Stockholms universitet

Modell för prissättning av sjukförsäkring

Rikard Hellman

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2012:4
Matematisk statistik
Maj 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2012:4**,
http://www.math.su.se/matstat

Modell för prissättning av sjukförsäkring

Rikard Hellman*

Maj 2012

Sammanfattning

I Sverige omfattas samtliga förvärvsarbetande anställda av den allmänna sjukförsäkringen som ersätter upp till cirka 80 % av inkomsten vid sjukdom. Som ett tillägg till denna kan man teckna privat sjukförsäkring som ersätter ytterligare. En viktig del i försäkringsbolagens arbete är att sätta rätt premie för privat tecknad sjukförsäkring. Premien ska återspegla risken för insjuknande, avveckling samt dödsfall. Ofta bestäms denna premie utifrån en persons kön och ålder. Med statistik från Försäkringskassan för den allmänna sjukförsäkringen tar vi fram en premieberäkningsmodell för privat sjukförsäkring som förutom kön och ålder även tar hänsyn till inkomstnivå och bostadsregion. Premieberäkningsmodellen delas upp i två delar, avvecklingsfunktion och insjuknandesannolikhet. På grund av skillnader i riskerna mellan beståndet i den allmänna sjukförsäkringen och de bestånd som återfinns inom privat försäkring bygger vår modell på de relativa effekter parametrarna har på avveckling respektive insjuknande. En utvidgad Cox-hazardmodell visar sig bäst beskriva avvecklingen utifrån duration, ålder, kön, bostadsregion och inkomstnivå som samspelar med ålder, kön och duration. Modellen visar att högavlönade avvecklas fortare än lågavlönade och att det något oväntat är större sannolikhet att avvecklas om man inte är bosatt i storstadsregioner. Modellen för insjuknande baseras på den relativa insjuknandesannolikheten som påverkas av ålder, kön och inkomstnivå som samspelar med ålder. De framtagna modellerna för avveckling och insjuknande binds till sist samman till en premieberäkningsmodell. I ett test jämförs den framtagna modellen mot en fiktiv premie baserad på Försäkringskassans data. Testet visar högst premiepåslag för lågavlönade män i åldrarna 30-54 år för vilka premien ökar med upp till 35 % mot den fiktiva premien. Lägst premie syns hos högavlönade kvinnor i 30-årsåldern där premiemodellen antyder en premie cirka 60 % lägre än den fiktiva.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: r.hellman@hotmail.com. Handledare: Andreas Nordvall Lagerås.

1 Förord och tack

Detta arbete utgör en masteruppsats i försäkringsmatematik om 30 högskolepoäng på Stockholms universitet. Arbetet är utfört i samarbete med min arbetsgivare Salus Ansvar. Data som används i arbetet kommer från Försäkringskassan.

Jag vill rikta ett stort tack till mina två handledare Andreas Nordvall Lagerås, docent i matematisk statistik vid Stockholms Universitet, och Jan Åke Persson, aktuarie PWC (tidigare chefaktuarie Salus Ansvar). Deras vägledning, kunskap och agerande som bollplank har varit ovärderliga för detta arbete. Jag vill även rikta ett stort tack till Eva De Val, chefaktuarie Salus Ansvar, vars uppmuntran varit ett enormt stöd under arbetets gång. Ett stort tack även till Bertil Thorslund, Johan Grum, Marie Mulder och Eva-Lo Ighe på Försäkringskassan för deras hjälp med att ta fram det data som används i arbetet. Slutligen riktas ett stort tack till mina vänner, familj och inte minst min sambo var stöd och uppmuntran gett mig kraft att färdigställa uppsatsen.

Innehåll

1 Inledning

En sjukförsäkring är en försäkring där försäkringstagaren får ersättning vid arbetsoförmåga på grund av sjukdom eller olycka. Ersättningen är vanligen en periodisk utbetalning som är tänkt att täcka inkomstbortfall. Men ersättning kan även utbetalas som ett engångsbelopp, en så kallad sjukkapitalförsäkring, om den försäkrade konstateras vara kvarstående långvarigt sjuk, det vill säga om den försäkrade inte förväntas komma tillbaka till arbete.

I Sverige har vi en allmän sjukförsäkring som innebär att samtliga förvärvsarbetande anställda i Sverige har en obligatorisk sjukförsäkring som förvaltas av staten. Den allmänna sjukförsäkringen ersätter upp till ca 80% av den sjukpenninggrundande inkomsten[1]. Utöver detta kan man teckna en privat sjukförsäkring hos ett försäkringsbolag som ger ytterligare ersättning. En person som tjänar mer än 7,5 prisbasbelopp, vilket under 2012 är 330 000 kr, kan alltså få ut en större del av sin lön om han eller hon vid sidan av den allmänna sjukförsäkringen även har tecknat en privat sjukförsäkring. För att ingen skall tjäna på att vara sjuk finns en utbredd branchpraxis som säger att en persons ersättning får uppgå till maximalt 90% av förvärvsinkomsten.

En viktig del i försäkringsbolagens arbete är att sätta rätt premie på den privata sjukförsäkringen. Premien ska återspegla risken, ersättningsnivån samt de kostnader som är förknippade med försäkringen. Risken i en sjukförsäkring består av tre komponenter, nämligen risken att bli sjuk, insjuknanderisken, risken att vara sjuk under en viss tid, avvecklingsrisken, samt risken att avlida, dödsfallsrisken. Att vara sjuk definieras i den här uppsatsen som att man från Försäkringskassan får ersättning på grund av arbetsoförmåga. Många försäkringsbolag skattar ovanstående risker utifrån sina egna försäkringsbestånd som för de mindre aktörerna ofta är begränsade till storleken vilket gör dessa skattningar mer eller mindre osäkra.

Denna uppsats syftar till att med hjälp av statistik från Försäkringskassan för den allmänna sjukförsäkringen skatta riskerna för insjuknande och avveckling för att utveckla premieberäkningsmodeller som är applicerbara på privat sjukförsäkring. Statistiken innehåller näst intill samtliga förvärvsarbetande anställda i Sverige och motsvarar således nästan hela den potentiella kundpopulationen för privat sjukförsäkring i Sverige. Förutom de vanliga parametrarna i premiemodeller kön och ålder är statistiken uppdelad på inkomst och vart man är bosatt. Huvuddelen av arbetet handlar om hur dessa parametrar påverkar avveckling och insjuknande. Förhoppningen är att en parametrisering av premieberäkningsmodellerna utifrån ovanstående parametrar kommer att resultera i bättre underlag vid premiesättning.

[1]Den sjukpenninggrundande inkomsten bestäms som det minsta av årslönen och inkomsttaket, 7,5 prisbasbelopp (1 prisbasbelopp 2012 = 44 000 kr).

Det är en allmänt accepterad sanning att försäkringsbolagens bestånd är bättre ur risksynpunkt än befolkningen i stort. Anledningen till detta är att försäkringsbolagen undviker att försäkra personer med höga risker genom att till exempel tillämpa hälsoprövning vid nyteckning. Vi skall undersöka sanningshalten i detta påstående genom jämförelser mellan beståndet i den allmänna sjukförsäkringen och vårt eget, Salus Ansvars, försäkringsbestånd.

Under sommaren 2008 infördes nya sjukskrivningregler för den allmänna sjukförsäkringen. Vi kommer att undersöka hur detta påverkat insjuknade och avveckling och hur detta påverkar premiemodellerna i privat sjukförsäkring.

2 Försäkringskassan

För att förstå datamaterialet som behandlas i denna uppsats krävs en inblick i regelverket kring den allmänna sjukförsäkringen. I detta avsnitt går vi igenom det nu gällande regelverket. I stort finns det två typer av ersättningar, sjukpenning som ger en tidsbegränsad ersättning samt sjuk- och aktivitetsersättning som ges om sjukperioden bedöms vara livsvarig. Nedan beskrivs först reglerna för sjukpenning och därefter för sjuk- och aktivitetsersättning.

2.1 Sjukpenning

När en person blir sjuk får han de 14 första dagarna ersättning från sin arbetsgivare (exklusive den första dagen som är en karensdag). Därefter görs en anmälan om sjukdom till försäkringskassan. Från försäkringskassan får man då vanligtvis sjukpenning. Sjukpenning kan i normalfallet fås i upp till 364 dagar under en period av 450 dagar. Sjukpenning ger en ersättning motsvaranade ca 80% av den sjukpenninggrundande inkomsten (SGI). SGI bestäms som personens lön eller maximalt 7,5 prisbasbelopp (ett prisbasbelopp 2012 är 44 000 kr). Om en person är sjuk längre än 364 dagar kan försäkringskassan bevilja förlängd sjukpenning i upp till 550 dagar. Under speciella omständigheter kan den förlängda sjukpenningen sakna tidsgräns. Ersättningen vid förlängd sjukpenning uppgår till strax under 75% av SGI.

Då sjukpenningdagarna tagit slut erbjuds arbetslivsintroduktion hos arbetsförmedlingen vilket i praktiken innebär att arbetsförmedlingen gör en utredning om behovet av fortsatt stöd.

2.2 Sjukersättning och Aktivitetsersättning

Sjukersättning är en ersättning för personer mellan 30 och 64 år som troligtvis aldrig mer kommer att kunna arbeta heltid. Full ersättning motsvarar 64% av antagandeinkomsten (den inkomst man troligen skulle ha haft om man arbetade) eller minst den så kallade garantiersättningen[2]. Innan den första juli 2008 fanns tidsbegränsad sjukersättning som nu är avskaffad. Personer som hade tidbegränsad sjukersättning vid övergången till de nya sjukreglerna i juli 2008 behåller sin tidsbegränsade ersättning till dess att den löper ut. Efter den tidsbegränsade perioden kan man ansöka om förlängd tidsbegränsad ersättning i upp till 18 månader, dock längst t.o.m. 2012.

Aktivitetsersättning är en ersättning för personer mellan 19 och 29 år som inte beräknas komma tillbaka till arbete på minst ett år på grund av sjukdom eller annan funktionsnedsättning. Ersättningsnivån uppgår även här till maximalt 64% av antagandeinkomsten eller som minst garantiersättningen.

Både sjuk- och aktivitetsersättning kan beviljas för hel, tre fjärdedels, halv eller en fjärdedels ersättning beroende på förmågan att försörja sig genom arbete. Personer som inte har hel sjuk- eller aktivitetsersättning och har inkomst av arbete har möjlighet att få sjukpenning om de blir mer sjuka, det vill säga man kan få både sjukpenning och sjuk- eller aktivitetsersättning samtidigt.

Om försäkringskassan bedömer att personer med sjukpenning troligtvis aldrig kommer att bli arbetsföra igen kan sjukpenningen övergå till sjukersättning. Samma sak gäller personer med sjukpenning som uppfyller kraven för aktivitetsersättning.

3 Beskrivning av data

Genom ett samarbete med Försäkringskassan har vi erhållit data över samtliga sjukfall i den allmänna sjukförsäkringen som inträffat från 2003 till och med andra kvartalet 2010. Dessutom har vi erhållit data över det försäkrade beståndet i den allmänna sjukförsäkringen. Vi kommer nedan att redovisa för dessa två dataset och syftet med dem.

3.1 Sjukfall

Att vara sjuk definieras i den här uppsatsen som att man av Försäkringskassan får ersättning på grund av arbetsoförmåga, det kan vara antingen

[2]Garantiersättning baseras på ålder och hur länge man bott i Sverige.

sjukpenning eller sjukersättning. Som vi nämnt tidigare ersätter sjukpenning upp till ca 450 dagar och om man därefter är fortsatt sjuk kan man erhålla sjukersättning. Sjukersättning kan även ges tidigare om man förväntas vara livsvarigt sjuk. Försäkringskassan delar vanligvis upp data på de två ersättningstyperna. I den här uppsatsen har vi däremot valt att definera ett sjukfall som att vara sjuk enligt definitionen ovan, det vill säga vi särskiljer inte sjukpenning från sjukersättning utan ser till en persons sammalagda tid som sjuk. Data innehåller samtliga sjukfall som påbörjats under perioden 2003-01-01 till 2010-09-30. Utsökningen av data gjordes 2010-11-30 vilket innebär att vi inte har någon information om sjukfallens utveckling efter denna tidpunkt. Sjukfallen är uppdelade på parametrarna ålder, kön, inkomstnivå, bostadsort samt kvartal då personen insjuknat. Försäkringskassan får enligt lag endast lämna ut aggregerad data där varje post måste inkludera tre eller fler individer. Detta har medfört att vi varit tvungna att göra grövre indelningar av parametrarna än vad som varit önskvärt. Nedan följer en beskrivning av den erhållna datatabellen.

Sjukfall	
Kolumnnamn	**Beskrivning**
Kvartal	Kvartal då personerna insjuknade, ex. 2003Q2.
Region	Anger om personerna är bosatta i storstadregioner eller inte[3]
Åldersgrupp	Åldersgrupper i 5-års intervall, $20-24, ..., 60-64$ år.
Kön	Kön på personerna i gruppen.
Inkomstnivå	Låg- eller höginkomsttagare, $< 240\ 000$ kr/år eller $\geq 240\ 000$ kr/år.
Duration	Antalet dagar sjukfallen varat uppdelat i 20-dagarsintervall upp till 360 dagar, $0-20, ..., 341-360, +360$.
Antal sjukfall	Antalet sjukfall i respektive grupp, om antal sjukfall < 3 lämnas fältet tomt.
Antal sjukdagar	Antalet sjukdagar totalt inom gruppen, om antal sjukfall < 3 lämnas fältet tomt.

Tabell 1: Sjukfall

Tabellen innehåller uppgifter med vilka vi har möjlighet att skatta avvecklingen för sjukfallen uppdelad på de olika parametrarna. Vi har ingen exakt duration på varje sjukfall men vi har möjlighet att beräkna medelduration inom varje grupp. Eftersom sjukfallen även är uppdelade på kvartal har vi möjlighet att undersöka om/hur de nya sjukskrivningreglerna som infördes

[3]Definitionen för storstadsregion är hämtad från Tillväxtverkets regionsfamiljer[2]. Storstadsregioner inkluderar regionerna Stockholm, Göteborg och Malmö.

den 1 juli 2008 har påverkat avvecklingen.

