienter.

Antag att det har gjorts en studie med $m$ st individer, och nu vill man jämföra hur stor den relativa risken är för individer med kovariater $\mathbf{Z}_1$ jämfört med individer med kovariater $\mathbf{Z}_2$. Det innebär att

$$\frac{\lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_1\}}{\lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_2\}} = \frac{\exp\{\boldsymbol{\beta}' \mathbf{Z}_1\}}{\exp\{\boldsymbol{\beta}' \mathbf{Z}_2\}} = \exp\{\boldsymbol{\beta}' (\mathbf{Z}_1 - \mathbf{Z}_2)\}$$

dvs. för två individer med olika kovariater kommer den relativa risken att vara konstant i tiden. Den här egenskapen gör att Cox modell brukar kallas för en proportionell hazard modell. Problemet

med den här egenskapen är att den inte alla gånger är rimlig, och leder till ett modellantagande som måste testas. Tester av modellens lämplighet kommer att tas upp lite längre fram.

Hur kommer då martingalen $M = N - A$ in i bilden igen? I Cox modell så är det ju intensitetsfunktionen $\lambda$ som man antar har ett visst utseende, och den är kopplad till processen $A$ vilket har visats tidigare. Om man utgår ifrån kompensatorn till en räkneprocess och antar att den har en viss form, under vissa förutsättningar, så kommer man fram till modeller som kallas för multiplikativa intensitetsmodeller, som just har egenskapen att de kan delas upp i $M = N - A$.

Följande definition av den multiplikativa intensitets modellen ges i Fleming och Harrington [5]:

Den multiplikativa intensitetsmodellen för observationer från $m$ oberoende individer, består av $m$ tripplar $(N_i, Y_i, \mathbf{Z}_i)$ av räkne-, censurerings- och kovariatprocesser, en högerkontinuerlig historia $\{\mathcal{F}_t : t \geq 0\}$, och de $m$ st intensitetsprocesserna $l_i(t) = \lambda_0(t)Y_i(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}$ under följande antaganden:

1. $\mathbf{N}$ är en simultan räkneprocess där inga två processer $N_i$ och $N_j$ hoppar samtidigt, dvs. $P(\Delta N_i(t) = \Delta N_j(t) = 1) = 0$ för alla $i \neq j$ och $t \geq 0$.
2. för alla $i$ gäller att $M_i = N_i - A_i$ är en martingal med avseende på $\mathcal{F}_t$, där
$$A_i(t) = \int_0^t l_i(u)du$$
är den kontinuerliga kompensatorn.
3. alla censureringsprocesser $Y_i$ och kovariatprocesser $\mathbf{Z}_i$ är bestämbara med avseende på $\mathcal{F}_t$. $\mathbf{Z}_i$:na är även lokalt begränsade.

Definitionen av den multiplikativa intensitetsmodellen för icke-informativt högercensurerade data med fixa kovariater säger att det existerar funktioner $l_i$ sådana att $A_i(t) = \int_0^t l_i(u)du$ som har formen

$$l_i(t) = \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}Y_i(t)$$

där $Y_i(t) = I(T_i \geq t, U_i \geq t)$ och uppfyller $M_i = N_i - A_i$. Med andra ord så är Cox modell en multiplikativ intensitetsmodell under dessa antaganden.

Eftersom intensitetsfunktionen $\lambda$ är betingad med avseende på kovariaterna $\mathbf{Z}$ kommer även överlevnadsfunktionen $S$ att vara betingad med avseende på dessa. Ur de samband mellan intensitetsfunktionen och överlevnadsfunktionen som visats tidigare så fås att

$$\begin{aligned} S(t \mid \mathbf{Z}) &= \exp\Big\{ -\int_0^t \lambda(u \mid \mathbf{Z})du\Big\} \\ &= \exp\Big\{ -\int_0^t \lambda_0(u)\exp\{\boldsymbol{\beta}'\mathbf{Z}\}du\Big\} \\ &= \exp\Big\{ -\exp\{\boldsymbol{\beta}'\mathbf{Z}\}\int_0^t \lambda_0(u)du\Big\} \\ &= S_0(t)^{\exp\{\boldsymbol{\beta}'\mathbf{Z}\}} \end{aligned}$$

där $S_0(t) = \exp\Big\{ -\int_0^t \lambda_0(u)du\Big\}$.

## 4.2 Likelihood, partiell likelihood samt skattning av $\boldsymbol{\beta}$

Nästa steg för att komma vidare är att kunna skatta den $p$-dimensionella vektorn $\boldsymbol{\beta}$, vilket man gör genom att lösa de $p$ stycken likelihoodekvationerna $\frac{\partial}{\partial\beta_g}\log L(\boldsymbol{\beta}) = 0$, för $g = 1, \ldots, p$, numeriskt med hjälp av t.ex. Newton-Raphsons metod.

Den vanliga likelihooden använder man ju till att hitta skattningar av olika slags parametrar, i vårt fall regressionsparametrarna $\boldsymbol{\beta}$. Men vad är då den partiella likelihooden för något? Tanken med den partiella likelihooden är att hitta en mindre likelihood än den vanliga, men som fortfarande har samma egenskaper. Anledningen till detta är att om man ser till de data som man får från en överlevnadsstudie, så kommer det att finnas faktorer som påverkar dess fullständighet, t.ex. censurering och trunkering. Det man nu vill göra är att få bort de data som inte tillför någon ytterligare information vad gäller skattningarna av regressionsparametrarna $\boldsymbol{\beta}$.

Antag att man har ett icke-informativt högercensurerat datamaterial som innehåller information om $(X_i, \mathbf{Z}_i, \delta_i)$ för individerna $i = 1, \ldots, m$. Antag att bara $D$ individer har upplevt händelsen vid studiens slut vid tiderna $t_1 < \cdots < t_D$ och låt $t_1 > t_0 = 0$ samt $t_D < t_{D+1} = \infty$. Antag även att informationen i $(X_i, \mathbf{Z}_i, \delta_i)$ kan skrivas om som händelser $G_i$, $H_i$, $i = 1, \ldots,$ D enligt följande:

Låt $G_i$ vara händelsen att den ordnade individ $(i)$ har upplevt händelsen vid tiden $t_i$, och låt $H_i$ beteckna händelsen för de individer som censureras i intervallet $[t_{i-1}, t_i)$ samt deras censureringstider och att precis en individ har upplevt händelsen vid tiden $t_i$. Nu kan man ställa upp likelihooden

$$\begin{aligned} L(\boldsymbol{\beta}) &\propto P(G_1, H_1, \ldots, G_D, H_D) \\ &= \left\{ \prod_{i=2}^{D} P(G_i \mid H_i, G_{i-1}, H_{i-1}, \ldots, G_1, H_1) \right\} P(G_1 \mid H_1) \\ &\times \left\{ \prod_{i=2}^{D} P(H_i \mid G_{i-1}, H_{i-1}, \ldots, G_1, H_1) \right\} P(H_1). \end{aligned}$$

Nu är likelihooden uppdelad i två delar. Under antagandet om icke-informativ högercensurering består den första delen av betingade sannolikheter som rör de exakta observationerna och den andra av de som berör censureringar. Om man nu försöker att dra paralleller mellan dessa två produkter och deras inverkan på $\boldsymbol{\beta}$, är det förhoppningsvis lättare att tänka på dessa två uttrycks förhållande till kovariaterna, som i sin tur är kopplade till $\boldsymbol{\beta}$. Att kovariaterna är kopplade till de exakta observationerna är rimligt, och det är även rimligt att tro att de censurerade observationerna inte har någon koppling till kovariaterna, eller endast har en svag koppling. Med hjälp av det här lösa resonemanget kan vi nu alltså stryka den andra produkten ur uttrycket ovan, och vi har då fått fram den partiella likelihooden

$$\mathcal{L}(\boldsymbol{\beta}) \propto \left\{ \prod_{i=2}^{D} P(G_i \mid H_i, G_{i-1}, H_{i-1}, \ldots, G_1, H_1) \right\} P(G_1 \mid H_1).$$

För att bestämma uttrycket för den partiella likelihooden för en Cox-modell för en kohortstudie med icke-informativ högercensurering där inga individer upplever händelsen samtidigt så kan man bestämma uttrycken för de betingade sannolikheterna såhär:

$$\begin{aligned} &P(\text{individ (i) upplever händelsen vid } t_i \mid \text{en individ upplever händelsen vid } t_i) = \\ &= \frac{P(\text{individ (i) upplever händelsen vid } t_i \mid \text{har inte upplevt händelsen vid } t_{i-})}{P(\text{en individ upplever händelsen vid } t_i \mid \text{har inte upplevt händelsen vid } t_{i-})} \\ &= \frac{\lambda(t|\mathbf{Z}_{(i)})}{\sum_{j \in \mathcal{R}_i} \lambda(t|\mathbf{Z}_j)} \\ &= \frac{\lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_{(i)}\}}{\sum_{j \in \mathcal{R}_i} \lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{Z}_j\}} \\ &= \frac{\exp\{\boldsymbol{\beta}' \mathbf{Z}_{(i)}\}}{\sum_{j \in \mathcal{R}_i} \exp\{\boldsymbol{\beta}' \mathbf{Z}_j\}} \end{aligned}$$

där $\mathcal{R}_i$ är riskmängden, mängden av individer som inte har upplevt händelsen vid $t_i$, för den ordnade individen $(i)$ som upplever händelsen vid tiden $t_i$. Då blir den partiella likelihooden

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\{\boldsymbol{\beta}' \mathbf{Z}_{(i)}\}}{\sum_{j \in \mathcal{R}_i} \exp\{\boldsymbol{\beta}' \mathbf{Z}_j\}}$$

På samma sätt som i den vanliga maximumlikelihoodteorin så fås de partiella maximum likelihood skattningarna genom att lösa de $p$ stycken ekvationerna $\frac{\partial}{\partial\beta_g}\log\mathcal{L}(\boldsymbol{\beta})=0$ för $g=1,\ldots,p$ där

$$\log\mathcal{L}(\boldsymbol{\beta})=\sum_{i=1}^{D}\left(\boldsymbol{\beta}'\mathbf{Z}_{(i)}-\log\sum_{j\in\mathcal{R}_i}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\}\right)$$

vilket ger

$$\frac{\partial}{\partial\beta_g}\log\mathcal{L}(\boldsymbol{\beta})=\sum_{i=1}^{D}z_{(i)g}-\sum_{i=1}^{D}\frac{\sum_{j\in\mathcal{R}_i}z_{jg}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\}}{\sum_{j\in\mathcal{R}_i}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\}}=U_g(\boldsymbol{\beta})$$

där $U_g$ är en komponent i score-funktionen $\mathbf{U}(\boldsymbol{\beta})$.