Vi får ett visst bortfall på grund av de sekretesslagar vi talat om tidigare. Totalt består tabellen ovan av 3 928 765 sjukfall och det är 3 991 fall som inte kommer med på grund av att de tillhör grupper med för få sjukfall, vilket ger ett bortfall på ungefär 0,10%. Det finns flera användbara metoder för att ta hänsyn till data som saknas varav imputation[1] är en av de vanligaste metoderna. Bortfallet av data i vårt fall är dock av sådan karaktär att dessa metoder inte på ett enkelt sätt kan appliceras. Detta i kombination med att bortfallet endast är 0,10% gör att vi tillämpar en metod som helt enkelt utesluter de poster som innehåller färre än tre sjukfall.

Med data över inträffade sjukfall i kombination med data över det försäkrade beståndet som beskrivs närmare nedan kan vi även skatta insjuknandesannolikheten.

3.2 Försäkrat bestånd

Samtliga arbetsföra individer i Sverige har rätt till ersättning från Försäkringskassan om de på grund av sjukdom inte kan arbeta. Det försäkrade beståndet innefattar därmed alla personer som har rätt till ersättning vid inträffat sjukfall. Vi har från Försäkringskassan fått uppgifter om detta försäkringsbestånd för åren 2003 fram till 2010, där vi valt att ta beståndet som det såg ut per den sista december varje år. Data är uppdelat på år som beståndet avser, åldersgrupp, kön, bostadsort samt inkomstnivå. Precis som tidigare beskrivs den erhållna datatabellen nedan.

Försäkrat Bestånd	
Kolumnnamn	**Beskrivning**
År	År som beståndet avser, bestånd per 2003-12-31 avser år 2003 osv.
Region	Anger om personerna är bosatta i storstadregioner eller inte
Åldersgrupp	Åldersgrupper i 5-års intervall, $20-24, ..., 60-64$ år.
Kön	Kön på personerna i gruppen.
Inkomstnivå	Låg- eller höginkomsttagare, $<$ 240 000 kr/år eller $\geq$ 240 000 kr/år.
Antal försäkrade	Antalet försäkrade i respektive grupp.

Tabell 2: Försäkrat bestånd

Med tabellerna ovan har vi nu möjlighet att skatta både avveckling och insjuknande för den allmänna sjukförsäkringen. Vi har även möjlighet att undersöka eventuella skillnader mellan kön, ålder, inkomstnivå och bostads-

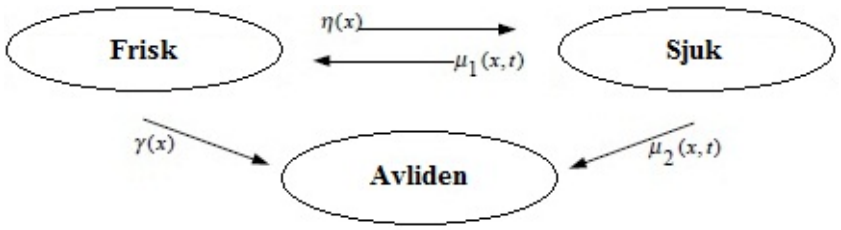
ort. Detta hjälper oss när vi nu ska utveckla de premieberäkningsmodeller som vi sedan kan använda för privat sjukförsäkring.

Även om det försäkrade beståndet inte är uppdelat på kvartal som sjukfallen ovan har vi möjlighet att undersöka hur även insjuknande har påverkats av de nya sjukskrivningreglerna. Men här blir vi då tvungna att använda oss av försäkringsår istället för kvartal.

4 Teori

Vid beräknandet av premien för en sjukförsäkring tar man hänsyn till risken för att bli sjuk, hur länge man förväntas vara sjuk, storleken på den ersättning man får vid sjukdom, risken att avlida under försäkringsperioden samt de kostnader försäkringsbolaget har för att administrera försäkringen. I den här uppsatsen bortser vi emellertid från de kostnader som uppstår i samband med administrationen och fokuserar endast på den så kallade aktuariella premien som beskriver försäkringens risker i relation till ersättningsnivån.

En persons hälsotillstånd kan beskrivas med hjälp av semi-Markovprocesser. En semi-Markovprocess är en stokastisk process som beskriver övergången mellan olika tillstånd där övergången beror både på nuvarande tillstånd och den spenderade tiden i nuvarande tillstånd. För de sjukfall vi studerar i den här uppsatsen gäller att en person kan befinna sig i tre olika tillstånd, nämligen frisk, sjuk eller avliden. En schematisk beskrivning av de tre möjliga tillstånden kan se ut på följande sätt



Figur 1: Schematisk process.

där
$\eta(x)$ - Insjuknandeintensiteten
$\mu_1(x,t)$ - Intensiteten med vilken en sjuk person blir frisk
$\mu_2(x,t)$ - Intensiteten med vilken en sjuk person avlider
$\gamma(x)$ - Dödlighetsintensiteten för en frisk person.

Variabeln x representerar personens ålder vid insjuknande eller dödsfall och t representerar den tid personen varit sjuk. Vanligtvis brukar man inte skilja

på $\mu_1(x,t)$ och $\mu_2(x,t)$ utan definerar istället $\mu(x,t) = \mu_1(x,t) + \mu_2(x,t)$. Funktionen $\mu(x,t)$ kallas avvecklingsintensiteten och är den intensitet med vilken en person går från att vara sjuk till endera av de två andra tillstånden.

Innan vi beskriver premieberäkningen definierar vi även sannoliketen att en person som blivit sjuk vid åldern x fortfarande är sjuk efter tiden t.

Definition Låt T och X vara två stokastiska variabler. Sannolikheten att fortfarande vara sjuk efter tiden t givet att man blivit sjuk vid ålder x definieras då som

$$\lambda(x,t) = P\left(T > t | X = x\right) = 1 - P\left(T \leq t | X = x\right) = 1 - F(t|x)$$

där $F(t|x)$ är den kumulativa fördelningsfunktionen för T givet x.

Funktionen $\lambda(x,t)$ brukar kallas för avvecklingsfunktionen.

4.1 Premieberäkning

Antag att en x-årig individ tecknar en sjukförsäkring som ger 1 krona per år i ersättning om han blir sjuk. Han måste dock ha varit sjuk minst tiden k innan ersättning betalas ut, en så kallad karenstid. Antag även att försäkringen gäller till och med z års ålder[4]. Premien för denna försäkring ges då av

$$E(x, z-x, k) = \int_0^{z-x-k} e^{-\delta \cdot s} \cdot \frac{l_{x+s}}{l_x} \cdot \eta(x+s) \cdot \lambda(x+s,k) \cdot \int_k^{z-x-s} \frac{\lambda(x+s,u)}{\lambda(x+s,k)} \cdot e^{-\delta \cdot u} du ds \tag{1}$$

där

- $e^{-\delta \cdot s}$ är diskonteringsfaktorn med ränteintensitet δ
- l_{x+s}/l_x är sannolikheten att en x-årig person fortfarande är vid liv vid ålder $x+s$.[5]
- $\eta(x+s) \cdot \lambda(x+s,k)$ är sannolikheten att en person blir sjuk vid ålder $x+s$ och fortfarande är sjuk efter tiden k.
- $\lambda(x+s,u)/\lambda(x+s,k)$ är sannolikheten att en person som blivit sjuk vid ålder $x+s$ fortfarande är sjuk efter tiden u givet att han var sjuk vid tiden k, där $k \leq u$.

[4] I praktiken tecknas sjukförsäkring ofta för ett år i sänder. För sådan sjukförsäkring sätts $z = x+1$.

[5] Denna sannolikhet utvecklas inte vidare här. Den intresserade läsaren hänvisas istället till Gunnar Anderssons bok Livförsäkringsmatematik[3]. I de fall då $z = x+1$ ligger denna sannolikhet nära 1.

Vid andra ersättningsbelopp än 1 krona per år multipliceras den så kallade enhetspremien, $E(x, z-x, k)$, med det försäkrade beloppet för att erhålla rätt premie.

För att beräkna enhetspremien behöver vi således ha kännedom om sjuklighetsintensiteten $\eta(x)$, avvecklingsintensiteten $\mu(x,t)$ samt dödlighetsintensiteten $\gamma(x)$. Den sistnämda kommer inte att behandlas i denna uppsats utan vi kommer anta att dödlighetsintensiteten är känd. Därmed begränsas problemet till att skatta $\eta(x)$ och $\mu(x,t)$.

4.2 Insjuknande

En av de huvudsakliga riskerna i en sjukförsäkring är hur stor sannolikheten är att bli sjuk, insjuknandesannolikheten. För att kunna skatta denna sannolikhet behöver vi veta hur många personer som blir sjuka under en viss period samt hur många personer som är under risk att bli sjuka under samma period. Perioden vi använder oss av är kalenderår. Vi har valt att definiera antalet personer under risk under perioden som antalet försäkrade personer vid utgången på året. Mer vanligt är att man istället använder sig av medelantalet av försäkrade personer under perioden. På grund av den stora mängden data som används ger de två definitionerna approximativt samma insjuknandesannolikhet. Vi inför beteckningarna $M(x,I)$ = antalet individer som blivit sjuka vid x års ålder under observationsintervall I, samt $N(x,I)$ = antalet x-åriga individer i det försäkrade beståndet under observationsintervall I. Utifrån dessa beteckningar kan vi skatta intensiteten med vilken en x-årig person blir sjuk som

$$\hat{\eta}(x,I) = \frac{M(x,I)}{N(x,I)}.$$

Eftersom observationsintervallet är satt till ett år blir den skattade intensiteten att insjukna approximativt densamma som sannolikheten att insjukna vid x års ålder.

Vi kommer senare att utgå från denna skattning då vi undersöker hur insjuknandet skiljer sig beroende på kön, ålder, inkomstnivå och vart man är bosatt. Tillvägagångssättet för detta beskrivs närmare i avsnitt 5.

4.3 Avveckling

Den andra stora sannolikheten i en sjukförsäkring är sannolikheten att sjukfallet avvecklas. Vår skattning av avvecklingen kommer att bygga på Kaplan-Meierskattningen. Kaplan-Meier skattar avvecklingsfunktionen $\lambda(x,t)$ för data. Skattningmetoden är mycket vanligt förekommande inom överlevnads-

analys och har fördelen att den tar hänsyn till eventuell censurering som finns i data.

Vi börjar med att definiera avvecklingsintensiteten som

$$\begin{aligned}
\mu(x,t) &= \lim_{h\to 0} \frac{P(t < T \leq t+h | T > t, X = x)}{h} \\
&= \lim_{h\to 0} \frac{1}{h} \frac{P(t < T \leq t+h, T > t | X = x)}{P(T > t | X = x)} \\
&= \frac{1}{P(T > t | X = x)} \lim_{h\to 0} \frac{P(T < t+h | X = x) - P(T < t | X = x)}{h} \\
&= \frac{1}{P(T > t | X = x)} \lim_{h\to 0} \frac{F(t+h|x) - F(t|x)}{h} = \\
&= \frac{1}{P(T > t | X = x)} \frac{d}{dt} F(t|x).
\end{aligned}$$

Tidigare definierade vi $\lambda(x,t) = P\left(T > t | X = x\right) = 1 - F(t|x)$. Deriveras detta uttryck fås

$$\frac{d}{dt}\lambda(x,t) = -\frac{d}{dt}F(t|x)$$

och vidare fås då att

$$\mu(x,t) = -\frac{1}{\lambda(x,t)}\frac{d}{dt}\lambda(x,t) = -\frac{d}{dt}\ln\left(\lambda(x,t)\right)$$

$$\Longleftrightarrow$$

$$\lambda(x,t) = \exp\left(-\int_0^t \mu(x,s)ds\right).$$

I litteraturen brukar man istället för avvecklingsintensitet använda benämningen hazardfunktion. Hazardfunktionen är ett mer generellt begrepp som syftar till intensiteten att förflytta sig mellan olika tillstånd i stokastiska processer. Man brukar även prata om den kumulativa hazardfunktionen vilken definieras som $H(x,t) = \int_0^t \mu(x,s)ds$.

Kaplan-Meierskattningen är den icke-parametriska maximum-likelihood-skattningen[4] av avvecklingsfunktionen $\lambda(x,t)$. Vi låter $t_1, ..., t_N$ vara tidpunkter för vilka vi obeserverat avveckling från sjukfall i ett stickprov av storlek N. Vidare definierar vi d_i som antalet individer som avveckas vid tidpunkten t_i och Y_i som antalet personer som är under risk att avvecklas precis innan t_i. Med detta kan vi definiera Kaplan-Meierskattningen av avvecklingsfunktionen som

$$\hat{\lambda}(t) = \begin{cases} 1 & \text{om } t < t_1 \\ \prod_{t_i \leq t} 1 - \frac{d_i}{Y_i} & \text{om } t \geq t_1. \end{cases} \tag{2}$$

4.4 Avvecklingsmodeller

Ett av huvudsyftena med arbetet är som sagt att undersöka hur avvecklingen påverkas av kön, ålder, duration, inkomst och vart man är bosatt. Vi väljer att presentera de modeller som anpassats till avvecklingen för att undersöka detta redan nu för att på så sätt underlätta för läsaren i ett senare skede. Tre olika modeller har undersökts för att modellera avvecklingsfunktionen och dess parametrar. Vi utgår från Cox Proportionella Hazardmodell som antar proportionalitet mellan hazardfunktionerna. I den andra modellen används samma grundstruktur som i Cox-modellen men vi lättar på proportionalitetsantagandet. Till sist redovisas för en modell, även den utan proportionalitetsantagande, som beskriver kontinuerliga variabler med hjälp av fraktionella polynom, den så kallade MFPT-modellen.

4.4.1 Cox Proportionella Hazardmodell

Cox-modellen är mycket vanligt förekommande inom överlevnadsanalys och väljs ofta på grund av dess enkelhet. Den innehåller ett proportionalitetsantagande som är mycket tilltalande. Man antar att hazardfunktionerna är proportionella mot varandra. I vårt fall innebär detta antagande att avvecklingsintensiteterna antas vara proportionella mot varandra och kvoten mellan två avvecklingsintensiteter antas oberoende av tiden man varit sjuk.

Definition Låt $y_1, y_2, ..., y_p$ vara värden för kovariaterna $Y_1, Y_2, ..., Y_p$. Enligt Cox Proportionella hazardmodell ges då hazardfunktionen av

$$h(t, \mathbf{y}) = h^0(t) \exp\left(\sum_{i=1}^{p} \beta_i y_i\right)$$

där $\beta_1, \beta_2, ..., \beta_p$ är modellens regressionsparametrar och $h^0(t)$ är baselinehazardfunktionen vid tiden t.

Med denna modell kan man beskriva avvecklingsfunktionen som

$$\lambda(t, \mathbf{y}) = \lambda^0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i y_i\right)}$$

eller uttryckt med kumulativa hazardfunktionen

$$H(t, \mathbf{y}) = H^0(t) \exp\left(\sum_{i=1}^{p} \beta_i y_i\right).$$

Med den proportionella Cox-modellen söker vi alltså en funktion $r(\mathbf{y}) = \exp\left(\sum_{i=1}^{p} \beta_i y_i\right)$ som minimerar likelihooden i vår modell

$$h(t) = h^0(t) \exp\left(\sum_{i=1}^{p} \beta_i y_i\right).$$

Proportionalitetsantagandet som görs innebär att $r(\mathbf{y})$ är oberoende av durationen t.

4.4.2 Utvidgad Cox Hazardmodell

I de fall då effekterna av parametrarna beror på durationen kan man inte anta proportionalitet hos avvecklingsintensiteterna. Modellen som beskrivs här har samma struktur som Cox proportionella hazardmodell men vi har lättat på proportionalitetsantagandet genom att tillåta tidsberoende effekter. Dessutom anpassas modellen till den kumulativa hazardfunktionen istället för som i fallet ovan till hazardfunktionen.

Definition Låt $y_1, y_2, ..., y_{p_1}$ vara variabler oberoende av tiden och $y_1(t), y_2(t), ..., y_{p_2}(t)$ vara variabler vars effekter beror på t. Låt vidare $\mathbf{y(t)} = \{y_1, ..., y_{p_1}, y_1(t), ..., y_{p_2}(t)\}$. Den utvidgade Cox-modellen definieras då som

$$H(t, \mathbf{y(t)}) = H^0(t) \exp\left(\sum_{i=1}^{p_1} \beta_i y_i + \sum_{j=1}^{p_2} \delta_j y_j(t)\right).$$

I uppsatsen låter vi de tidsberoende variablerna verka multiplikativt med en funktion $g_j(t)$ vilket gör att vi får en något förenklad modell enligt

$$H(t, \mathbf{y}) = H^0(t) \exp\left(\sum_{i=1}^{p} \beta_i y_i g_i(t)\right)$$

där vi låtit $p_1 = p_2 = p$ och $\mathbf{y} = y_1, ..., y_p$.

Avvecklingsfunktionen fås då som

$$\lambda(t, \mathbf{y}) = \lambda^0(t)^{\exp\left(\sum_{i=1}^{p} \beta_i y_i g_i(t)\right)}.$$

Precis som tidigare söker vi en funktion $r(t, \mathbf{y}) = \sum_{i=1}^{p} \beta_i y_i g_i(t)$ som beskriver avvecklingen. Skillnaden från tidigare är att vid anpassningen av denna modell används minstakvadratmetoden.

4.4.3 MFPT-modellen

MFPT-modellen[6] är egentligen ingen modell utan en algoritm för att finna en modell som är en ytterligare utvidgning av Cox-modellen. Grundmodellen som används är den vanliga proportionella Cox-hazardmodellen som sedan byggs på i tre steg till en MFPT-modell som modellerar de kontinuerliga

[6]MFPT är en förkortning för Multivariable fractional polynomial time

variablerna med hjälp av fraktionella polynom. Slutprodukten blir en modell som baseras på de kumulativa hazardfunktionerna enligt

$$H(t, \mathbf{y}) = H^0(t) \exp\left(\sum_{i=1}^{p} f_i(y_i)\beta_i(t)\right)$$

där $f_i(y_i)$ och $\beta_i(t)$ är första eller andra ordningens fraktionella polynom. I de fall y_i är en kategorisk variabel antas $f_i(y_i) = y_i$.

Fraktionella polynom är mer flexibla än vanliga polynom och första ordningens fraktionella polynom defineras som

$$g_i(z) = \gamma_{i0} + \gamma_{i1} z^p,\ p \in S$$

och andra ordningens som

$$g_i(z) = \begin{cases} \gamma_{i0} + \gamma_{i1} z^{p_1} + \gamma_{i2} z^{p_2} & \text{om } p_1 \neq p_2,\ p_1,\ p_2 \in S \\ \gamma_{i0} + \gamma_{i1} z^{p_1} + \gamma_{i2} z^{p_2} \ln(z) & \text{om } p_1 = p_2 = p \in S \end{cases}$$

där $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Mängden S ger oss $8+36 = 44$ möjliga polynom för att beskriva effekten av kontinuerliga variabler.

I steg ett av MFPT-algoritmen väljer vi ut de kategoriska variabler som skall ingå i modellen. I detta steg bestäms även den funktionella formen för kontinuerliga variabler utifrån de möjliga fraktionella polynomen. För att finna det polynom som bäst beskriver effekten beräknas samtliga 44 möjliga modeller, en för varje fraktionellt polynom. Modellerna jämförs sedan med χ^2-test och den modell som ger bäst anpassning väljs.

I steg två undersöks om det finns variabler med korttidseffekter, det vill säga variabler som enbart påverkar avvecklingen under ett visst durationsintervall. Data delas upp i mindre durationsintervall för att se om effekterna av parametrarna skiljer sig åt under olika intervall och signifikanta korttidseffekter läggs till i modellen.

I steg tre som är det sista steget i algoritmen identifieras eventuella durationseffekter. Durationseffekterna modelleras även de med fraktionella polynom som bestäms på samma sätt som i steg ett genom χ^2-test.

Avvecklingsfunktionen får samma utseende som i den utvidgade Cox-modellen

$$\lambda(t, \mathbf{y}) = \lambda^0(t)^{\exp\left(\sum_{i=1}^{p} f_i(y_i)\beta_i(t)\right)}$$

och den sökta funktionen $r(t, \mathbf{y}) = \exp\left(\sum_{i=1}^{p} f_i(y_i)\beta_i(t)\right)$ bestäms även här med hjälp av minsta kvadratmetoden.

5 Modellanpassning

I detta avsnitt redovisas för de metoder som använts i utvecklandet av de modeller för avveckling och insjuknande som tagits fram. Vi börjar med att beskriva de metoder som använts för att finna en modell som beskriver avvecklingen vilket visat sig vara det mest omfattade arbetet både teoretiskt och praktiskt. Därefter går vi igenom skattningen av insjuknandefrekvens som till stor del bygger på de modeller vi tagit fram för avvecklingen. Slutligen binds avvecklingmodellen och modellen för insjuknande samman och resultatet presenteras i form av den premieberäkningsmodell som är det slutgiltiga målet med uppsatsen.

5.1 Avveckling

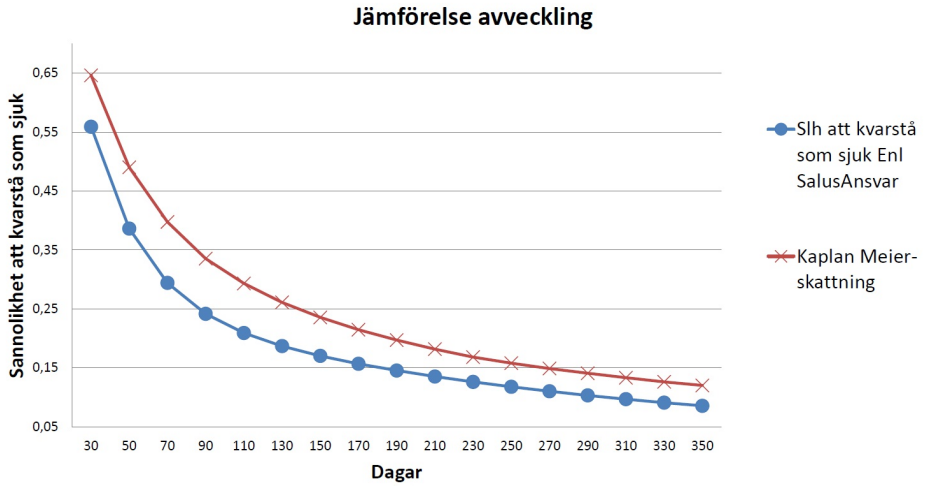
De avvecklingsmodeller som används idag tar ofta bara hänsyn till personens ålder, kön och hur länge personen i fråga varit sjuk. Vi kommer här att ta fram en modell som beskriver avvecklingen med hjälp av just nämnda parametrar samt personens inkomst och var personen är bosatt. Vi kommer till att börja med analysera data från hela sjukbeståndet representerat av tabellen sjukfall 1 som finns beskriven i avsnitt 3. Denna tabell innehåller samtliga sjukfall som inträffat från 2003 fram till och med andra kvartalet 2010 uppdelat på parametrarna åldersgrupp, kön, inkomstnivå, region samt duration som är uppdelad i 20-dagarsintervall. Vi har information om antalet sjukfall för varje uppsättning parametrar och dessutom den sammanlagda durationen av sjukfallen för varje uppsättning. Med detta har vi förutom antalet sjukfall i varje grupp möjlighet att beräkna medeldurationen för sjukfallen inom varje grupp. Vid skattning av avvecklingen har vi utgått ifrån just den beräknade medeldurationen som vi anser ger den mest korrekta skattningen utifrån de förutsättningar vi har. Utöver ovanstående innehåller data även information om vilket kvartal sjukfallen inträffat. Detta ger oss möjlighet att undersöka eventuella trender i avvecklingen över tiden. Vi vet att man i juli 2008 införde nya sjukskrivningsregler och det kan vara intressant att se om/hur detta påverkat avvecklingen.

För att skatta avvecklingsfunktionen använder vi Kaplan-Meierskattningen, en vanligt förekommande skattningsmetod inom överlevnadsanalys. Vi minns definition 2 från teoriavsnittet

$$\hat{\lambda}(t) = \prod_{t_i \leq t} 1 - \frac{d_i}{Y_i}$$

där Y_i= antal personer under risk precis innan tidpunkt t_i och d_i = antalet personer som avvecklas vid tiden t_i. Som duration t_i används medeldurationen inom varje grupp.

En vanlig föreställning är att ett försäkringsbolags sjukbestånd ofta är en betydligt bättre riskgrupp än befolkningen i allmänhet. Med bättre riskgrupp menas att personer som har en privat tecknad sjukförsäkring skulle återhämta sig från sjukdom snabbare än någon som enbart får ersättning från försäkringskassans allmänna sjukförsäkring. I grafen nedan jämförs den skattade avvecklingsfunktionen från försäkringskassans bestånd mot en av de avvecklingskurvor som Salus Ansvar tillämpar.



Figur 2: Jämförelse mellan skattning av avveckling i data och Salus Ansvars avvecklingskurva givet att sjukfallet varat minst 20 dagar, $\lambda(t)/\lambda(20)$, för åldersgruppen 35-39 år.

Vi ser att sannolikheten att kvarstå som sjuk givet att sjukfallet varat minst 20 dagar är högre för försäkringskassans bestånd än för det bestånd Salus Ansvar förväntas ha vilket styrker påståendet om att bestånd från privat tecknad sjukförsäkring skulle vara en bättre riskgrupp.

Det faktum att avvecklingstakten skiljer sig åt mellan privat och allmän sjukförsäkring innebär att vi inte kommer att kunna undersöka de direkta effekterna som parametrarna har på avvecklingen. Vi har alltså ett bestånd vi vill finna en modell för och ett annat bestånd att skatta modellen ifrån. Om vi till exempel kommer fram till att sannolikheten att avvecklas ökar med 10 procentenheter i den allmänna sjukförsäkringen om man är höginkomsttagare istället för låginkomsttagare behöver inte detta innebära en ökning med 10 procentenheter för privat sjukförsäkring. För att få en modell som är applicerbar på avvecklingen i privat sjukförsäkring kommer vi därför undersöka parametrarnas relativa effekter på avvecklingen.

Definition Låt $\lambda_x(t, k, \theta)$ = sannolikheten att en individ av ett visst kön, representerat av k, med parameteruppsättning θ = (Inkomstnivå, Region)

som är x år gammal vid insjuknande fortfarande är sjuk vid tidpunkten t. Låt vidare $\lambda_x^0(t,k)$ = sannolikheten att en individ med kön = k som är x år gammal vid insjuknande fortfarande är sjuk vid tidpunkten t oberoende av θ. Då defineras den relativa överlevnadssannolikheten som

$$R_x(t,k,\theta) = \frac{\lambda_x(t,k,\theta)}{\lambda_x^0(t,k)}.$$

$R_x(t,k,\theta)$ kan betraktas som den faktor vi är tvungna att multiplicera grundavvecklingen $\lambda_x^0(t,k)$ (även kallad baseline) med för att skatta avvecklingsfunktionen givet θ. Om vi nu gör antagandet att de relativa effekterna är desamma för privat sjukförsäkring som för den allmänna försäkringen kan vi använda $R_x(t,k,\theta)$ direkt på vår egen avvecklingsfunktion.

Inom överlevnadsanalys är det vanligare att man anpassar modeller till hazardfunktionen istället för som ovan till avvecklingsfunktionen. Ett tredje alternativ är att anpassa modellen till den kumulativa hazardfunktionen. Hazardfunktionen som bara är en annan benämning på avvecklingsintensiteten är besvärligare att skatta än avvecklingfunktionen och den kumulativa hazardfunktionen varför vi valt att anpassa modellerna i den här uppsatsen till de två senare istället. På samma sätt som ovan för avvecklingsfunktionen definierar vi här den relativa kumulativa hazardfunktionen.