Informationsmatrisen kommer även att behövas, bland annat för att kunna använda Newton-Raphsons metod. Som vanligt kommer elementen i informationsmatrisen att ges av

$$\begin{aligned}\mathcal{I}_{gh}(\boldsymbol{\beta}) &= -\frac{\partial^2}{\partial\beta_g\partial\beta_h}\log\mathcal{L}(\boldsymbol{\beta})\\ &= \sum_{i=1}^{D}\frac{\sum_{j\in\mathcal{R}_i}z_{jg}z_{jh}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\}}{\sum_{j\in\mathcal{R}_i}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\}}-\sum_{i=1}^{D}\frac{(\sum_{j\in\mathcal{R}_i}z_{jg}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\})(\sum_{j\in\mathcal{R}_i}z_{jh}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\})}{(\sum_{j\in\mathcal{R}_i}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\})(\sum_{j\in\mathcal{R}_i}\exp\{\boldsymbol{\beta}'\mathbf{Z}_j\})}\end{aligned}$$

Vad gäller skattningarna av $\boldsymbol{\beta}$ så kommer dessa att vara asymptotiskt normalfördelade enligt

$$\hat{\boldsymbol{\beta}}\in N(\hat{\boldsymbol{\beta}},\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}).$$

Eftersom den partiella likelihooden inte är en fullständig likelihood, så kommer dess egenskaper inte heller att bete sig som sig bör i alla lägen, men för de fall som är intressanta kommer inga problem att uppstå.

Det här sättet att angripa den partiella likelihooden har ju inte använt sig av martingaler över huvudtaget, men har förhoppningsvis givit en lite bättre förståelse för hur den partiella likelihooden är uppbyggd och fungerar. För en härledning av den partiella likelihooden utgående ifrån intensitetsprocessen hänvisas till t.ex. Fleming och Harrington [5].

## 4.3 Tester och skattningar

Det här kapitlet kommer i princip bara att gå ut på att definiera ett par vanliga signifikanstester av $\boldsymbol{\beta}$. Anledningen till detta är att för en mer komplett genomgång kommer man att behöva diverse asymptotiska resultat som går utanför uppsatsens omfattning. För en ordentlig genomgång hänvisas till Fleming och Harrington [5] eller Andersen et al. [1].

Skattningar av den absoluta risken och överlevnadsfunktionen för givna kovariater kommer inte att presenteras i det här kapitlet, p.g.a. av att förfarandet som kommer att presenteras i kapitlet som rör inbäddad fall-kontroll studien kommer att vara en omväg jämfört med det traditionella.

## 4.4 Score-test, LR-test och Wald's test

För att testa hypotesen $H_0:\boldsymbol{\beta}=\boldsymbol{\beta}_0$ för $\boldsymbol{\beta}=\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_p\}'$, finns det ett antal olika tester att ta till. De tester som presenteras här, bygger på att statistikorna är asymptotiskt $\chi^2$ - fördelade med $p$ frihetsgrader under $H_0$.

Statistikorna för dessa tester är:

$$\begin{aligned}\text{Score-test: } \chi^2_{SC} &= (\mathbf{U}(\boldsymbol{\beta}_0))'\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{U}(\boldsymbol{\beta}_0)\\ \text{Likelihood Ratio-test: } \chi^2_{LR} &= 2\{\log\mathcal{L}(\hat{\boldsymbol{\beta}})-\log\mathcal{L}(\boldsymbol{\beta}_0)\}\\ \text{Walds's test: } \chi^2_{W} &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\end{aligned}$$

## 4.5 Diagnostik, modellantaganden, martingalresidualer

För att kunna förlita sig på de resultat som man har fått fram vid test av olika kovariaters inverkan på modellen, så måste man undersöka hur väl modellantagandena är uppfyllda. För Cox modell är det, som sagts tidigare, antagandet om proportionell risk som måste undersökas. Om man inte vet hur kovariaterna kan tänkas bete sig är det även av intresse att undersöka deras form, dvs. om de bör transformeras eller inte.

## 4.6 Martingalresidualer

Om man ser till de $m$ st lokala martingalerna $M_i = N_i - A_i$ för Cox modell med fixa kovariater för icke-informativt högercensurerat datamaterial så är

$$A_i(t) = \int_0^t I(T_i \geq u, U_i \geq u) \exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}\lambda_0(u)du.$$

Låt $\Lambda(t)$ beteckna den kumulativa intensitetsfunktionen och $\Lambda_0(t)$ beteckna den kumulativa baselineintensitetsfunktionen. Nu kan man skriva

$$A_i(t) = \int_0^t I(T_i \geq u, U_i \geq u) \exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}d\Lambda_0(u).$$

Nu ska vi ta oss en titt på Breslow's estimator

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{j=1}^m dN_j(u)}{\sum_{k=1}^m Y_k(u)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}}$$

För en härledning av Breslow's estimator så hänvisas till t.ex. Fleming och Harrington [5]. Breslow's estimator får följande utseende för icke-informativt högercensurerade data

$$\begin{aligned}\hat{\Lambda}_0(t) &= \int_0^t \frac{\sum_{j=1}^m dN_j(u)}{\sum_{k=1}^m I(T_k \geq u, U_k \geq u)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}} \\ &= \sum_{j=1}^m \int_0^t \frac{1}{\sum_{k=1}^m I(T_k \geq u, U_k \geq u)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}} dN_j(u)\end{aligned}$$

men eftersom alla $N_j(u)$ är högerkontinuerliga $0/1$ processer så är

$$\int_0^t f_j(u)dN_j(u) = f_j(u_j)\Delta N_j(u_j) \text{ om } u_j \in [0,t] \text{ och } 0 \text{ annars.}$$

$u_j$ motsvarar här hopptiden för processen $N_j$ och $\Delta N_j(u_j) = N_j(u_j) - N_j(u_j-) = 1$.
Nu följer att

$$\begin{aligned}\hat{\Lambda}_0(t) &= \sum_{u_j \leq t} \frac{1}{\sum_{k=1}^m I(T_k \geq u_j, U_k \geq u_j)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}} \\ &= \sum_{u_j \leq t} \frac{1}{\sum_{k\in\mathcal{R}_j}\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}}\end{aligned}$$

Om man nu ser till

$$d\hat{\Lambda}_0(t) = \frac{\sum_{j=1}^m dN_j(t)}{\sum_{k=1}^m I(T_k \geq t, U_k \geq t)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_k\}}$$

så kan man nu ställa upp

$$\begin{aligned}\hat{M}_i(t) &= N_i(t) - \int_0^t I(T_i \geq u, U_i \geq u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} d\hat{\Lambda}_0(u) \\ &= N_i(t) - \int_0^t I(T_i \geq u, U_i \geq u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} \frac{\sum_{j=1}^m dN_j(u)}{\sum_{k=1}^m I(T_k \geq u, U_k \geq u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_k\}} \\ &= N_i(t) - \sum_{j=1}^m \int_0^t \frac{I(T_i \geq u, U_i \geq u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\}}{\sum_{k=1}^m I(T_k \geq u, U_k \geq u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_k\}} dN_j(u)\end{aligned}$$

och på samma sätt som ovan kommer integralen att kunna skrivas om, vilket ger

$$\begin{aligned}\hat{M}_i(t) &= N_i(t) - \sum_{u_j \leq t} \frac{I(T_i \geq u_j, U_i \geq u_j)}{\sum_{k=1}^m I(T_k \geq u_j, U_k \geq u_j) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_k\}} \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} \\ &= \{I(T_i \geq u_j, U_i \geq u_j) = 0 \text{ för alla } 0 \leq u_i \leq u_j \leq t \text{ där } u_i \text{ är hopptiden för } N_i\} \\ &= N_i(t) - \sum_{u_j \leq u_i \leq t} \frac{1}{\sum_{k \in \mathcal{R}_j} \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_k\}} \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\}\end{aligned}$$

Tanken är nu att undersöka $\hat{M}_i$ då $t \to \infty$.

**Definition**: Martingalresidualer för fixa kovariater

$$\hat{M}_i = N_i(\infty) - \int_0^\infty Y_i(u) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} d\hat{\Lambda}_0(u).$$

För icke-informativt högercensurerade data fås då att

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} = \delta_i - r_i$$

där $\delta_i = 0$ om individ i är censurerad och 1 annars. $t_i$ motsvarar här tiden då individ i upplever händelsen eller censureras. Nu kan man se på uppdelningen $\hat{M}_i = \delta_i - r_i$ som att $\delta_i$ motsvarar det observerade antalet fall och $r_i$ motsvarar de förväntade av modellen. De $m$ st $\hat{M}_i$ kommer även ha den egenskapen att de summerar till noll.