Definition Låt $H_x(t,k,\theta)$ = den kumulativa hazardfunktionen för en individ av ett visst kön, representerat av k, med parameteruppsättning θ som är x år gammal vid insjuknande och har varit sjuk tiden t. Låt vidare $H_x^0(t,k)$ = den kumulativa hazardfunktionen för en individ med kön = k som är x år gammal vid insjuknande och har varit sjuk tiden t oberoende av θ. Då definieras den relativa kumulativa hazardfunktionen som

$$r_x(t,k,\theta) = \frac{H_x(t,k,\theta)}{H_x^0(t,k)}.$$

Skattningen av $r_x(t,k,\theta)$ kan sedan fås från Kaplan-Meierskattningen som

$$\hat{r}_x(t,k,\theta) = \frac{\hat{H}_x(t,k,\theta)}{\hat{H}_x^0(t,k)} = \frac{-\ln\left(\hat{\lambda}_x(t,k,\theta)\right)}{-\ln\left(\hat{\lambda}_x^0(t,k)\right)}.$$

Det har under arbetets gång visat sig att modeller baserade på $r_x(\cdot)$ generellt sätt gett bättre anpassning till data. En bidragande orsak till detta är att de absoluta effekterna sett över tid ökar betydigt mer i $R_x(t,k,\theta)$ än i $r_x(t,k,\theta)$.

Nuvarande avvecklingsmodell tar som vi redan nämnt hänsyn till de tre faktorerna ålder, kön och duration. Det är således enbart faktorerna inkomst

och region som tillkommer i en utvidgning av nuvarande modell. Av den anledningen görs antaganden om att kön, ålder och duration endast påverkar $r_x(t, k, \theta)$ i interaktion med antingen inkomst- eller regionsparametern. Det vill säga vi antar att de tre faktorerna inte påverkar själva avvecklingsfunktionen ytterligare utan snarare påverkar de effekter inkomstnivå och region har på avvecklingsfunktionen.

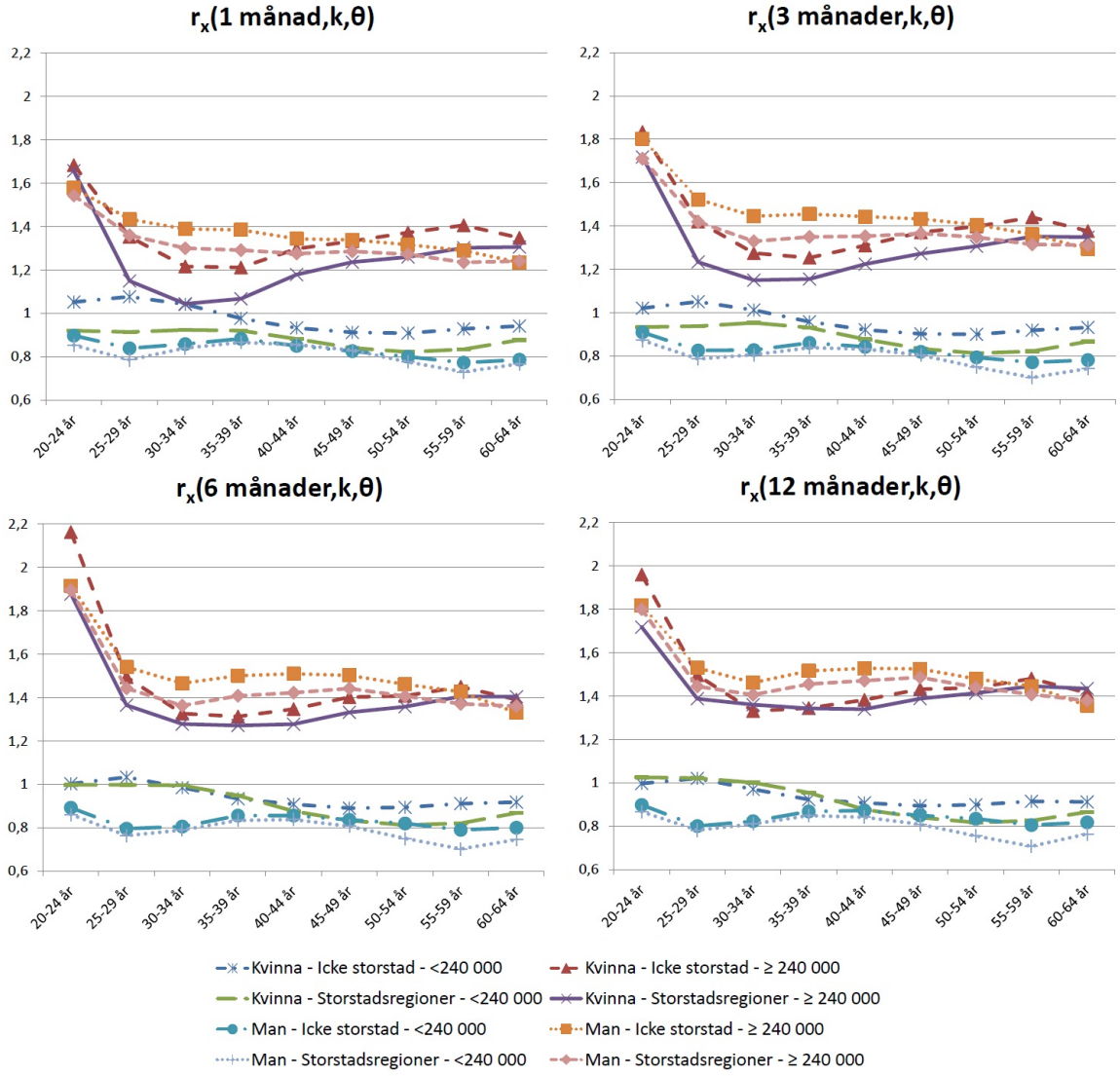
Två stycken kontinuerliga variabler återfinns i datamaterialet, nämligen duration och ålder. Dessa är i och för sig redovisade som diskreta men är i grunden kontinuerliga. Ålder är uppdelad i femårsintervall 20-24 år, ..., 60-64 år och duration är uppdelad i 20-dagarsintervall. I arbetet med modellanpassning har vi modellerat ålder både som en kontinuerlig variabel och som en faktorvariabel. Vid användandet av kontinuerlig ålder har vi antagit en ålder som ligger i mitten av femårsintervallet, det vill säga individer i intervallet 20-24 år antas alla vara 22,5 år gamla och så vidare. Duration har enbart modellerats som kontinuerlig då vi har kännedom om medeldurationen för sjukfall inom samma grupp. Även inkomst är i grunden en kontinuerlig variabel men i vår data är den binär då den endast kan anta värdena $< 240\ 000$ kr eller $\geq 240\ 000$ kr som representerar låg- respektive höginkomsttagare. Att modellera inkomst som kontinuerlig är därför inte aktuellt i vårt fall.

För att undersöka hur avvecklingen påverkas av ålder visas skattningen $\hat{r}_x(t, k, \theta)$ i figur 3 nedan för fyra olika durationer.

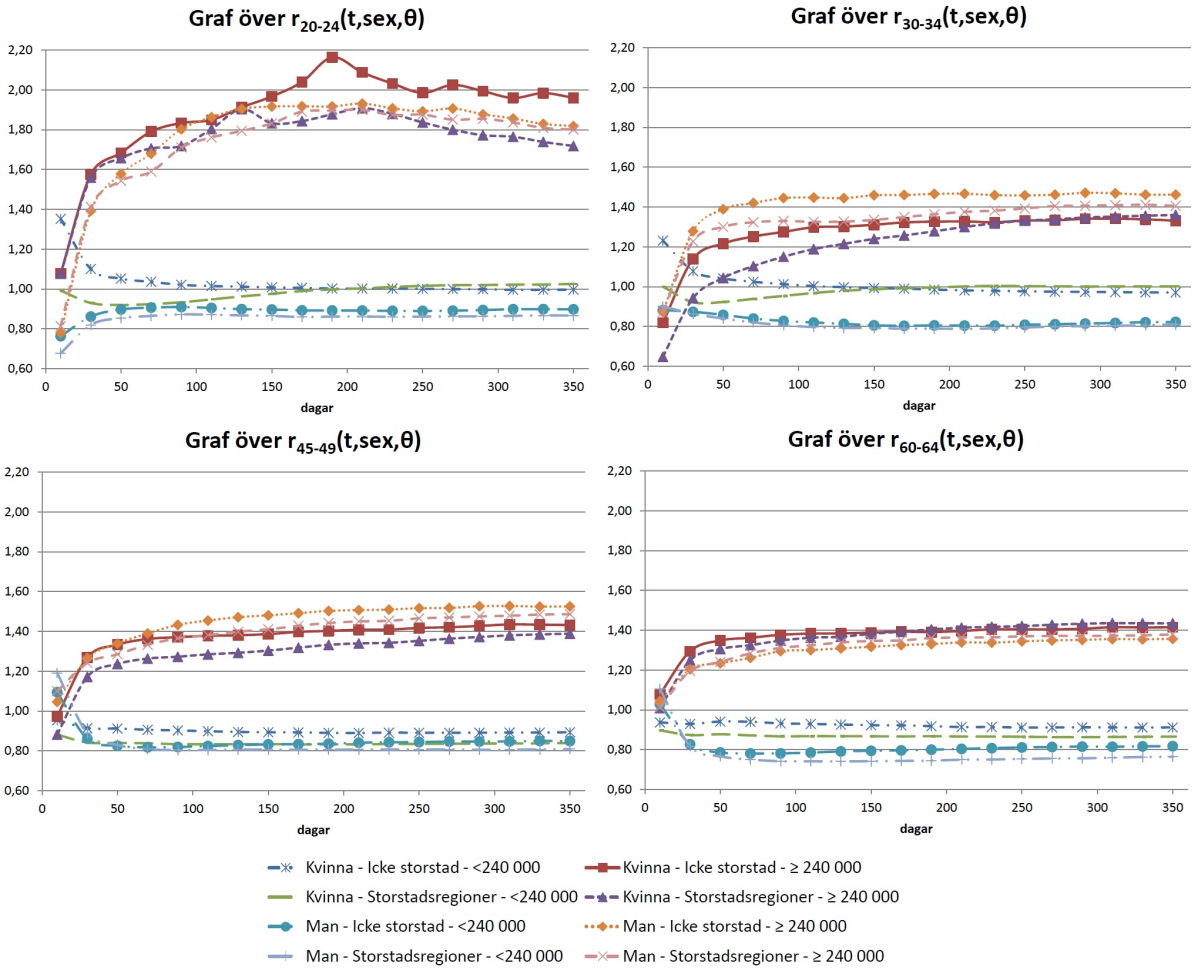
Graferna antyder att det finns ett åldersberoende att ta hänsyn till. Möjligtvis skulle ålderseffekterna kunna modelleras med hjälp av polynom. I graferna syns även tydliga skillnader mellan låg- och höginkomsttagare där höginkomsttagare avvecklas snabbare än låginkomsttagare. Dessutom finns det en antydan till att ålder och inkomst interagerar med varandra eftersom $\hat{r}_x(t, k, \theta)$ påverkas olika av ålder beroende på inkomstnivå. Vi kommer att återkomma till detta senare.

För att undersöka hur vår andra kontinuerliga variabel, duration, påverkar avvecklingen visas i figur 4 nedan en graf över hur $\hat{r}_x(t, k, \theta)$ utvecklas över tiden för fyra olika åldersgrupper.

Vad som framgår tydligt är att $\hat{r}_x(t, k, \theta)$ ser ut att konvergera i samtliga av dessa fyra grafer vilket är en önskvärd egenskap. Huvudsyftet med arbetet är som bekant att ta fram en ny prissättningsmodell som tar hänsyn till inte bara den försäkrades ålder och kön utan även dennes inkomst och var han eller hon är bosatt. Ett delmål i detta är att hitta en modell med vilken vi kan prediktera avvecklingsfunktionen utifrån nämnda parametrar. Eftersom vi endast har data för durationer upp till ett år kommer vår modell att vara anpassad efter den avveckling som sker under första året som sjuk. Om vi då kan anta att eventuella samspelseffekter inkomst och region har med durationen avtar ju högre durationen är kommer vi kunna göra säkrare pre-



Figur 3: Illustration av åldersberoende i $\hat{r}_x(t,k,\theta)$ vid fyra olika durationspunkter.



Figur 4: Illustration av tidsberoende i $\hat{r}_x(t, k, \theta)$ för fyra olika åldersgrupper.

diktioner av avvecklingsfunktionen för $t > 1$ år. Att man inte kan se samma konvergensegenskaper för ålder i figur 3 ovan spelar mindre roll då vi sällan är intresserade av att prediktera avvecklingen för individer utanför aktuellt åldersintervall. En sjukförsäkring är avsedd att ersätta för inkomstförlust då en person blir sjuk och inte har möjlighet att arbeta. De flesta arbetsföra individer är idag mellan 20 och 65 år gamla och det är därför inte så intressant att veta avvecklingen för personer utanför intervallet 20-65 år.

En annan observation vi gör i grafen ovan är att $\hat{r}_x(t, k, \theta)$ varierar mycket för åldersgruppen 20-24 år, speciellt för höginkomsttagare. Detta är något vi sett genomgående i all analys som gjorts under arbetets gång vilket medfört vissa svårigheter då vi försökt anpassa modeller till de lägre åldersgrupperna. En teori om varför det förhåller sig på detta vis skulle kunna vara att det inte finns så många unga sjuka personer som samtidigt är höginkomsttagare och att underlaget för denna grupp således är statistiskt instabilt. Denna teori styrks av tabellen nedan.

	Antal sjukfall			
Ålder	**<240 000**	**≥240 000**	**Totalt**	**Andel ≥240 000**
20-24	200 610	23 882	224 492	11%
25-29	307 473	63 028	370 501	17%
30-34	388 427	106 458	494 885	22%
35-39	380 503	139 592	520 095	27%
40-44	337 693	168 844	506 537	33%
45-49	297 243	175 247	472 490	37%
50-54	288 015	193 238	481 253	40%
55-59	294 026	202 012	496 038	41%
60-64	219 224	139 259	358 483	39%

Ju fler observationer vi har desto säkrare blir våra skattningar. Vi har därför valt att använda oss av modeller som är viktade efter antalet sjukfall inom varje grupp. På så sätt lägger man större vikt på de skattningar av avvecklingen som baseras på ett större antal obeservationer och mindre vikt för de skattningar där vi har färre observationer.

En av de enklare modellerna som beskrivs i litteraturen är Cox Proportionella Hazardmodell[5] som antar att hazardfunktionerna är proportionella mot varandra. Vi tänker oss två individer med parameteruppsättning θ_1 repektive θ_2, proportionalitetsantagandet innebär att kvoten av deras hazardfunktioner, $h_x(t, k, \theta_1)$ och $h_x(t, k, \theta_2)$, är oberoende av durationen t. Ekvivalent innebär detta antagande även att kvoten av de kumulativa hazardfunktionerna är oberoende av durationen och alltså att $r_x(t, k, \theta)$ ovan kan skrivas $r_x(k, \theta)$.

Tre olika modeller har undersökts för att modellera avvecklingsfunktionen. Det är ovanstående Cox proportionella hazardmodell samt två modeller som

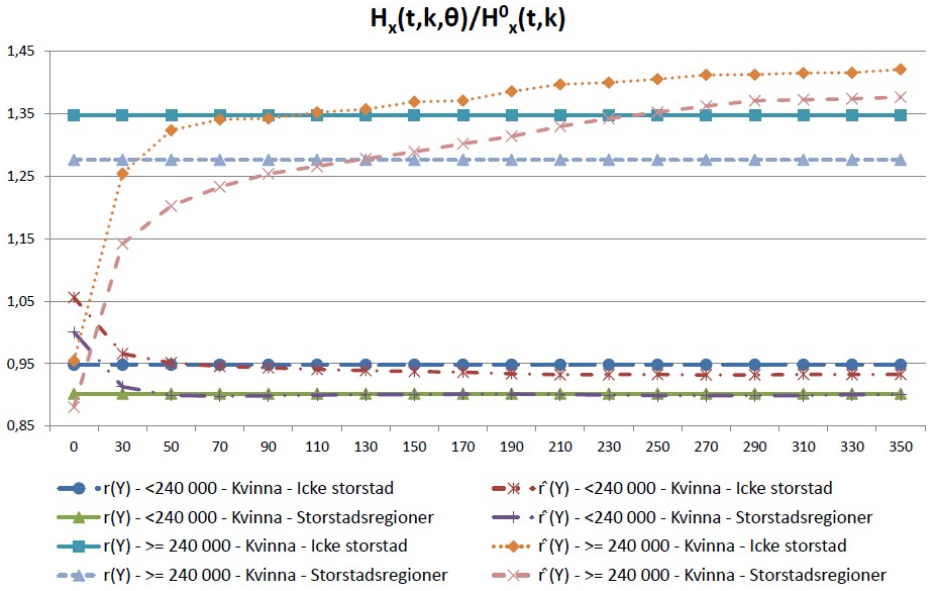
är en utvidgning av Cox-modellen ovan där man frångått proportionalitetsantagandet. Dessa två modeller är den utvidgade Cox-modellen samt MFPT-modellen.

Uttryckt i $r_x(t,k,\theta)$ får Cox proportionella hazardmodell utseendet

$$r(Y) = \exp\left(\sum_{i=1}^{p} \beta_i y_i\right)$$

där $Y = \{y_1, ..., y_p\}$ och y_i = värdet av våra parametrar. Vi har tagit oss friheten att tillåta transformationer av åldersvariabeln x för att bättre passa data, exempelvis $x_2 = x^2, x_3 = x^3, ...$. Även tänkbara dummy-variabler har tillåtits som till exempel uppdelningar på ung och gammal. Som vi såg i figur 3 ovan antar åldersberoendet inte någon enkel funktionell form och det visar sig att transformationerna av x är till stor hjälp.

För att enkelt kunna se det proportionalitetsantagande som Cox-hazard-modellen antar visas i figur 5 nedan den skattade modellen $r(Y)$ och dess responsvariabel $\hat{H}_x(t,k,\theta)/\hat{H}^0_x(t,k)$.



Figur 5: Cox-modell från första sjukdagen och responsvariabeln. $\hat{r}(Y) = \hat{H}_x(t,k,\theta)/\hat{H}^0_x(t,k)$ och $r(Y)$ = skattad Cox-modell. Endast kvinnor redovisas i grafen.

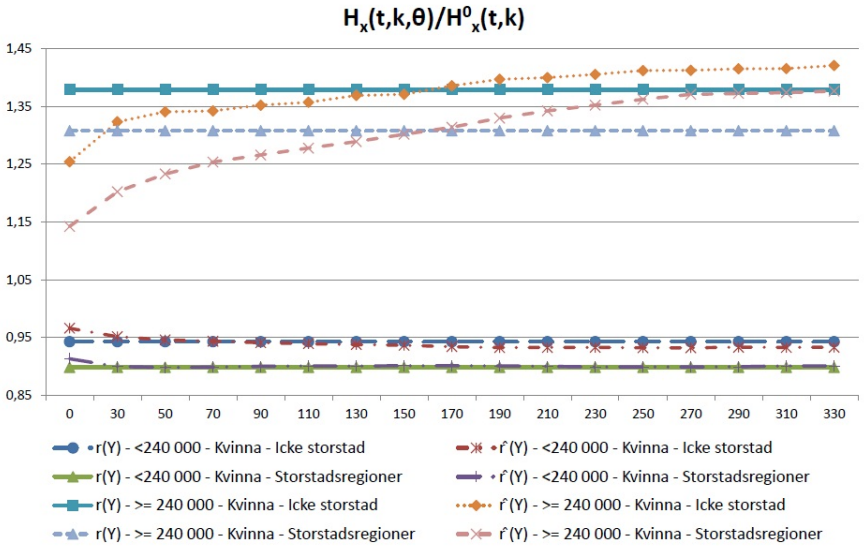
De horisontella linjerna i grafen representerar den skattade Cox-modellen $r(Y)$ som kan sättas i relation till den inte lika horisontella responsvariabeln $\hat{r}_x(t,k,\theta)$. Vi kan se att Cox-modellen ger en ganska bra anpassning för låginkomsttagare. För dem påverkar inte sjukfallets längd avvecklingen

nämnvärt, förutom möjligtvis under de 30 första dagarna. Däremot finner vi ett betydligt större tidsberoende hos höginkomsttagare där den relativa kumulativa hazardfunktionen skiljer sig relativt mycket beroende på sjukfallets längd. Däremot minskar durationseffekten ju längre sjukfallet varat precis som vi kunde se i figur 4 ovan.

I de flesta sjukförsäkringar har man en karenstid som uppgår till minst en månad och ofta ännu längre. Det betyder att en person måste vara sjuk minst en månad innan han eller hon kan få ut någon ersättning från sin försäkring. Man kan därför argumentera för att hur avvecklingen ser ut under den första månaden inte är så intressant utan det enda som är av intresse är om personen fortfarande är sjuk då karenstiden löpt ut. I syfte att minimera beroendet av durationen och få en enklare modell har vi därför valt att anpassa modellen till en delmängd av data där vi bortser från skattningarna för $t < 30$ dagar. Detta betyder inte att vi bortser från de sjukfall vars duration är kortare än 30 dagar. Kaplan-Meierskattningen som används för att skatta avvecklingsfunktionen och således även vår responsvariabel $\hat{r}_x(t, k, \theta)$ är en kumulativ funktion som bygger på tidigare skattningar. Exempelvis gäller att

$$\hat{\lambda}_x(t_2) = \hat{\lambda}_x(t_1) \prod_{t_i \leq t_2} 1 - \frac{d_i}{Y_i}.$$

Vi tar således hänsyn till de sjukfall som avslutats inom 30 dagar men vi bortser från $\hat{\lambda}_x(30 \text{ dagar})$ när vi gör vår modellanpassning.



Figur 6: Cox-modell anpassad från dag 30 och responsvariabeln. $\hat{r}(Y) = \hat{H}_x(t, k, \theta)/\hat{H}^0_x(t, k)$ och $r(Y) =$ skattad Cox-modell. Endast kvinnor redovisas i grafen.

I grafen ovan visas den proportionella Cox-modellen anpassad till den reducerade datamängden. Vi ser att anpassningen är betydligt bättre nu när vi bortser från de 30 första dagarna, vilket även framgår i en jämförelse av modellerna i tabellen nedan.

Modell	**Cox-Prop från $t = 0$**	**Cox-Prop från $t = 30$**
Antal parametrar	11	11
Antal obs.	1 296	1 224
SSE r(Y)	27,70	17,46
MSE r(Y)	0,0214	0,0143
SST r(Y)	122,24	119,43

Tabellen visar att vi har lika många signifikanta parametrar i båda modellerna. Felkvadratsumman (SSE) för den senare modellen är betydligt lägre än för den första modellen anpassad från första sjukdagen. Detta trots att totalkvadratsumman (SST) är likvärdig i båda modellerna. Observera även att vi med antalet observationer här menar antalet unika uppsättningar av parametrarna duration, ålder, kön, inkomstnivå och region.

Tyvärr antyder figur 6 att det fortfarande finns ett visst tidsberoende hos höginkomsttagare som den proportionella Cox-hazardmodellen inte lyckas fånga upp. För att få en bättre anpassad modell har vi således lättat på det proportionalitetsantagande som Cox-modellen gör och infört en durationsvariabel. Men det görs med, som tidigare nämnts, begränsningen att durationen endast tillåts verka i interaktion med variablerna inkomstnivå och region. Modellen som vi kallar den utvidgade Cox-modellen får utseendet

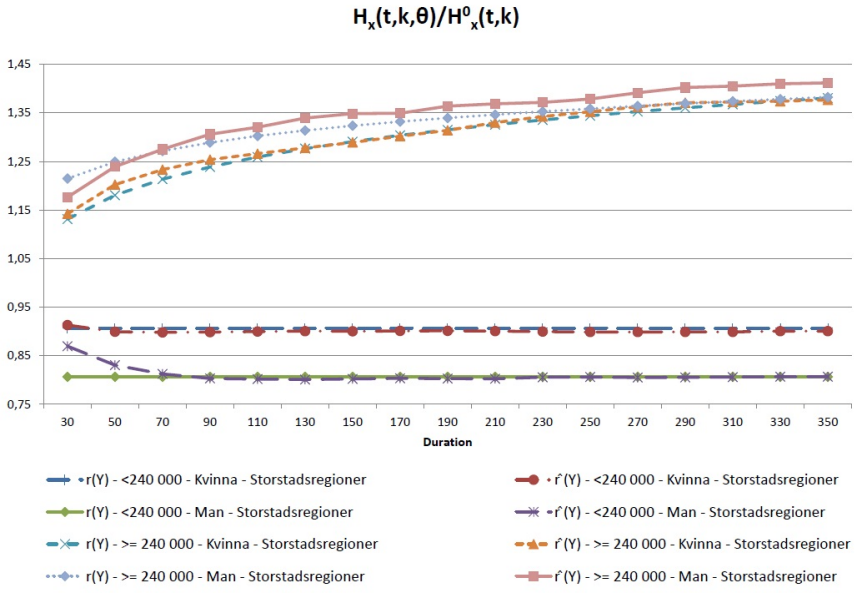
$$r(t, Y) = \exp\left(\sum_{i=1}^{p} \beta_i y_i g_i(t)\right)$$

där $g_i(t)$ är en funktion av tiden. Den slutgiltigt valda avvecklingsmodellen är en modell med detta utseende där vi som diskuterats ovan valt att inte ta med skattningar av avvecklingsfunktionen för den första månaden av sjukperioden i anpassningen. Det visar sig att tiden bäst modelleras med en funktion $g_i(t) = \ln(t)$, där t mäts i år, som interagerar med inkomstnivå och kön där

$$g_i(t) = \begin{cases} \ln(t) & \text{om inkomst } \geq 240\ 000 \\ 0 & \text{om inkomst } < 240\ 000 \end{cases}$$

och kön endast påverkar koefficienten β_i. Att inkomstnivån interagerar med duration på det här sättet stämmer väl överens med vad vi sett i flera av graferna ovan där durationseffekten till synes varit större för höginkomsttagare. Att tidsberoendet kan beskrivas med funktionen $\ln(t)$ är trevligt i flera avseenden, $\ln(t)$ är en konkav funktion vilket även durationseffekten visat sig vara, dessutom blir $\ln(t) = 0$ vid $t = 1$. Om vi sätter villkoret i vår modell

att $g(t) = 0$ då $t \geq 1$ får vi en durationseffekt som enbart påverkar avvecklingen för de durationer vi har data för. Grafen nedan visar den utvidgade Cox-modellens anpassning till data.

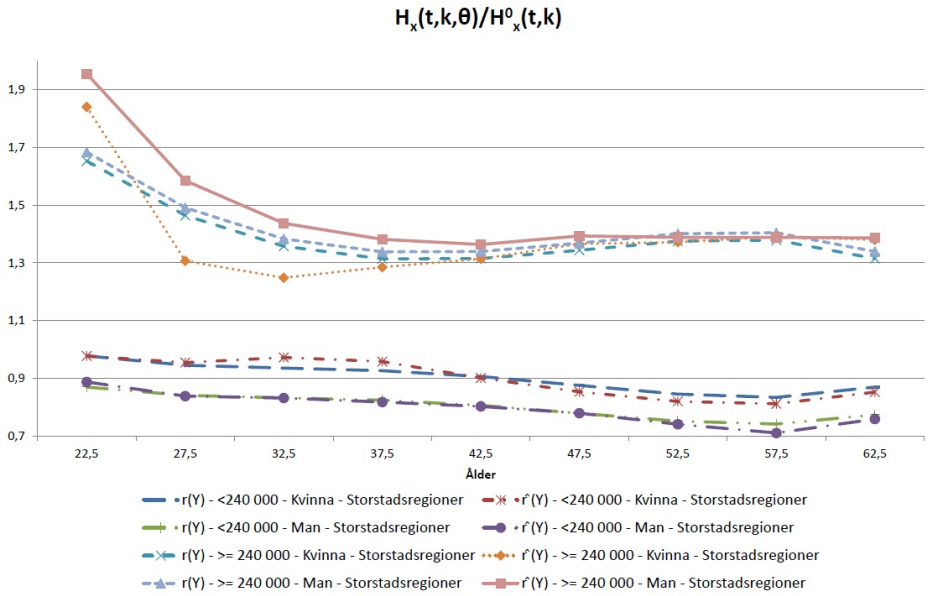


Figur 7: Utvidgad Cox-modell anpassad från dag 30 och responsvariabeln med $x = 40 - 44$ år. $\hat{r}(Y) = \hat{H}_x(t,k,\theta)/\hat{H}^0_x(t,k)$ och $r(Y) =$ skattad modell. Endast storstadsregion redovisas.