Det fina med martingalresidualerna är att man kan undersöka kovariaters funktionella form. Med en kovariat $Z_i$:s funktionella form, menas dess beteende, som eventuellt följer någon funktion av typen $Z_i^2$, $\log Z_i$ osv. Om man utgår från att man vet den funktionella formen hos kovariaterna $\mathbf{Z}^*$ och har motsvarande koefficienter $\boldsymbol{\beta}^*$ så kan man göra följande uppställning

$$\Lambda(t \mid \mathbf{Z}^*, Z_i) = \Lambda_0(t) \exp\{\boldsymbol{\beta}^{*\prime} \mathbf{Z}^*\} \exp\{g(Z_i)\}$$

Eftersom $\hat{M}_i$ kommer att vara centrerad kring nollan för den korrekta modellen, vilket den antas vara för $\mathbf{Z}^*$, så kommer en avvikelse att antyda den funktionella formen $g$ hos den nya kovariaten $Z_i$.

## 4.7 Test av proportionalitet, Andersen-plottar

För att undersöka antagandet om proportionell risk kommer vi att använda en grafisk metod som kallas Andersen-plottar. Tanken bakom Andersen-plottar är att man undersöker den kumulativa intensitetsfunktionen för olika värden på kovariaterna, och plottar dessa mot varandra. Om antagandet om proportionalitet är korrekt bör man få räta linjer som går genom origo. Rent praktiskt så gör man på liknande sätt som för martingalresidualerna. Man börjar med att anta att man har tillgång till 'korrekta' kovariater $\mathbf{Z}^*$ och lägger till en ny kovariat $Z_i$ som bara kan anta ett fixt antal värden. Om $Z_i$ är kontinuerlig får man göra någon lämplig stratifiering. Metoden förutsätter även att kovariaterna är tidsoberoende, dvs. fixa.

# 5 Inbäddad fall-kontroll studie

Skillnaden mellan en fall-kontroll studie och en kohortstudie, är att man till varje fall som upplever händelsen vid tiden $t_i$, samplar en eller flera kontroller som är vid risk vid tiden $t_i$. Jämförelsen som görs kommer då att vara den mellan fallet och dess samplade kontroller, istället för mellan fallet och alla övriga individer som är vid risk. Anledningen till att man säger att studien är inbäddad, kommer av att den är inbäddad i en större kohort, för man gör ju ett val av både fall och kontroller ur en större population. Det här leder till att ett fall kan ha varit en kontroll till ett annat fall tidigare i studien. Fördelen med ett sådant här försök är att man inte behöver samla in bakgrundsinformation angående alla individer i kohorten, utan att det räcker att göra det för de samplade individerna. För att kompensera för de personer som inte samplas kommer det in vikter som beskriver förhållandet mellan antal samplade individer och hela kohorten. Dessa vikter kommer att visa sig viktiga vid skattningen av överlevndsfunktionen m.m.

## 5.1 Utvidgning av Cox modell till inbäddad fall-kontroll

För att få en bättre känsla för hur inbäddad fall-kontroll studien kommer att påverka räkneprocesserna kommer följande avsnitt att vara en förenklad framställning av den som ges i Borgan och Langholz [2]. Tyngdpunkten kommer att läggas på hur intensitetsprocessen byggs upp och på hur vikterna kommer in i bilden.

Låt $\tilde{\mathcal{R}}(t)$ vara den riskmängd av storlek $k$ som samplas utan återläggning ur den vanliga riskmängden $\mathcal{R}(t)$, och som innehåller fallet samt $k-1$ st kontroller. För att underlätta beräkningarna kommer antalet kontroller som samplas att antas vara fixt, ty om antalet kontroller får variera blir man tvungen att skatta sannolikheten för att man just har fått $k(t)-1$ st kontroller vid tiden $t$.

Tanken med en inbäddad fall-kontroll studie är ju att man bara behöver samla in bakgrundsinformation om de individer som finns med i $\tilde{\mathcal{R}}(t)$, men detta innebär att man har begränsat sin ursprungliga information. För att kompensera för detta inför man vikter som beskriver relationen mellan $\tilde{\mathcal{R}}(t)$ och $\mathcal{R}(t)$. Dessa vikter kommer att visa sig bli $n(t)/k$ där $n(t) =$ antal individer i $\mathcal{R}(t)$.

För att göra utvidgningen från en kohort till inbäddad fall-kontroll studie kommer man i princip att genomföra samma steg som tidigare, men med skillnaden att man nu även måste ta hänsyn till samplingen av riskmängderna $\tilde{\mathcal{R}}(t)$. För att få en bättre förståelse för hur hur det här går till kommer här en rätt kortfattad, men förhoppningsvis illustrativ, genomgång av hur intensitetsprocessen byggs upp.

## 5.2 Intensitetsprocessen

Om man utgår från samma processer som för kohortstudien, med tillhörande historier $\mathcal{F}_{t-}$, och den vanliga riskmängden $\mathcal{R}(t)$, och skapar en ny riskmängd $\tilde{\mathcal{R}}(t) \subset \mathcal{R}(t)$ som innehåller $k$ st individer, där individerna som ingår i den nya riskmängden är dragna utan återläggning ur den tidigare. En nödvändig åtgärd är nu att skapa en ny historia $\mathcal{H}_{t-}$, som är historien $\mathcal{F}_{t-}$ utvidgad med informationen som rör urvalet till $\tilde{\mathcal{R}}(t)$.

För en räkneprocess $N_i$ definierad som tidigare blir

$$A_i(t)dt = l_i(t)dt = P(dN_i(t) = 1 \mid \mathcal{F}_{t-}) = I(T_i \geq t, U_i \geq t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}\lambda_0(t)dt$$

och antagandet man nu gör, är att $P(dN_i(t) = 1 \mid \mathcal{H}_{t-}) = P(dN_i(t) = 1 \mid \mathcal{F}_{t-})$, dvs. den ytterligare informationen om vilka individer som har valts vid tiden $t$ påverkar inte intensiteten, och urvalet sägs vara oberoende.

Nu har vi kommit fram till den stora skillnaden mot tidigare, nämligen att sätta upp den betingade urvalssannolikheten $P(\tilde{\mathcal{R}}(t) = \mathbf{r} \mid \Delta N_i(t) = 1, \mathcal{H}_{t-})$, där $\mathbf{r} \subset \mathcal{R}(t)$, $\mathbf{r} \in \mathcal{P}^{(k)}$ och $i \in \mathbf{r}$. $\mathcal{P}^{(k)}$ är här mängden av alla delmängder av $\{1, \ldots, m\}$ av storlek $k$. Eftersom riskmängden $\mathbf{r}$ dras utan återläggning, och det är givet att individ $i$ upplever händelsen i intervallet $\Delta N_i(t) = N_i(t) - N_i(t-)$, så får man att

$$P(\tilde{\mathcal{R}}(t) = \mathbf{r} \mid \Delta N_i(t) = 1, \mathcal{H}_{t-}) = \binom{n(t)-1}{k-1}^{-1}$$

För att hålla reda på informationen om urvalet inför man nu räkneprocessen $N_{i,\mathbf{r}}(t) = I(X_i \leq t, \delta_i = 1)I(\tilde{\mathcal{R}}(t) = \mathbf{r})$, och under antagandet om oberoende urval kan man visa att

$$\begin{aligned} l_{i,\mathbf{r}}(t) &= P(dN_{i,\mathbf{r}} = 1 \mid \mathcal{H}_{t-}) \\ &= I(T_i \geq t, U_i \geq t) \exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}\lambda_0(t)dt \binom{n(t)-1}{k-1}^{-1} I(\mathbf{r} \subset \mathcal{R}(t), \mathbf{r} \in \mathcal{P}^{(k)}, i \in \mathbf{r}) \end{aligned}$$

och att

$$M_{i,\mathbf{r}}(t) = N_{i,\mathbf{r}}(t) - \int_0^t l_{i,\mathbf{r}}(u)du = N_{i,\mathbf{r}}(t) - A_{i,\mathbf{r}}(t)$$

## 5.3 Partiell likelihood och tester av $\boldsymbol{\beta}$

Man kan nu visa att den partiella likelihooden blir

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\{\boldsymbol{\beta}'\mathbf{Z}_{(i)}\}}{\sum_{l \in \tilde{\mathcal{R}}_i} \exp\{\boldsymbol{\beta}'\mathbf{Z}_l\}}$$

som har samma form som den vanliga partiella likelihooden för en kohortstudie, men med skillnaden att man nu har en annan riskmängd. Ur detta samband kan man på samma sätt som tidigare få ut score-funktionen och informationsmatrisen, samt de tester som nämnts i tidigare kapitel. Även här kommer $\boldsymbol{\beta}$-skattningarna att vara asymptotiskt normalfördelade enligt

$$\hat{\boldsymbol{\beta}} \in N(\hat{\boldsymbol{\beta}}, \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}).$$

## 5.4 Skattning av överlevnadsfunktionen

Anledningen till att skattningen av överlevnadsfunktionen inte togs upp för den vanliga kohortstudien är att det förfarandet som presenteras i Langholz och Borgan [9] för inbäddad fall-kontroll studien är lite annorlunda.

Istället för att göra som för fallet med en kohortstudie där man sätter upp

$$\hat{S}(t \mid \mathbf{Z}) = \hat{S}_0(t)^{\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}\}}$$

där

$$\hat{S}_0(t) = \exp\{-\hat{\Lambda}_0(t)\}$$

så går man via

$$\hat{S}(t \mid \mathbf{Z}) = \prod_{t_j \leq t} (1 - \hat{\lambda}(t_j \mid \mathbf{Z}))$$

där

$$\hat{\lambda}(t_j \mid \mathbf{Z}) = \frac{\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}\}}{\sum_{l \in \tilde{\mathcal{R}}_j} (n(t_j)/k) \exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_l\}}.$$

$t_j$ är här den j:te ordnade tiden för händelse.