Vi ser i den övre halvan av grafen att modellen $r(Y)$ beror på durationen för höginkomsttagare precis som vi nämner ovan.

Ålderseffekten är det vi haft störst problem med att modellera då den till synes inte har någon enkel funktionell form. Vi argumenterade tidigare för att vi inte var intresserade av vad som hände utanför aktuellt åldersintervall 20-64 år. Med detta argument skulle man kunna beskriva åldersberoendet med hjälp av en relativt komplicerad funktion utan att tänka på vad som händer då åldern är lägre än 20 år eller högre än 64 år. Man skulle även kunna tänka sig att modellen inte beror på ålder alls och det vi ser bara är slumpmässiga variationer. Men på grund av den stora datamängd vi har, ca 4 miljoner sjukfall, och det faktum att ålderseffekten till synes inte är helt oregelbunden avslås denna tanke. Efter att ha undersökt flera olika modeller där ålder har beskrivits både med hjälp av dummyvariabler och med funktioner av kontinuerlig ålder har vi till slut kommit fram till en modell där ålder beskrivs av fjärdegradspolynom som verkar i interaktion med inkomstnivå. Modellen blir på så sätt aningen mer komplicerad än vad vi hade önskat men tar ändå hänsyn till ett viktigt beroende som ålder har på avvecklingen. Åldereffekten i form av fjärdegradspolynom illustreras i grafen

över den utvidgade Cox-modellen nedan.



Figur 8: Utvidgad Cox-modell anpassad från dag 30 och responsvariabeln, för $t = 190$ dagar. $\hat{r}(Y) = \hat{H}_x(t,k,\theta)/\hat{H}^0_x(t,k)$ och $r(Y) =$ skattad modell. Endast storstadsregion redovisas.

Vad som inte diskuterats så mycket än är hur avvecklingen påverkas av vart man är bosatt. Parametern som används är region och är precis som inkomstnivå en binär parameter uppdelad på storstadsregion och icke storstadsregion. Effekten av inkomst är som vi sett väldigt tydlig vilket gjort det svårt att urskilja några direkt synbara effekter av regionsparametern. Ett log-rank test[7] där vi jämfört avvecklingen mellan storstadregion och icke storstadsregion visar dock något oväntat att avvecklingen är lägre i storstadsregioner. Log-rank testet ger en normalfördelad teststatistika vars värde blir $Z = 25{,}13$, ett tvåsidigt test ger ett p-värde mindre än 0,0001. Testet säger oss alltså att det finns signifikanta skillnader i avvecklingen mellan personer bosatta i storstadsregioner och icke storstadsregioner. En modell för avvecklingen bör därför ta hänsyn till detta. Införandet av en regionsparameter i vår modell visar sig ge en bättre anpassning. Värt att nämna är också att i ingen av de undersöka avvecklingsmodellerna har regionsparametern visat sig samspela med någon av de övriga effekterna, vilket antyder att den effekt bostadsort har på avvecklingen är helt oberoende av övriga parametrar.

Den utvidgade Cox-modellen är alltså den modell som vi anser bäst beskriver den relativa kumulativa hazardfunktionen. Det är inte den modell som

[7]Log-rank testet är ett icke-parametriskt test som ofta används inom överlevnadsanalys för att jämföra överlevnadssannolikheterna mellan två stickprov[8].

ger bäst anpassning till data men det är den modell som har bäst anpassning i förhållande till antalet förklarande variabler. Modellen innehåller tolv parametrar, ålder modelleras med fjärdegradspolynom som interagerar med inkomstnivå, effekten av inkomstnivå påverkas av ålder, kön och duration medan parametern region är oberoende av övriga parametrar. Modellen får utseendet

$$\begin{aligned} r_x(t,k,\theta) = \exp(&1{,}452 + 0{,}047R_I \\ &+ I_S(-0{,}041x - 0{,}0006x^2 + 0{,}00004x^3 - 0{,}0000003x^4) \\ &+ I_L(-0{,}174x + 0{,}007x^2 - 0{,}0001x^3 + 0{,}0000007x^4) \\ &+ 0{,}116I_LK_K + 0{,}052I_SK_M\ln(t) + 0{,}079I_SK_K\ln(t)) \end{aligned} \tag{3}$$

där

$$R_I = \begin{cases} 1 & \text{om icke storstad} \\ 0 & \text{om storstad} \end{cases}$$

$$I_S = \begin{cases} 1 & \text{om inkomst } \geq 240\ 000 \\ 0 & \text{om inkomst } < 240\ 000 \end{cases}$$

$$I_L = \begin{cases} 1 & \text{om inkomst } < 240\ 000 \\ 0 & \text{om inkomst } \geq 240\ 000 \end{cases}$$

$$K_K = \begin{cases} 1 & \text{om kvinna} \\ 0 & \text{om man} \end{cases}$$

$$K_M = \begin{cases} 1 & \text{om man} \\ 0 & \text{om kvinna} \end{cases}$$

$$t = \text{duration}$$

$$x = \text{ålder}$$

En jämförelse mot de två proportionella Cox-hazardmodellerna visar på en betydlig skillnad i anpassning.

Modellnamn	**C-PH$_0$**	**C-PH$_{30}$**	**Utvidgad Cox**
Antal parametrar	11	11	12
antal observationer	1 296	1 224	1 224
SSE r(Y)	27,70	17,46	4,88
MSE r(Y)	0,0214	0,0143	0,0040
SST r(Y)	122,24	119,43	119,43

Den tredje metoden som använts för att finna lämplig modell är MFPT-algoritmen[8]. Även här utgår vi från den proportionella Cox-hazardmodellen

[8]MFPT-algoritmen finns beskriven i avsnitt 4.

men effekten för de kontinuerliga variablerna ålder och duration modelleras med hjälp av en uppsättning fördefinierade polynom, så kallade fraktionella polynom. Fraktionella polynom som är mer flexibla än vanliga polynom kan användas inom regressionsanalys där vanliga ”traditionella” polynom är otillräckliga för att beskriva data. MFPT-algoritmen som används här begränsar sig till att använda första och andra ordningens fraktionella polynom vilket ger $8+36 = 44$ möjliga polynom för att beskriva ålder och duration. Algoritmen är en relativt enkel metod där man inte på förhand bestämmer den funktionella formen för de effekter man vill modellera. Den durationseffekt som fås i denna modell har samma struktur som i den utvidgade Cox-modellen ovan, det vill säga $g_i(t) = \ln(t)$. Men en brist hos MFPT-algoritmen som gjort att vi valt en annan modell som den slutgiltiga är att andra ordningens fraktionella polynom visat sig vara otillräckliga för att fånga upp de åldersvariationer som finns i data. Som den utvidgade Cox-modellen ovan visat krävs polynom av fjärde ordningen för att fånga upp ålderseffekterna. Man skulle i teorin kunna utveckla MFPT-algoritmen till att omfatta högre ordningens fraktionella polynom och på så sätt få en modell som bättre fångar upp åldereffekterna. En sådan utveckling är dock både omständig och tidskrävande att översätta till praktiken.

MFPT-algoritmen applicerad på data från dag 30 och framåt ger en modell med hela 43 förklarande variabler vilket anses alldeles för många. Anledningen till det stora antalet variabler är att algoritmen jämför de framtagna modellerna med likelihood-test och väljer den mer kompicerade modellen om det finns signfikanta skillnader mellan modellerna. Problemet är att mängden data gör att även små skillnader mellan modellerna blir signifikanta.

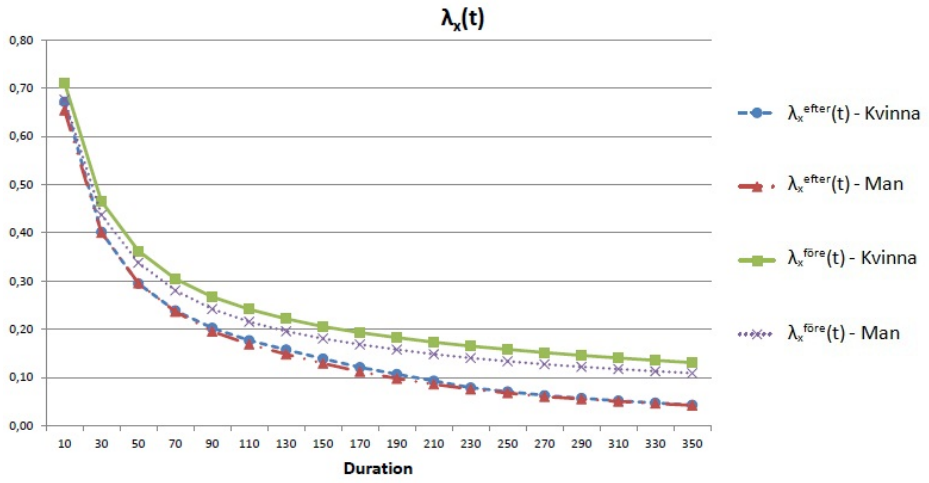
Av de ovan framtagna modellerna är den modell som bäst förklarar variationerna i den relativa kumulativa hazardfunktionen MFPT-modellen men då antalet parametrar i modellen är väsentligt högre än i övriga modeller anses den utvidgade Cox-hazardmodellen vara den bästa modellen. Nedan har vi sammanställt samtliga fyra modeller.

Modellnamn	**C-PH$_0$**	**C-PH$_{30}$**	**Utvidgad Cox**	**MFPT**
Antal parametrar	11	11	12	43
antal observationer	1 296	1 224	1 224	1 224
SSE r(Y)	27,70	17,46	4,88	3,47
MSE r(Y)	0,0214	0,0143	0,0040	0,0028
SST r(Y)	122,24	119,43	119,43	119,43

5.1.1 Skillnader i avveckling mellan gamla och nya sjukskrivningsregler

Hittills när vi tittat på avvecklingen har vi använt oss av hela datasetet med sjukfall från 2003 fram till och med andra kvartalet 2010. I juli 2008 infördes nya sjukskrivningregler och det kan vara intressant att se om avvecklingen har påverkats av detta. I detta avsnitt undersöker vi om avvecklingen för sjukfall inträffade före juli 2008 skiljer sig från avvecklingen för sjukfall som inträffat efter juli 2008.

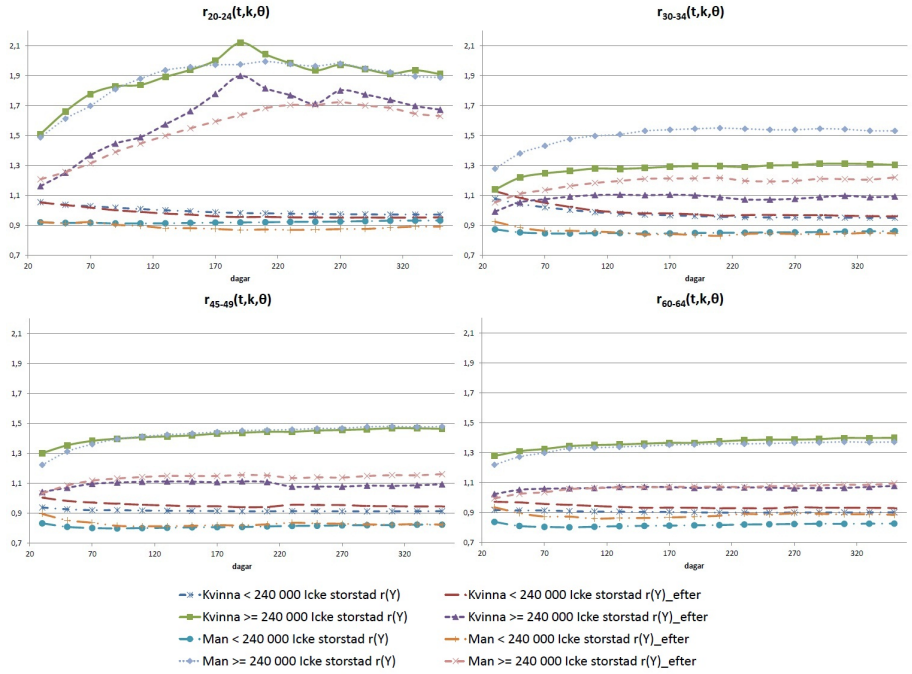
Följande graf visar den skattade avvecklingsfunktionen $\hat{\lambda}_x^0(t,k)$ före respektive efter de nya sjukskrivningsreglerna.



Figur 9: Avvecklingsfunktionen $\lambda_x^0(t,k)$ före och efter införandet av de nya sjukreglerna.

Det framgår tydligt att de nya reglerna har betytt mycket för avvecklingen. Sannolikheten att avvecklas är betydligt högre nu än vad den var tidigare, efter ett års duration är sannolikheten mer än tre gånger så stor efter mot innan. Ett log-rank test som jämför de två periodernas avveckling ger ett p-värde $< 0{,}0001$, alltså högst signifikanta skillnader. Men även om detta resultat är intressant i sig är det inte relevant för vår modell som bygger på relationen mellan avvecklingsfunktionerna $\lambda_x^0(t,k)$ och $\lambda_x(t,k,\theta)$. Frågan som bör ställas är således inte om avvecklingstakten skiljer sig mellan de två perioderna utan snarare om effekten av parametrarna skiljer sig mellan perioderna. Det vill säga får vi samma modell $r_x(t,k,\theta)$ vid anpassning till data före och efter de nya reglerna?

Nedan har vi undersökt detta grafiskt genom att jämföra den relativa kumulativa hazardfunktionen skattad utifrån hela datamängden mot samma skattning för data från perioden efter att de nya reglerna införts.