Det intressanta nu är att räkna ut konfidensband för överlevnadsfunktionen. För att beräkna dessa behöver vi använda oss av följande uttryck för variansen som ges i [9]:

$$Var(\hat{S}(t \mid \mathbf{Z})) = \hat{S}^2(t \mid \mathbf{Z}) Var(\hat{\Lambda}(t \mid \mathbf{Z}))$$

Den metod som nu presenteras för att beräkna konfidensband kallas Hall-Wellners metod, men detaljer som förklarar hur dessa fungerar är utanför ramarna för denna uppsats. För mer detaljer se t.ex. Klein och Moeschberger [8].

Log-transformerade Hall-Wellner band för överlevnadsfunktionen har följande utseende

$$[\hat{S}(t \mid \mathbf{Z})^{1/\theta}, \hat{S}(t \mid \mathbf{Z})^{\theta}]$$

där

$$\theta = \exp\left\{ \frac{k_\alpha(a_L, a_U)[1 + n\sigma^2(t \mid \mathbf{Z})]}{n^{1/2} \log \hat{S}(t \mid \mathbf{Z})} \right\}$$

$\sigma^2(t \mid \mathbf{Z})$ är variansen för $\hat{S}(t \mid \mathbf{Z})$, $k_\alpha(a_L, a_U)$ är konfidenskoefficienten och den fås ur tabell, t.ex. i Klein och Moeschberger [8].

$a_L$ och $a_U$ beräknas enligt följande:

$$\begin{aligned} a_L &= \frac{n\sigma^2(t_L \mid \mathbf{Z})}{1 + n\sigma^2(t_L \mid \mathbf{Z})} \\ a_U &= \frac{n\sigma^2(t_U \mid \mathbf{Z})}{1 + n\sigma^2(t_U \mid \mathbf{Z})} \end{aligned}$$

där $t_L$ är den undre tidsgränsen för konfidensbandet, vanligtvis 0, och $t_U$ är den övre, som vanligtvis är tiden för den sista observerade händelsen. Man bör dock ha i åtanke att längden på konfidensbanden påverkar dess bredd.

Anledningen till att man använder log-transformerade Hall-Wellner band, är för att man kan få värden som går över 1 om man använder de vanliga linjära banden.

## 5.5 Martingalresidualer

För att kunna få fram martingalresidualer måste man komma åt den kumulativa intensitetsfuntionen. Dessvärre kommer vi inte längre att kunna att använda oss av Breslow's estimator, utan blir tvugna att ta till följande skattning

$$\hat{\Lambda}_0(t) = \sum_{u_j \leq t} \frac{1}{\sum_{l \in \tilde{\mathcal{R}}_j} (n(u_j)/k) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_l\}}$$

som föreslås i Borgan och Langholz [2]. På liknande sätt som tidigare får man fram att martingalresidualerna blir

$$\hat{M}_i(t) = \delta_i - \sum_{u_j \leq t_i} \frac{1}{\sum_{l \in \mathcal{R}_j} (n(u_j)/k) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_l\}} \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\} = \delta_i - \hat{\Lambda}_0(t_i) \exp\{\hat{\boldsymbol{\beta}}' \mathbf{Z}_i\}$$

## 5.6 Test av proportionalitet, Andersen-plottar

Förfarandet blir det samma som för den vanliga kohortstudien, men med den nya skattningen av den kumulativa intensitetsfunktionen istället för Breslow's skattning.

### 5.7 Absolut risk, skattning och definition

En annan intressant storhet att skatta i överlevnadsstudier som vi inte har tagit upp tidigare är den absoluta risken. Den absoluta risken är den obetingade risken för att en godtycklig individ med vissa riskfaktorer kommer att uppleva händelsen som vi vill observera vid någon viss tidpunkt. Skattningen av den absoluta risken mellan tiden $s$ och $t$ för en individ med kovariaterna $\mathbf{Z}_0$ ges av:

$$\hat{\pi}(s,t|\mathbf{Z}_0) = \sum_{s<t_j\leq t} \hat{S}(s,t_j|\mathbf{Z}_0)\hat{\lambda}(t_j|\mathbf{Z}_0)$$

där

$$\hat{S}(s,u|\mathbf{Z}_0) = \prod_{s<t_j\leq u} (1-\hat{\lambda}(t_j|\mathbf{Z}_0))$$

och

$$\hat{\lambda}(t_j|\mathbf{Z}_0) = \frac{\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_0\}}{\sum_{l\in\mathcal{R}_j}(n(t_j)/k)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_l\}}$$

Uttrycket för variansen skattas med:

$$\hat{V}ar(\hat{\pi}(s,t|\mathbf{Z}_0)) = \hat{V}ar_1(\hat{\pi}(s,t|\mathbf{Z}_0)) + \hat{V}ar_2(\hat{\pi}(s,t|\mathbf{Z}_0))$$

där

$$\hat{V}ar_1(\hat{\pi}(s,t|\mathbf{Z}_0)) = \hat{\mathbf{G}}(s,t|\mathbf{Z}_0))'\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}\hat{\mathbf{G}}(s,t|\mathbf{Z}_0))$$

och

$$\hat{\mathbf{G}}(s,t|\mathbf{Z}_0)) = \sum_{s<t_j\leq t} \hat{S}(s,t_j|\mathbf{Z}_0)[1-\hat{\pi}(t_j,t|\mathbf{Z}_0)]\hat{\mathbf{b}}(t_j|\mathbf{Z}_0)$$

med

$$\hat{\mathbf{b}}(t_j|\mathbf{Z}_0) = \frac{(\partial/\partial\boldsymbol{\beta})\exp\{\boldsymbol{\beta}\}\hat{}'\mathbf{Z}_0}{\sum_{l\in\mathcal{R}_j}(n(t_j)/k)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_l\}} - \frac{\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_0\}\sum_{l\in\mathcal{R}_j}(n(t_j)/k)(\partial/\partial\boldsymbol{\beta})\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_l\}}{\{\sum_{l\in\mathcal{R}_j}(n(t_j)/k)\exp\{\hat{\boldsymbol{\beta}}'\mathbf{Z}_l\}\}^2}$$

och slutligen är

$$\hat{V}ar_2(\hat{\pi}(s,t|\mathbf{Z}_0)) = \sum_{s<t_j\leq t} \{\hat{S}(s,t_j|\mathbf{Z}_0)[1-\hat{\pi}(t_j,t|\mathbf{Z}_0)]\hat{\lambda}(t_j|\mathbf{Z}_0)\}^2.$$

Alla $t_j$ motsvarar här tid för en distinkt händelse.

## 6 HPV studien

Den här studien som vi nu ska tillämpa de metoder som vi har gått igenom på handlar om att undersöka sambandet mellan mängden av HPV-virus, *Human Papilloma Virus*, och tiden till diagnosen invasiv cervical carcinoma, livmoderhalscancer. Övriga faktorer som kan tänkas samverka är rökning, sexuell aktivitet, social bakgrund, och historia om tidigare könssjukdomar. För mer information om studien än vad som kommer att redovisas här hänvisas till Ylitalo [10], ur vilken all information om urvalsförfarande med mera är hämtat, och där även den medicinska bakgrunden finns presenterad.

## 6.1 Bakgrund

Under 1967 så påbörjades insamling av cytologiska prover angående livmoderhalscancer i Uppsala län, och från och med 1969 har prover samlats in från alla kvinnor som varit bosatta i Uppsala län. Cytologiska prover görs även vid gynekologiska undersökningar, vilket gör att olika kvinnor har tagit olika många prover vid olika tidpunkter, relativt deras födsel. Mellan 1969 och 1995 har det samlats in prover från totalt 146 889 kvinnor, och alla dessa prover finns lagrade på Uppsala Universitetssjukhus, patologiska institutionen.

## 6.2 Datainsamling och urval

Ur materialet här ovan har sedan en studiekohort skapats utgående ifrån vissa kriterier. De kvinnor som först togs in i kohorten måste ha följande kriterier uppfyllda:

- hon har minst ett registrerat cytologiskt prov mellan 1969 och 1995
- det första provet måste vara osmittat
- hon är född i Sverige
- hon var yngre än 50 då hennes första prov togs
- hon har inte haft någon 'malignant disease in cervix'
- hon var vid liv och tillgänglig för en personlig intervju då studien påbörjades (1 januari, 1996)

Det första provet motsvarar tiden då en kvinna kom in i kohorten.

Ur denna kohort har sedan alla fall som utvecklat 'cervical carcinoma in situ' mellan 1969 och 1995 identifierats genom att jämföra registret på Uppsala Universitetssjukhus med det nationella cancerregistret. Till varje fall har det sedan matchats fem potentiella kontroller ur studiekohorten. Dessa har valts slumpmässigt ur kohorten och matchats efter

- punkten för att komma in i kohorten, $\pm 90$ dagar
- födelseår
- kontrollen är smittfri vid tiden då motsvarande fall fått sin diagnos

Sedan har alla första prov hos varje fall och en slumpmässigt vald kontroll kontrollerats, utan att kontrollanten haft vetskap om fall/kontroll-status. Ett fall med smittat första prov utesluts ur studien, och om en kontroll är smittad, tas den bort och en ny väljs slumpmässigt ut av de kvarvarande kontrollerna.

Slutligen bestod datamaterialet av endast 373 fall och 373 matchade kontroller för vilka information angående demografiska och socioekonomiska egenskaper, sexuellt och reproduktivt beteende, rökvanor, preventivmedel samt historia om gynekologiska sjukdomar och sexuellt överförbara sjukdomar, inhämtats genom personliga intervjuer.