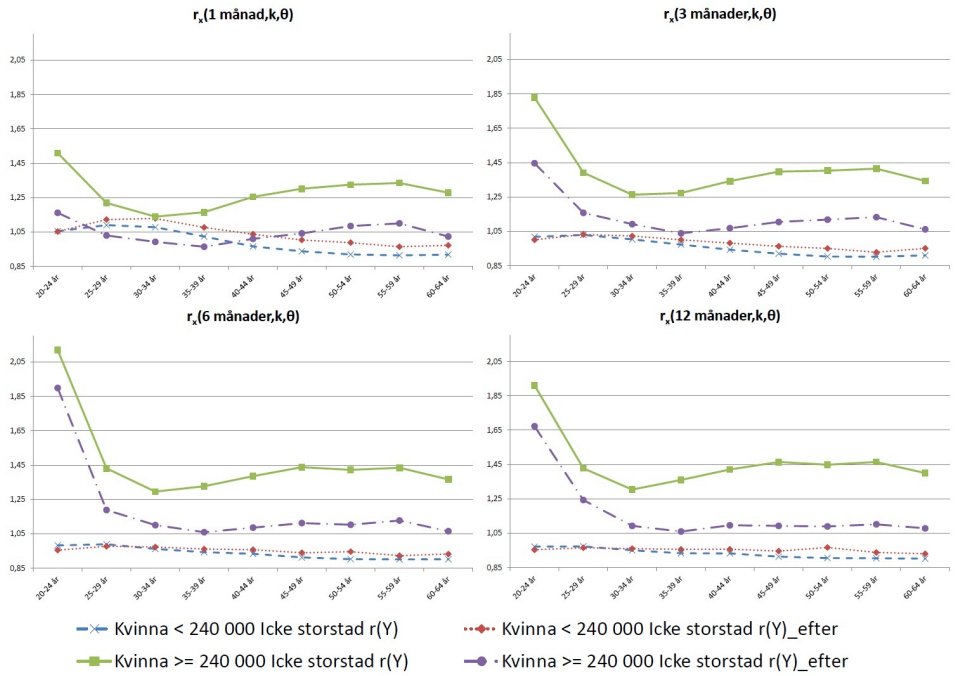


Figur 10: Jämförelse av $\hat{r}_x(t,k,\theta)$ för hela datamängden mot data efter införandet av de nya sjukreglerna.

I grafen ser vi att $\hat{r}_x(t,k,\theta)$ utvecklas på liknande sätt oberoende av vilken period som avses. Skillnaderna mellan de två perioderna ser ut att ligga i storleken av effekternas påverkan på avvecklingen snarare än i deras funktionella form. Detsamma tycks gälla för åldersberoendet som visas i figur 11 nedan.

Modellanpassning till ”efter”-data visar också precis som väntat att de modeller vi kommit fram till ovan även kan användas här. Den utvidgade Cox-modellen anpassad till data från den senare perioden har som enda skillnader en ytterligare konstant effekt för låginkomsttagare samt skillnader i regressionskoefficienterna. En jämförelse mellan denna modell och tidigare modeller baserade på hela datamaterialet visar att denna modell har en lägre felkvadratsumma än den Utvidgade Cox-modellen baserad på hela datasetet. Men då skall man tänka på att SST även är betydligt lägre för denna modell.

Modell	**C-PH$_0$**	**C-PH$_{30}$**	**U-Cox$_{hela}$**	**MFPT**	**U-Cox$_{efter}$**
Antal par.	11	11	12	43	13
antal obs.	1 296	1 224	1 224	1 224	1 224
SSE r(Y)	27,70	17,46	4,88	3,47	4,06
MSE r(Y)	0,0214	0,0143	0,0040	0,0028	0,0033
SST r(Y)	122,24	119,43	119,43	119,43	38,09

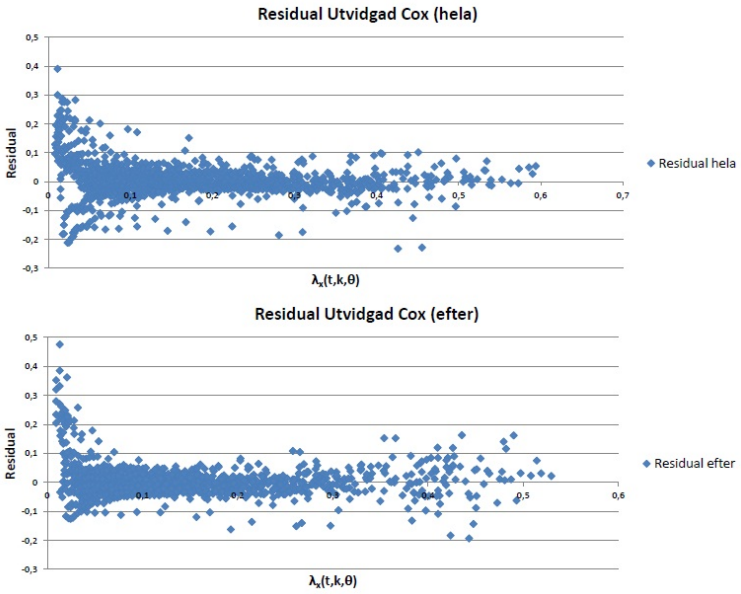


Figur 11: Jämförelse av $r_x(t, k, \theta)$ för hela datamängden mot data efter införandet av de nya sjukreglerna. Enbart kvinnor redovisas.

Förklaringsgraden, $R^2 = 1 - SSE/SST$, för den utvidgade $\text{Cox}_{\text{efter}}$-modellen är 0,89 och kan jämföras med förklaringsgraden för den utvidgade Cox-modellen för hela datamängden $R^2 = 0{,}96$. Modellens anpassning i förhållande till den totala avvikelsen i data är alltså något bättre vid användandet av hela datamängden. Den senare modellen passar ändå relativt bra till data. Detta är även något som framgår i graferna av residualerna för de två modellerna i figur 12 nedan.

I graferna ser vi till och med att anpassningen är bättre i den senare modellen för små värden på avvecklingsfunktionen $\lambda_x(t, k, \theta)$, vilket innebär att avvikelsen i modellen är lägre för höga durationer.

Eftersom effekterna av de undersökta parametrarna skiljer sig mellan de två perioderna innan och efter de nya sjukskrivningsreglerna anses en modell anpassad till hela den observerade perioden vara otillräcklig för vårt ändamål trots att modellen passar data något bättre än modellen anpassad till den senare perioden. Eftersom förutsättningarna ändrats på grund av de nya reglerna behöver vi en modell som också tar hänsyn till detta. Den utvidgade Cox-hazardmodellen anpassad till den senare periden anses vara tillräcklig för att beskriva avvecklingen utifrån de nya förutsättningarna och det är också den modell som väljs som den slutgiltiga modell som kommer ingå



Figur 12: Residualer av $r_x(t,k,\theta)$ mot Kaplan-Meierskattningen $\hat{\lambda}_x(t,k,\theta)$ på x-axeln.

i vår premieberäkningsmodell. Modellen har som tidigare nämnts samma utseende som modellen i ekvation 3 med skillnaden att en extra effekt för låginkomsttagare tillkommer. Modellen får således följande utseende

$$\begin{aligned} r_x(t,k,\theta) = \exp(&4{,}89 + 0{,}02R_I - 4{,}41I_L \\ &+ I_S(-0{,}41x + 0{,}01x^2 - 0{,}0002x^3 + 0{,}000001x^4) \\ &+ I_L(-0{,}08x + 0{,}003x^2 - 0{,}0001x^3 + 0{,}0000004x^4) \\ &+ 0{,}11I_LK_K + 0{,}02I_SK_M\ln(t) + 0{,}05I_SK_K\ln(t)) \end{aligned} \tag{4}$$

med samma beteckningar som ovan.

5.2 Insjuknande

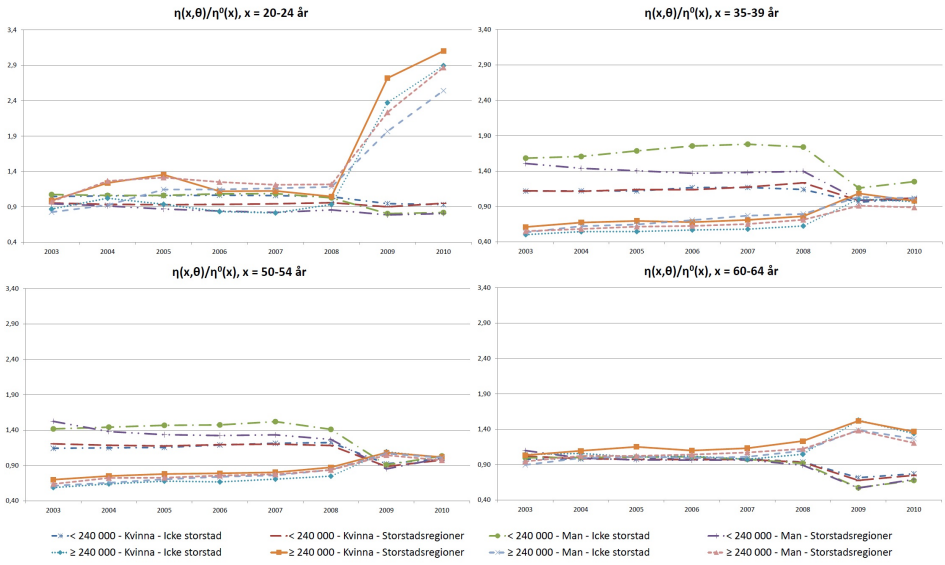
När vi nu gått igenom avvecklingen ovan har vi kommit en lång bit på vägen för att finna en premieberäkningsmodell. Sista steget är nu att finna en modell för att skatta insjuknandeintensiteten, eller sannolikheten att bli sjuk. För att skatta insjuknande tittar vi på hur många personer som blivit sjuka, det vill säga antalet sjukfall, och hur många personer som varit under risk för att bli sjuka, alltså antalet försäkrade personer. Vi använder de två tabellerna sjukfall och försäkrat bestånd till detta. Tabellen sjukfall innehåller samma data som användes för att skatta avvecklingen, skillnaden här är att vi inte tar hänsyn till durationen i sjukfallen då vi endast är intresserade

av om ett sjukfall inträffat inte hur länge det varat. Det försäkrade beståndet representerar alltså antalet personer under risk, tabellen är uppdelad på parametrarna åldersgrupp, kön, inkomstnivå, region och det år beståndet avser. Precis som i fallet med avvecklingen ovan undersöks även här hur de nya sjukskrivningsreglerna påverkat insjuknandet. Vi arbetar på ett liknande sätt med insjuknandet som vi gjort med avvecklingen och mäter de relativa skillnaderna i insjuknande snarare än de faktiska. En tänkbar modell för insjuknande är

$$\eta(x,k,\theta) = \eta^0(x,k)s(x,k,\theta)$$

där $\eta(x,k,\theta)$ är intensiteten med vilken en x-årig person med parameteruppsättning θ och kön k blir sjuk och $\eta^0(x,k)$ är intensiteten med vilken en x-årig person med kön k blir sjuk oberoende av θ, en så kallade baseline. $s(x,k,\theta)$ är en godtycklig funktion som beskriver intensitetens beroende av parametrarna inkomstnivå, region och de eventuellt samspelande köns- och ålderseffekterna.

Första steget blir att undersöka om det finns skillnader mellan skadeåren, det vill säga om effekterna av parametrarna skiljer sig åt mellan olika år. Detta undersöks i graferna över $\hat{s}(x,k,\theta) = \hat{\eta}(x,k,\theta)/\hat{\eta}^0(x,k)$ nedan.

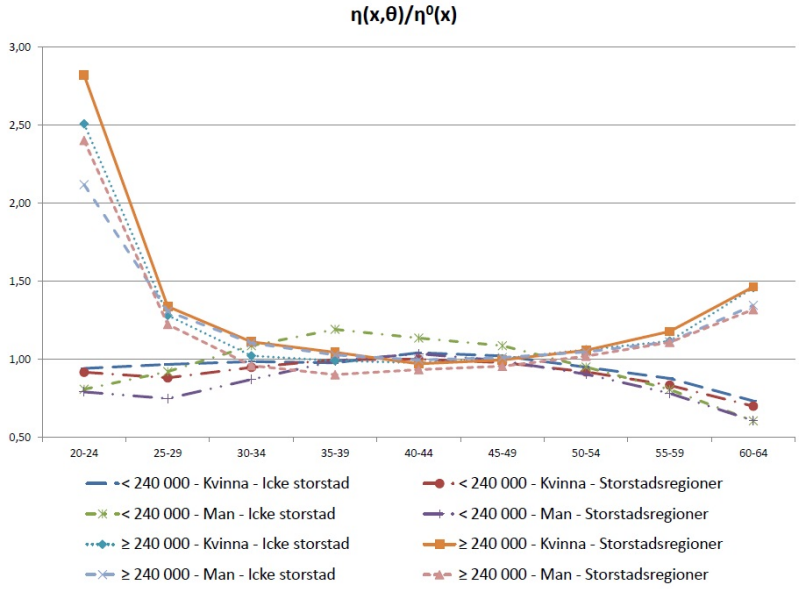


Figur 13: Kvot mellan insjuknandefrekvens och baseline uppdelat per år.

Vad som framgår tydligt av graferna är att någonting verkar hända 2009, eventuellt redan under 2008 men det är ingenting som syns säkert. Sjukskrivningsreglerna ändrades inför tredje kvartalet 2008 och vi kan tänka oss att det vi ser är en effekt av detta. Effekten av hög- respektive låginkomsttagare har ändrats dramatiskt 2009 och 2010. Det faktum att effekterna av

parametrarna ändrats så dramatiskt efter 2008 gör att vi enbart använder data från åren 2009 och 2010.

Möjliga variabler i vår modell är kön, ålder, inkomstnivå och region där ålder är vår enda kontinuerliga variabel. Precis som i fallet med avvecklingen måste vi finna åldersberoendets funktionella form, i grafen nedan undersöker vi ålderberoendet för responsvariabeln $\hat{\eta}(x,k,\theta)/\hat{\eta}^0(x,k)$.



Figur 14: Kvot mellan insjuknandefrekvens och baseline per åldersgrupp.

Vi ser direkt att ålderseffekten samspelar med inkomstnivå där höginkomsttagare i de yngsta och äldsta åldersgrupperna löper en större risk att insjukna än låginkomsttagare i samma åldersgrupper. För personer mellan 30-54 år ser det ut som att insjuknandefrekvensen inte påverkas av inkomstnivå.