## 6.3 Frågeställningar och analys av datamaterialet

Om man ser till hur urvalet är gjort så ser man på en gång att datamaterialet är högercensurerat och saknar competing risk, vilket gör det möjligt att använda de metoder som är föreslagna. Dessvärre har man även villkoret att det första provet för varje individ som blir intagen i studien är smittfritt. Det villkoret kommer därför att leda till att vi även får att datamaterialet är vänstertrunkerat, eftersom det blir en tidsberoende kontrollhändelse som måste vara uppfylld. Som tur är kommer det här inte att påverka metoderna särskilt mycket, utan det som händer är att räkneprocessen måste definieras om så att den även tar hänsyn till det här villkoret.

Effekten av den här restriktionen kommer alltså endast att leda till att tolkningen av reslutaten som fås blir annorlunda, p.g.a. att vi inte längre kommer att skatta rena sannolikheter, utan betingade sannolikheter m.a.p. den här restriktionen.

En ytterligare komplikation som finns i datamaterialet är att kontrollerna är matchade efter födelseår och tid då individen kom in i studien. Detta innebär att om man utgår från hela kohorten och vid varje tidpunkt $t$ som en individ upplever händelsen kommer $k-1$ st kontroller att samplas ur den sub-kohort som utgörs av de individer som är möjliga att matcha med fallet. Det man kan göra är då att dela upp studie-kohorten i olika strata med avseende på dessa matchningsfaktorer. Om man nu gör som i Langholz och Borgan [9] och antar att sannolikheten att individer kommer att dras från olika strata är den samma, så kommer man att kunna använda vikterna $n(t)/k$ som om samplingsförfarandet skulle varit helt slumpmässigt. Det här kanske inte är det bästa möjliga förfarandet, men det underlättar en hel del. Det mest önskvärda vore att kunna låta varje strata ha en individuell vikt som beskriver dess förhållande till studie-kohorten. Den stora fördelen med det här förfarandet är att den partiella likelihooden inte kommer att ändras.

Ett problem som vi har är att vi saknar information om vikterna som rör urvalsförfarandet för de matchade kontrollerna. Det kommer att påverka analysen, men inte all analys, t.ex. skattningen av regressionsparametrarna är inte beroende av den informationen. Det här problemet återkommer vi till lite längre fram där ett enkelt, men inte alltför tillfredställande, resonemang förs.

Så som data är insamlat kommer tiderna som ska modelleras att vara tiden från då en individ kommer in i studien tills dess att hon upplever händelsen eller blir censurerad.

Datamaterialet innehåller heller inga situationer där en kontroll till ett fall senare kommer att vara fall, men det är inte särskillt underligt, i och med att studiepopulationen är såpass stor.

Det som ska undersökas är hur exponeringen av HPV 16-virus påverkar risken, den relativa och den absoluta, för att en individ ska utveckla cancer. På grund av att alla individer har tagit olika många prover är det svårt att hitta ett bra mått på en individs exponering för HPV 16. Givetvis skulle det bästa vara om man kunde använda viral load vid varje registrerad tidpunkt som en tidsberoende kovariat, men problemet då blir att proverna inte rimligtvis kan ses som oberoende, vilket blir en annan faktor som måste kompenseras för. Ett sätt som förhoppningsvis är bättre än att begränsa sig till endast ett provtillfälle, men antagligen sämre än att ta med alla provtillfällen är att införa en ny förklarande variabel som är snittet av viral load från och med det första smittade provet fram till och med diagnos, samt att även ta med tiden från första smittade provet fram till och med diagnos som en förklarande variabel. På det här viset kan man nu betrakta både HPV 16-snittet och tiden från smitta till cancer som tidsoberoende.

## 6.4 Resultat

Den modell som vi nu ska titta på är Cox-modell med följande variabler:

- $hpv16$, indikator
- $hpvexp$, snittet av de exponerade proverna från och med det första smittade till och med diagnos, kontinuerlig
- $hpvtime$, tiden från första smittade provet till diagnos, kontinuerlig
- $sexcon$, antal sexualpartners per tidsenhet, kontinuerlig

Eftersom vi saknar tillgång till de vikter som beskriver förhållandet till den population som varje matchad kontroll har dragits ur, så börjar vi med den del av analysen som inte behöver tillgång till dessa, nämligen skattningen av regressionsparametrarna, test av dessa samt skattning av den relativa risken.

Skattningarna av $\boldsymbol{\beta}$ blev följande:

| | Skattning | Standardavvikelse |
|---|---|---|
| $\beta_1(hpv16)$ | 0.87 | 0.59 |
| $\beta_2(hpvexp)$ | -0.37 | 0.061 |
| $\beta_3(hpvtime)$ | -3.13e-05 | 5.87e-05 |
| $\beta_4(sexcon)$ | 1.48 | 0.69 |

Test av den globala hypotesen $H_0 : \boldsymbol{\beta} = \mathbf{0}$:

| | Teststatistika | p-värde |
|---|---|---|
| $\chi^2_W$ | 59.22 | 4.24e-12 |
| $\chi^2_{LR}$ | 195.73 | 0 |
| $\chi^2_{SC}$ | 144.33 | 0 |

och vi förkastar således nollhypotesen.
De lokala testerna av hypoteserna $\beta_i = 0$:

| | $\chi^2_W$ (p-värde) | $\chi^2_{LR}$ (p-värde) | $\chi^2_{SC}$ (p-värde) | $95\%KI_\beta$ |
|---|---|---|---|---|
| $\beta_1$ | 2.21 (0.14) | 2.37 (0.12) | 1.54 (0.21) | (-0.28, 2.02) |
| $\beta_2$ | 36.62 (1.44e-09) | 91.62 (0) | 44.63 (2.38e-11) | (-0.49, -0.25) |
| $\beta_3$ | 0.29 (0.59) | 0.29 (0.59) | 0.16 (0.69) | (-0.00015, 8.36e-05) |
| $\beta_4$ | 4.53 (0.033) | 4.83 (0.028) | 4.71 (0.030) | (0.12, 2.84) |

Vi ser nu att det bara är $\beta_2$ (*hpvexp*) och $\beta_4$ (*sexcon*) som är signifikanta på 5%-nivån. Vi kan även notera att de olika teststatistikorna ligger närmre varandra då resultaten inte är så väldigt signifikanta eller icke-signifikanta, samt att Wald- och likelihood ratio-statistikorna ligger närmre varandra än vad de gör i förhållande till statistikan för score-testet.

Den variabel som det verkar mest intressant att undersöka den relativa risken för är *hpvexp*, och då rimligtvis mellan en individ som är osmittad, dvs. $hpvexp = 50$, och t.ex. en individ med $hpvexp = 40$, givet att övriga bakgrundsvariabler är fixa. Den relativa risken blir då: $RR_{hpvexp}(40, 50) = \exp\{\hat{\beta}_2(40 - 50)\} = 40.21$ med det 95%KI: $(12.15, 133.03)$. Med andra ord är risken ca 40ggr högre för att få livmoderhalscancer om du har en hpv-exponering på 40, jämfört med om du är osmittad.

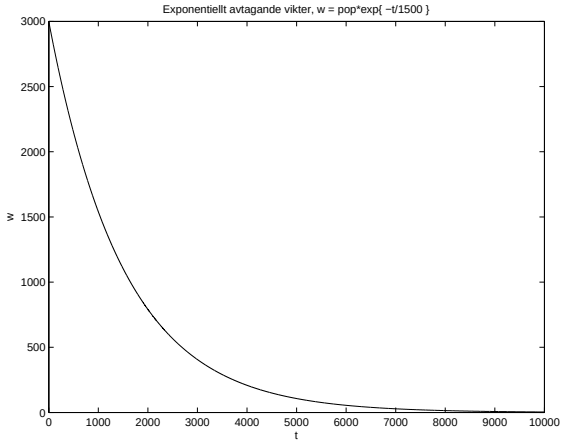
## 6.5 Vikter

Som tidigare nämnts så har vi inte tillgång till de vikter som beskriver förhållandet till den populationen som varje matchad kontroll har dragits ur. Det är ett problem, eftersom nästan alla av de metoder som vi tidigare har tittat på behöver dessa vikter för att kunna ge några resultat. Det som vi faktiskt kan göra utan de riktiga vikterna är att undersöka viktkänsligheten hos metoderna genom att välja olika vikter. Två fall som vi kan undersöka relativt lätt är dels konstanta vikter samt vikter som är exponentiellt avtagande i förhållande till tiden till diagnos utgående från någon konstant populationsstorlek enligt

$$w_i = popsize * \exp\{-t_i/1500\}$$

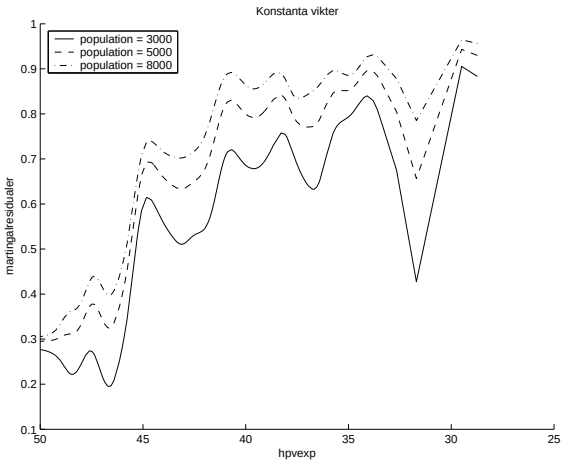
där *popsize* är konstant och $t_i$ är tiden från det att individ $i$ kommer in i studien till och med diagnos. Anledningen till att man delar tiden med 1500 är bara för att kunna få en rimlig hastighet på populationsminskningen, se fig 3.

Nu är det dags att börja undersöka modellantagandena, och det gör vi genom att titta på martingalresidualerna för de olika variablierna samt gör Andersenplottar. Martingalresidualerna som beräknas ger ett antal punkter som bör ligga någorlunda jämnt spridda kring nollan om man plottar dessa mot den variabel som man vill undersöka den funktionella formen för, om antagandet om linjaritet stämmer. Om det linjära antagandet inte verkar stämma ger dessa plottar, förhoppningsvis, en känsla för hur den funktionella formen kan tänkas se ut. För att underlätta analysen av dessa plottar bör man använda någon form av utjämning. I den analysen som presenteras här har det använts en Gaussisk utjämning.




Figur 3: Vikterna avtar exponentiellt med tiden, här är grundpopulationsstorleken = 3000.

Vi börjar med att undersöka den funktionella formen för *hpvexp*, som är den variabeln som är mest signifikant. För konstanta vikter ser vi att för större vikter trycks grafen ihop och lyfts upp en aning, men att den bibehåller formen, se fig 4.

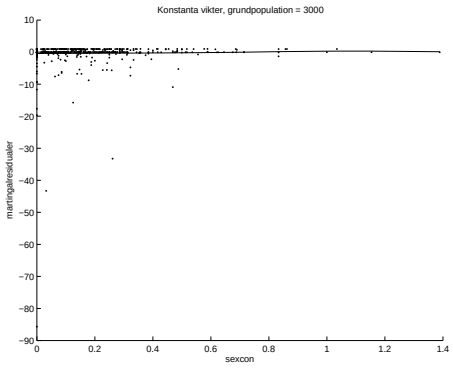



Figur 4: Här ser vi exempel på utjämnade martingalresidualer för några olika val av konstanta vikter.

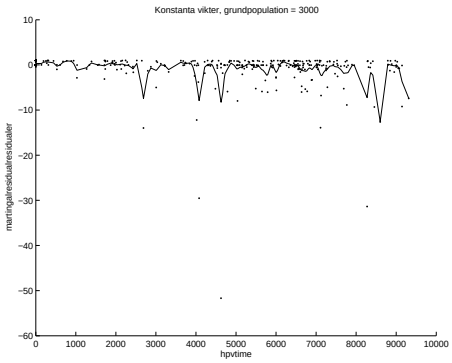
Ur graferna kan det vara svårt att se något tydligt mönster, men med lite god vilja skulle man kunna tänkas se att graferna höjs upp en aning för värden lägre än ca 45. Tanken med det här är att om vi tror på att 45 är något slags tröskelvärde för *hpvexp* så skulle vi kunna skapa en ny variabel som bara indikerar huruvida exponeringen är hög eller låg utifrån detta värde. Martingalresidualerna för *hpvtime* och *sexcon* ses i fig 5 och 6 och de ser någorlunda ok ut.

För Andersenplottarna har följande stratifiering använts för att få en någorlunda jämn fördelning i de olika strata:

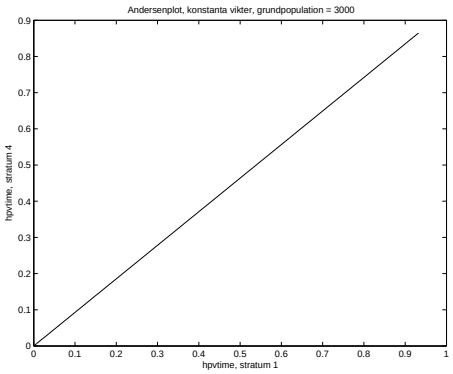
| Variabel | Stratifiering |
|---|---|
| hpv16 | indikator |
| hpvexp | 40<, 40-45<, 45-50<, 50 |
| hpvtime | 141<, 141-4500<, 4500-6500<, ≥6500 |
| sexcon | 0, <0-0.12<, 0.12-0.25<, ≥0.25 |



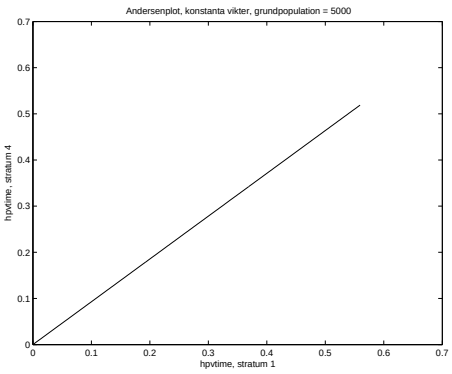

Figur 5: Martingalresidualerna för *sexcon*




Figur 6: Martingalresidualerna för *hpvtime*




Figur 7: Andersenplot för *hpvtime* då grundpopulationen är 3000.




Figur 8: Andersenplot för *hpvtime* då grundpopulationen är 5000.

Andersenplottarna följer samma mönster som martingalresidualerna, större vikter ger mer ihoptryckta grafer, se exempel i fig 7 och 8. Antagandet om proportionalitet ser ut att vara uppfyllt för alla variabler.

Om vi inför variabeln *hpvind* som är 1 om HPV 16-exponeringen är hög, 0-45, och 0 annars, och gör om analysen får vi följande parameterskattningar:

| | Skattning | Standardavvikelse |
|---|---|---|
| $\beta_1(hpv16)$ | 0.614 | 0.53 |
| $\beta_2(hpvind)$ | 2.79 | 0.44 |
| $\beta_3(hpvtime)$ | 8.79e-05 | 5.06e-05 |
| $\beta_4(sexcon)$ | 1.24 | 0.65 |

Test av den globala hypotesen $H_0 : \boldsymbol{\beta} = \mathbf{0}$:

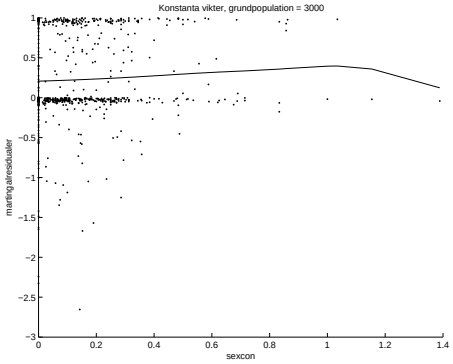
| | Teststatistika | p-värde |
|---|---|---|
| $\chi^2_W$ | 64.56 | 3.19e-13 |
| $\chi^2_{LR}$ | 180.47 | 0 |
| $\chi^2_{SC}$ | 141.96 | 0 |

och vi förkastar således nollhypotesen.
De lokala testerna av hypoteserna $\beta_i = 0$:

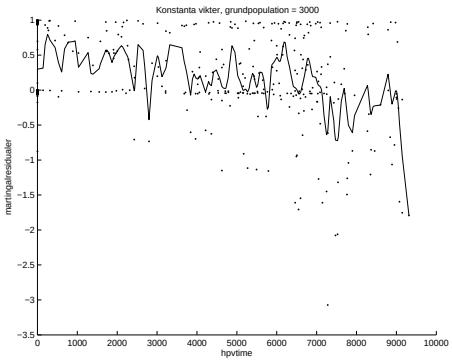
| | $\chi^2_W$ (p-värde) | $\chi^2_{LR}$ (p-värde) | $\chi^2_{SC}$ (p-värde) | $95\%KI_\beta$ |
|---|---|---|---|---|
| $\beta_1$ | 1.32 (0.25) | 1.38 (0.24) | 0.96 (0.33) | (-0.43, 1.66) |
| $\beta_2$ | 40.43 (2.04e-09) | 76.36 (0) | 46.21 (1.06e-11) | (1.93, 3.65) |
| $\beta_3$ | 3.02 (0.082) | 3.12 (0.078) | 2.13 (0.14) | (-1.12e-05, 0.00019) |
| $\beta_4$ | 3.61 (0.057) | 3.72 (0.054) | 3.68 (0.055) | (-0.038, 2.52) |

Ur tabellen ovan ser vi att *hpvind* är starkt signifikant, men till skillnad mot tidigare är *sexcon* ej signifikant på 5%-nivån, men däremot börjar *hpvtime* närma sig signifikans på 5%-nivån.

Den relativa risken för om du har en låg HPV-exponering, $hpvind = 0$ jämfört med om du har en hög blir nu: $RR_{hpvind}(1,0) = \exp\{\hat{\beta}_2\} = 16.28$ med 95%KI: $(6.89, 38.48)$, dvs. en högexponerad individ har ca 16ggr högre risk än en lågexponerad att utveckla livmoderhalscancer.




Figur 9: Martingalresidualerna för *sexcon*, med *hpvind*.




Figur 10: Martingalresidualerna för *hpvtime*, med *hpvind*.

Vad gäller martingalresidualerna ser vi i fig 9 och 10 hur det ser ut för *sexcon* och *hpvtime*, och det verkar vara ok. Eventuellt skulle kan man kanske kunna göra någon slags transformering av variabeln *hpvtime*, eftersom det ser ut att finnas en avtagande tendens i grafen, men det är något svårttolkat. Andersenplottarna visar sig också vara ok.

En alternativ analys som vi kan göra om vi inte tror på vårt antagande om att det finns något tröskelvärde för *hpvexp* är att ersätta *hpvexp* med $hpv1$, där $hpv1$ endast är HPV 16-exponeringen vid det första mättillfället. En sådan analys ger följande parameterskattningar:

| | Skattning | Standardavvikelse |
|---|---|---|
| $\beta_1(hpv16)$ | 0.71 | 0.42 |
| $\beta_2(hpv1)$ | -0.016 | 0.018 |
| $\beta_3(hpvtime)$ | 0.00026 | 4.51e-05 |
| $\beta_4(sexcon)$ | 1.41 | 0.59 |

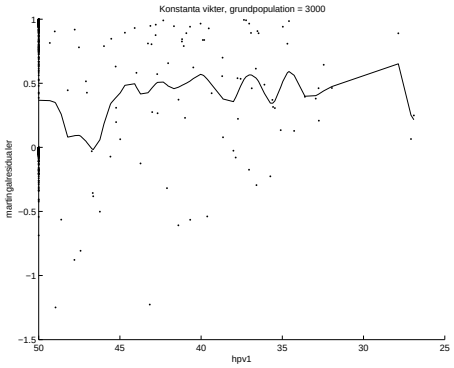
Test av den globala hypotesen $H_0 : \boldsymbol{\beta} = \mathbf{0}$:

| | Teststatistika | p-värde |
|---|---|---|
| $\chi^2_W$ | 63.21 | 6.13e-13 |
| $\chi^2_{LR}$ | 104.92 | 0 |
| $\chi^2_{SC}$ | 89.54 | 0 |

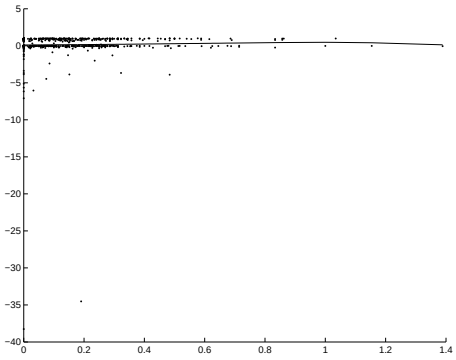
och vi förkastar således nollhypotesen.

De lokala testerna av hypoteserna $\beta_i = 0$:

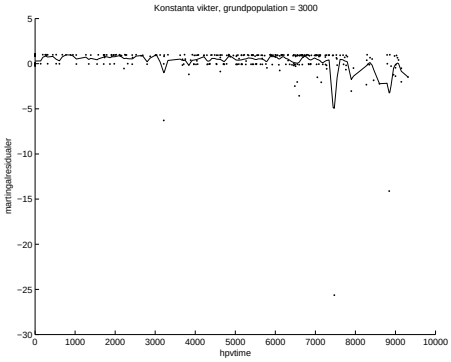
| | $\chi^2_W$ (p-värde) | $\chi^2_{LR}$ (p-värde) | $\chi^2_{SC}$ (p-värde) | $95\% KI_\beta$ |
|---|---|---|---|---|
| $\beta_1$ | 2.83 (0.093) | 3.02 (0.082) | 1.98 (0.16) | (-0.12, 1.53) |
| $\beta_2$ | 0.79 (0.37) | 0.81 (0.37) | 0.73 (0.39) | (-0.051, 0.019) |
| $\beta_3$ | 33.84 (5.97e-09) | 44.15 (3.039e-11) | 31.62 (1.88e-08) | (0.00017, 0.00035) |
| $\beta_4$ | 5.80 (0.016) | 5.99 (0.014) | 6.10 (0.014) | (0.26, 2.56) |




Figur 11: Martingalresidualerna för *hpv*1



Figur 12: Martingalresidualerna för *sexcon*

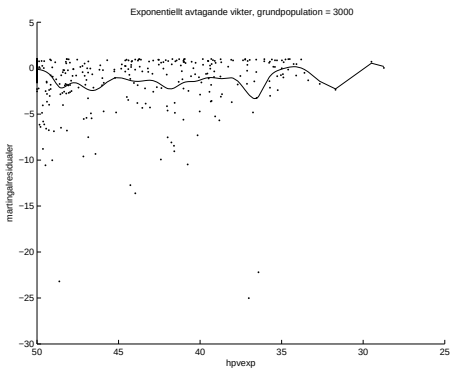



Figur 13: Martingalresidualerna för *hpvtime*

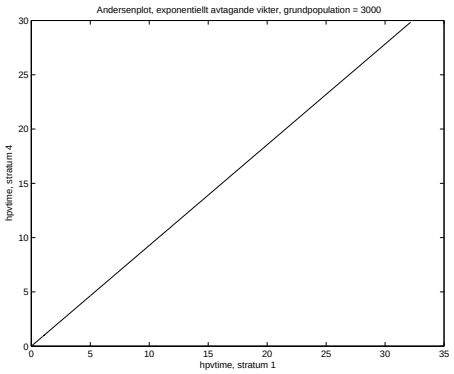
Nu har det hänt en hel del. Variabeln som har med HPV 16-exponeringen är inte längre signifikant, men det är däremot *hpvtime* och *sexcon*. Martingalresidualerna har ett liknande utseende med dom från den första analysen, se fig 11-13, och därmed är det svårt att veta hur man ska bete sig. Antagligen skulle vi kunna göra på samma sätt som vi gjorde för *hpvexp* även med *hpv*1 genom att införa en indikatorvariabel som endast beskriver hög eller låg exponering, men det känns då rimligare att göra det för *hpvexp*, eftersom den variabeln innehåller mer information. Andersenplottarna ser ut ungefär som för den första analysen, så antagandet om proportionalitet är ok. Problemet med hur man ska göra med HPV 16-exponeringsvariablen kvarstår dock.

Om vi gör samma analys som ovan, men med vikter som avtar exponentiellt i förhållande till tiden från då en individ kommer in i studien tills dess att hon får diagnos, så kommer vi att se samma mönster som för fallet med konstanta vikter, nämligen att för martingalresidualerna blir graferna mer ihoptryckta med större vikter, dvs. med större grundpopulation, samt att dom höjs upp en aning. Om vi ser till martingalresidualerna för *hpvexp* så ser vi att det skulle kunna vara möjligt att införa en indikatorvariabel även här, men nu med ett tröskelvärde runt 47, se fig 14. För *hpvtime* och *sexcon* har residualerna liknande utseende som tidigare.

För Andersenplottarna verkar det inte heller uppstå några problem, se ex. i fig 6.5.




Figur 14: Martingalresidualerna för *hpvexp*




Figur 15: Andersenplot för *hpvtime*

Om vi ersätter variabeln *hpvexp* med indikatorvariabeln *hpvind* som är 1 i intervallet 0-47 och 0 annars, och gör om analysen, får vi följande parameterskattningar:

| | Skattning | Standardavvikelse |
|---|---|---|
| $\beta_1(hpv16)$ | 0.54 | 0.49 |
| $\beta_2(hpvind)$ | 2.17 | 0.35 |
| $\beta_3(hpvtime)$ | 6.73e-05 | 5.36e-05 |
| $\beta_4(sexcon)$ | 1.12 | 0.63 |

Test av den globala hypotesen $H_0 : \boldsymbol{\beta} = \mathbf{0}$:

| | Teststatistika | p-värde |
|---|---|---|
| $\chi^2_W$ | 75.85 | 1.33e-15 |
| $\chi^2_{LR}$ | 158.90 | 0 |
| $\chi^2_{SC}$ | 131.50 | 0 |

och vi förkastar således nollhypotesen.
De lokala testerna av hypoteserna $\beta_i = 0$:

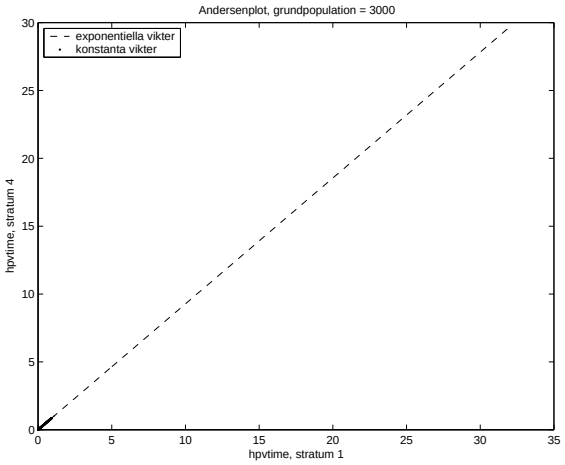
| | $\chi^2_W$ (p-värde) | $\chi^2_{LR}$ (p-värde) | $\chi^2_{SC}$ (p-värde) | $95\%KI_\beta$ |
|---|---|---|---|---|
| $\beta_1$ | 1.20 (0.27) | 1.25 (0.26) | 0.84 (0.36) | (-0.42, 1.50) |
| $\beta_2$ | 39.07 (4.10e-10) | 54.80 (1.34e-13) | 30.90 (2.71e-08) | (1.49, 2.85) |
| $\beta_3$ | 1.58 (0.21) | 1.60 (0.21) | 0.93 (0.33) | (-3.77e-05, 0.00017) |
| $\beta_4$ | 3.15 (0.076) | 3.21 (0.073) | 3.18 (0.075) | (-0.12, 2.35) |

Den enda variabeln som nu är signifikant är *hpvind*. Den relativa risken för en individ med $hpvind = 1$ mot en med $hpvind = 0$ blir nu $RR_{hpvind}(1, 0) = 8.78$ och 95%KI: $(4.44, 17.36)$.

Martingalresidualerna för *sexcon* och *hpvtime* följer samma mönster som för de konstanta vikterna. Problemet från analysen med de konstanta vikterna finns dock kvar när det gäller *hpvtime*, finns det någon trend eller inte? Nu ser Andersenplottarna bättre ut, och antagandet om proportionalitet är ok.

Om vi resonerar på samma sätt som tidigare som när vi använde konstanta vikter, och anser att det inte är godtagbart att ersätta *hpvexp* med *hpvind*, utan istället gör om analysen med variablen $hpv1$, så kommer resultaten att bli väldigt snarlika, och inga andra slutsatser än de som kunde dras då kan dras nu.

Vad kan man då säga om vikternas inverkan? För både martingalresidualerna och Andersenplottarna verkar vikterna ha en markant inverkan. För martingalresidualerna får vi liknande mönster för båda val av vikter, vilket även gäller för Andersenplottarna, se. fig. 16. Och utifrån

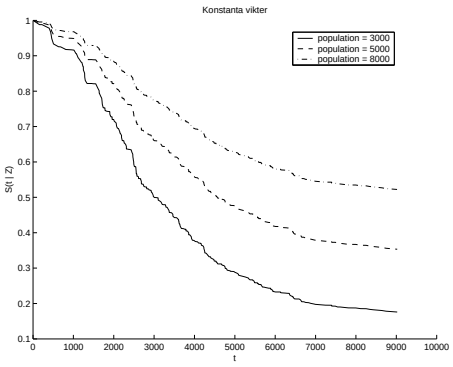



Figur 16: Andersenplot för *hpvtime* med både konstanta och exponentiellt avtagande vikter

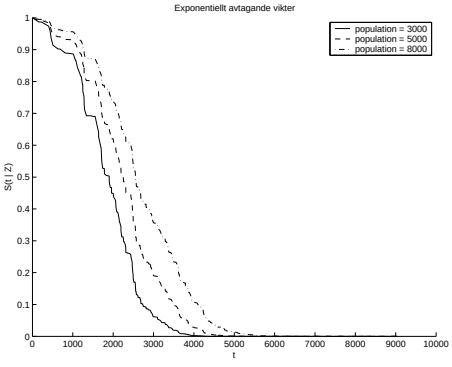
den här analysen är det svårt att dra så många fler slutsatser än så.

## 6.6 Skattning av överlevnadsfunktionen

Eftersom de sanna vikterna saknas, behöver vi inte lägga någon större vikt på den modellanalys som gjordes här ovan, utan kan anta att modellen med *hpv*16, *hpvexp*, *hpvtime* och *sexcon* är den rätta.



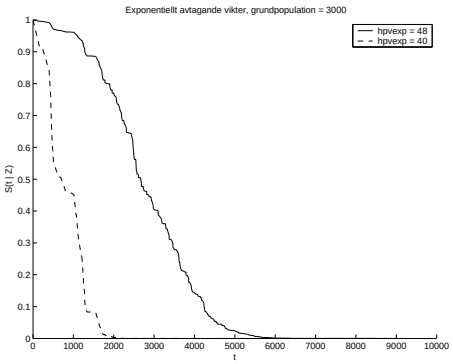

Figur 17: Överlevnadsfunktioner för olika val av konstanta vikter då $hpv16 = 1$, $hpvexp = 45$, $hpvtime = 3500$ och $sexcon = 0.7$



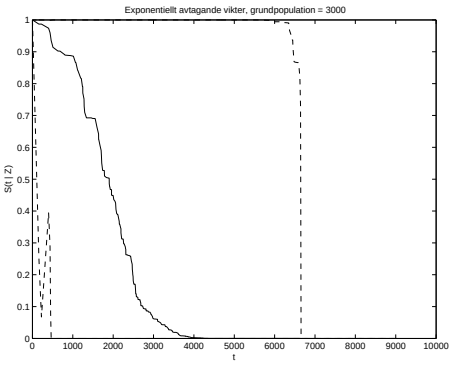

Figur 18: Överlevnadsfunktioner för olika val av exponentiellt avtagande vikter då $hpv16 = 1$, $hpvexp = 45$, $hpvtime = 3500$ och $sexcon = 0.7$

I fig 17 och 18. ser vi hur några överlevnadskurvor kan tänkas se ut för olika val av vikter, då kovariaterna hålls fixa. Vi ser att för de exponentiellt avtagande vikterna går överlevnadsfunktionen snabbt mot noll medan överlevnadsfunktionen ser ut att plana ut för de konstanta vikterna. Kom ihåg att Andersenplottarna urartade för långa observationstider då vi hade exponentiellt avtagande vikter, det beror just på dessa vikters beteende för överlevnadsfunktionen, eftersom Andersenplottarna bygger på sambandet $H(t) = -\log S(t)$.

Om man undersöker olika kovariaters inverkan på överlevnaden, visar det sig att det är *hpvexp* som är den som påverkar mest, vilket känns rimligt i och med att det är den variabeln som är




Figur 19: Överlevnadsfunktion för exponentiellt avtagande vikter då $hpv16 = 1$, $hpvtime = 3500$ och $sexcon = 0.7$ är fixa.



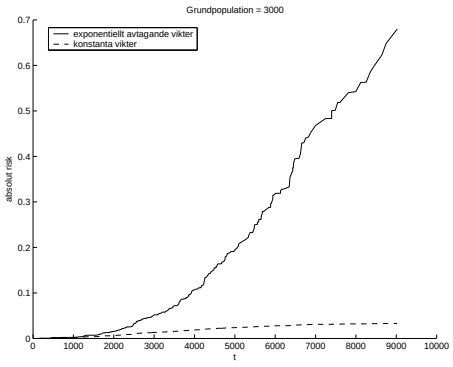

Figur 20: Överlevnadsfunktion för exponentiellt avtagande vikter då $hpv16 = 1$, $hpvexp = 45$, $hpvtime = 3500$ och $sexcon = 0.7$, med tillhörande 95% Hall-Wellner band

mest signifikant, se fig 19.

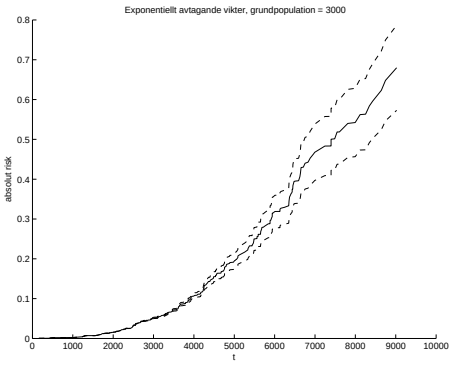
I fig. 20 finns ett exempel på en överlevnadskurva med tillhörande Hall-Wellner band. Vi ser att det övre bandet går väldigt snabbt mot ett, medan det undre går mot noll, men detta är inget märkligt, eftersom banden ska innefatta hela överlevnadskurvan med 95% sannolikhet. Det som till synes är något märkligt är att den övre gränsen gör en brant dykning ner till noll då tiden närmar sig 6000, men det beror på att skattningen av överlevnadsfunktionen är *lika* med noll för punkter bortom den tiden. Detta innebär inte att variansen är noll, utan bara att beräkningen av konfidensbanden har havererat, ty banden beräknas enligt $\hat{S}(t \mid \mathbf{Z})^{1/\theta}$ och $\hat{S}(t \mid \mathbf{Z})^{\theta}$.

## 6.7 Skattning av den absoluta risken

Om vi använder samma modell som ovan, så kommer vi få kurvor som beter sig på ett snarlikt sätt som för de olika valen av vikter gjorde för överlevnadsfunktionen, dvs. för konstanta vikter verkar den absoluta risken plana ut med tiden, men för exponentiellt avtagande vikter så accelererar risken med tiden, se fig 21. I fig 22. ser vi ett exempel på hur den absoluta risken beter sig över tiden, för en neutral individ samt tillhörande punktskattningar.




Figur 21: Den absoluta risken för en neutral individ, dvs. då $hpv16 = 0$, $hpvexp = 50$, $hpvtime = 10000$ och $sexcon = 0.0$, för olika val av vikter.




Figur 22: Den absoluta risken över tiden för en neutral individ med tillhörande 95% punktskattningar.

## 6.8 Sammanfattning

För att sammanfatta resultaten lite kort, så har vi sett att vikterna har en tydlig inverkan på de olika metoderna. För de konstanta vikterna så uppstår det inga större problem med någon av metoderna, medan de exponentiellt avtagande vikterna gör att det kan uppstå problem med att överlevnadsfunktionen snabbt går mot noll och att den absoluta risken ökar kraftigt. Vi har även undersökt några tänkbara omparametriseringar, för olika val av vikter, för att visa principen, men inga relevanta slutsatser kan egentligen dras med hjälp av dessa. Vad gäller skattningar av regressionsparametrar har vi sett att den enda kovariaten som är klart signifikant är den som har att göra med HPV-exponeringen. Den relativa risken för en högexponerad individ mot en lågexponerad, alternativt, en oexponerad mot en exponerad tyder på att de HPV 16-exponerade individerna har en kraftigt ökad risk att utveckla livmoderhalscancer.

# Referenser

[1] Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). Statistical Models Based on Counting Processes. Springer-Verlag: Springer Series in Statistics.

[2] Borgan, Ø. and Langholz, B. Nonparametric Estimation of Relative Mortality from Nested Case-Control Studies. *Biometrics 49, 593-602, June 1993.*

[3] Borgan, Ø., Goldstein, L and Langholz, B. Methods for the Analysis of Sampled Cohort Data in the Cox Proportional Hazards Model *The Annals of Statistics, 1995, Vol. 23, No. 5, 1749-1778.*

[4] Djehiche, B. (2000). Stochastic Calculus. An Introduction with Applications. Department of Mathematical Statistics, KTH, Stockholm, Sweden.

[5] Fleming, T. R. and Harrington, D. P. (1991). Counting Processes and Survival Analysis. Wiley Series in Probability and Mathematical Statistics.

[6] Gut, A. (1995). An Intermediate Course in Probability. Springer-Verlag.

[7] Jacod, J. and Protter, P. (2002). Probability Essentials, 2 ed. Springer Verlag: Universitext Series.

[8] Klein, J. P. and Moeschberger, M. L. (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer: Statistics for Biology and Health.

[9] Langholz, B. and Borgan, Ø. Estimation of Absolute Risk from Nested Case-Control Data. *Biometrics 53, 767-774, June 1997.*

[10] Ylitalo, N. (2000). Human Papillomavirus and Cervical Carcinoma in Situ: Implications for Future Screening. Department of Medical Epidemiology, Karolinska Institutet, Stockholm, Sweden.