Precis som grafen antyder visar det sig vid modellerande av insjuknande att ålderseffekten bäst beskrivs av två polynom, ett för låginkomsttagare och ett för höginkomsttagare. Det är två fjärdegradpolynom som bäst beskriver ålderseffekten i interaktion med inkomstnivå. Med samma resonemang kring ålderseffekterna som vi hade för avvecklingen ovan spelar det mindre roll vad som händer utanför observerat ålderintervall. Det vill säga hur fjärdegradspolynomen beter sig för åldrar lägre än 20 år och högre än 64 år är av mindre vikt för vårt ändamål. De övriga två parametrarna kön och region visar sig inte signifikanta i modelleringen. Vår slutgiltigt valda modell här består således av två stycken fjärdegradspolynom, ett för höginkomsttagare

och ett för låginkomsttagare. Modellen får följande utseende

$$\begin{aligned} s(x,\theta) = 3{,}41 &+ I_H(27{,}17 - 2{,}70x + 0{,}09x^2 - 0{,}001x^3 + 0{,}000007x^4) \\ &+ I_L(-0{,}31x + 0{,}01x^2 - 0{,}0002x^3 + 0{,}000001x^4) \end{aligned} \tag{5}$$

Modell	**Antal par.**	**antal obs.**	**SSE**	**MSE**	**SST**	$\mathbf{R^2}$
$\mathbf{s}(\mathbf{x},\theta)$	9	72	0,53	0,0085	10,27	94,88%

5.3 Premieberäkning

De två modellerna för avveckling och insjuknande som precis är framtagna kan nu användas för att utveckla en premieberäkningsmodell som förutom kön och ålder även tar hänsyn till parametrarna inkomstnivå och region. Premien i en sjukförsäkring beror på sannolikheten att bli sjuk, sannolikheten att avvecklas samt sannolikheten att avlida. Sannolikheten att avlida behandlas inte i denna uppsats utan antas vara känd sedan tidigare. Vi minns definition 1 av enhetspremien ovan

$$E(x, z-x, k) = \int_0^{z-x-k} e^{-\delta\cdot s}\cdot\frac{l_{x+s}}{l_x}\cdot\eta(x+s)\cdot\lambda(x+s,k)\cdot\int_k^{z-x-s}\frac{\lambda(x+s,u)}{\lambda(x+s,k)}\cdot e^{-\delta\cdot u}duds.$$

Vi tänker oss att λ och η i uttrycket ovan representerar avvecklingsfunktion och insjuknandeintensitet oberoende av θ = (inkomstnivå, region). Det vill säga $\lambda(x+s,k) = \lambda^0_{x+s}(t,k)$ och $\eta(x+s) = \eta^0(x+s,k)$ där vi låtit karenstiden k vara lika med t och istället låtit k representera kön för att använda samma beteckningar som tidigare. När vi har vetskap om inkomstnivå och region för en person vi vill beräkna premie för kan vi nu uttrycka avvecklingsfunktionen med hjälp av vår framtagna modell $r_x(t,k,\theta)$ för den relativa kumulativa hazardfunktionen enligt

$$\lambda_{x+s}(t,k,\theta) = e^{-H^0_{x+s}(t,k)r_{x+s}(t,k,\theta)} = \lambda^0_{x+s}(t,k)^{r_{x+s}(t,k,\theta)}$$

där $r_x(t,k,\theta)$ är den Utvidgade Cox-hazardmodellen 4 som är framtagen ovan.

$$\begin{aligned} r_x(t,k,\theta) = \exp(&4{,}89 + 0{,}02R_I - 4{,}41I_L \\ &+ I_S(-0{,}41x + 0{,}01x^2 - 0{,}0002x^3 + 0{,}000001x^4) \\ &+ I_L(-0{,}08x + 0{,}003x^2 - 0{,}0001x^3 + 0{,}0000004x^4) \\ &+ 0{,}11I_LK_K + 0{,}02I_SK_M\ln(t) + 0{,}05I_SK_K\ln(t)). \end{aligned}$$

På samma sätt kan vi uttrycka insjuknandeintensiteten som

$$\eta(x+s,k,\theta) = \eta^0(x+s,k)s(x+s,\theta).$$

där den relativa insjuknandeintensiteten

$$\begin{aligned} s(x,\theta) = 3,41 &+ I_H(27{,}17 - 2{,}70x + 0{,}09x^2 - 0{,}001x^3 + 0{,}000007x^4) \\ &+ I_L(-0{,}31x + 0{,}01x^2 - 0{,}0002x^3 + 0{,}000001x^4). \end{aligned}$$

Enhetspremien för en person med parameteruppsättning (x,k,θ) blir således

$$\begin{aligned} E(x,z-x,t,k,\theta) = &\int_0^{z-x-t} e^{-\delta\cdot s}\cdot\frac{l_{x+s}}{l_x}\cdot\eta(x+s,k,\theta)\cdot\lambda_{x+s}(t,k,\theta) \\ &\cdot\int_t^{z-x-s}\frac{\lambda_{x+s}(u,k,\theta)}{\lambda_{x+s}(t,k,\theta)}\cdot e^{-\delta\cdot u}duds. \end{aligned} \tag{6}$$

där t är den karenstid man måste vara sjuk innan ersättning ges.

6 Slutsatser och Diskussion

Syftet med arbetet var att använda statistik från den allmänna sjukförsäkringen för att utveckla en premiemodell för privat sjukförsäkring som förutom kön och ålder även tar hänsyn till den försäkrades inkomstnivå och bostadsort. Denna modell skulle sedan kunna användas till att beräkna premier för enskilda försäkringstagare eller grupper av försäkrade, så kallade gruppförsäkringar, där vi har kännedom om inkomstnivå och bostadsregion för att få mer individuellt anpassade premier. Statistiken från den allmänna sjukförsäkringen är inte representativ för beståndet i privat sjukförsäkring. Insjuknanderisk och avvecklingsrisk är inte desamma i de två bestånden utan skiljer sig ganska markant från varandra. För att ändå kunna använda statistik från Försäkringskassan för modeller som sedan kan appliceras på privat sjukförsäkring har vi undersökt de relativa riskerna för insjuknande och avveckling och hur parametrarna för kön, ålder, duration, inkomstnivå och region påverkar dessa relativa risker. Men med detta måste vi även göra antagandet att de undersökta parametrarna påverkar de relativa riskerna på samma sätt i båda bestånden. Exempelvis innebär detta att vi gör antagandet att om insjuknandesannolikheten ökar med 5% om en person är höginkomsttagare istället för låginkomsttagare i den allmänna sjukförsäkringen så ökar sannolikheten med 5% även för privat sjukförsäkring. Detta är ett antagande som bör undersökas innan implementation av den framtagna premieberäkningsmodellen genomförs, en undersökning som ligger utanför ramarna för

detta arbete. Anledningen till att vi kan använda de relativa riskerna överhuvudtaget är att vi inte är intresserade av storleken på avvecklingrisk och insjuknanderisk utan enbart hur de undersökta parametrarna påverkar dessa risker.

Vi kommer inte spekulera kring orsakerna till varför modellerna ser ut som de gör, varför höginkomsttagare avvecklas snabbare än låginkomsttagare eller varför höginkomsttagare i de äldre åldersgrupperna insjuknar i större utstäckning än låginkomsttagare i samma åldrar. Detta skulle som sagt enbart vara spekulationer och för den intresserade finns åtskilliga undersökningar där dessa frågor behandlas, bland annat diskuterar Peter Lundborg m fl [10] i en mycket intressant rapport från IFAU[9] sambanden mellan socioekonomiska skillnader och hälsa.

För avvecklingen har vi använt oss av den relativa kumulativa hazardfunktionen som vår responsvariabel i modellerna. En annan möjlighet är att använda den relativa avvecklingsfunktionen som reponsvariabel. Det har dock visat sig att modeller baserade på den relativa kumulativa hazardfunktionen generellt sätt gett bättre anpassning än modeller baserade på den relativa avvecklingsfunktionen.

Av de tre undersökta modellerna för avveckling är det den utvidgade Cox-hazardmodellen som vi anser beskriver avvecklingen bäst. Detta är inte helt oväntat då det är den av våra modeller som är mest flexibel. Den proportionella Cox-hazardmodellen antar proportionalitet, det vill säga att effekterna av undersökta parametrar inte beror på durationen, ett antagande som inte visat sig tillämpbart på vårt data där effekten av inkomst i allra högsta grad beror på sjukfallets längd. I vilket fall för lägre durationer, vid ett års duration antas i vår modell för avvecklingen ovan att inkomsteffekten inte längre påverkas av durationen. Detta är ett antagande som görs då den relativa kumulativa hazardfunktionen som är vår responsvariabel ser ut att konvergera med tiden i kombination med att vi inte har data för högre durationer än ett år.

Den andra modellen som valts bort är MFPT-modellen som bygger på MFPT-algoritmen där antaganden görs att duration och ålder påverkar avvecklingen genom första eller andra ordningens fraktionella polynom. Fördelen med algoritmen är att användandet av fraktionella polynom ger en relativt flexibel modell där vi ändå har begränsat antalet möjliga modeller till en hanterbar mängd. Algoritmen ger även en enkel process för att välja lämplig modell då alla möjliga modeller jämförs och bästa modell väljs med hänsyn tagen till både modellens enkelhet och anpassning till data. MFPT-modellen visar sig dock inte vara tillräckligt flexibel då vi kommit fram till att ålder bäst modelleras med fjärde ordningens polynom. Som nämnts tidigare är det möjligt

[9]Institutet för arbetsmarknads- och utbildningspolitisk utvärdering

att utvidga MFPT-algoritmen till att omfatta fjärde ordningens fraktionella polynom. Att utöka antalet möjliga polynom på detta sätt ger dessvärre snabbt ett oöverskådligt antal möjliga modeller. Med fraktionella polynom upp till andra ordningen hade vi 8+36=44 möjliga modeller för ålder och samma antal för duration. Med tredje ordningens polynom skulle vi få 120 möjliga modeller för respektive parameter och med fjärde ordningens skulle vi få 330 stycken.

Vi har sett tydliga skillnader i både avveckling och insjuknande innan och efter införande av de nya sjukskrivningsreglerna. Avvecklingssannolikheten är högre efter att de nya reglerna införts vilket i sig inte betyder att personer tillfrisknar snabbare nu än tidigare. Det enda vi med säkerhet kan säga är att personer snabbare än tidigare blir utförsäkrade och inte längre får ersättning från Försäkringskassan. Ersättningsbedömningen i privat sjukförsäkring baseras oftast på Försäkringskassans bedömning varför det är mindre relevant från försäkringsbolagens sida huruvida personerna faktiskt blir friska eller bara utförsäkrade. För insjuknande finns inte lika tydliga mönster att sannolikheten skulle vara lägre eller högre innan mot efter de nya reglerna. Som vi såg i figur 13 ser förändringen ut att bero på de undersöka parametrarna i större utstäckning än för avvecklingen, exempelvis fanns tidigare stora skillnader i insjuknande mellan låg- och höginkomsttagare i åldersgruppen 35-39 år medan det nu inte föreligger några större skillnader i insjuknande mellan inkomstnivåerna.

En intressant fråga är naturligtvis hur premien faktiskt påverkas av den nya premieberäkningsmodellen. För att undersöka detta har vi beräknat premier utifrån den avveckling och den insjuknandeintensitet som vi skattat från Försäkringskassans bestånd. Vi använder den skattade avvecklingen och den skattade insjuknandesannolikheten, $\hat{\lambda}_x^0(t,k)$ och $\hat{\eta}^0(x,k)$, utan hänsyn tagen till inkomstnivå och region för att beräkna premien för en sjukförsäkring med en månads karenstid som ger 1000 kr per månad i ersättning under ett år. Därefter har vi jämfört dessa ”grundpremier” mot de premier vi får då vi använder den framtagna premieberäkningsmodellen som tar hänsyn till inkomst och region. Tabell nedan visar hur de nya premierna skiljer sig från grundpremien i relativa tal.

I tabellen kan vi se hur insjuknade och avveckling samverkar i den nya modellen. Sannolikheten att bli sjuk för höginkomsttagare i åldersgruppen 20-24 år är nästan 2,5 gånger så hög som för höginkomsttagare i åldersgruppen 40-44 år. Men avvecklingssannolikheten för samma inkomstgrupp är betydligt lägre för 20-24 åringar än för 40-44 åringar varför kvoten mellan grundpremie och modellpremie för de två grupperna inte skiljer sig nämnvärt från varandra. Dyrast premie i förhållande till grundpremien får låginkomsttagare i åldersgruppen 40-44 år, även låginkomsttagare i åldersgrupperna 30-34 år och 50-54 år som ej redovisas i tabellen ovan får en högre premie med

den nya premiemodellen. Eftersom sannolikheten för insjuknande inte skiljer sig nämnvärt mellan de två inkomstnivåerna för dessa åldersgrupper är den förhöjda premien främst en konsekvens av avvecklingsmodellen.

Region/Inkomst	**Kön**	**20-24 år**	**40-44 år**	**60-64 år**
Icke storstad - $\geq$ 240 000	Kvinna	46%	45%	73%
	Man	62%	53%	89%
Storstad - $\geq$ 240 000	Kvinna	49%	47%	76%
	Man	66%	56%	93%
Icke Storstad - $<$ 240 000	Kvinna	89%	119%	76%
	Man	99%	131%	82%
Storstad - $<$ 240 000	Kvinna	92%	123%	78%
	Man	103%	135%	84%

Tabell 3: Kvot mellan premie enligt framtagen premieberäkningsmodell mot premie utan hänsyn tagen till region/inkomst.

Som avslutning vill vi återigen betona vikten av att undersöka sanningshalten i antagandet att effekterna av undersökta parametrar är de samma i det bestånd man vill använda premieberäkningmodellen för som i Försäkringskassans bestånd. Används premiemodellen felaktigt kan konsekvensen bli en premie som inte täcker kostnaderna för försäkringen vilket i värsta fall kan leda till att försäkringsbolaget går med förlust.

Referenser

[1] Jeffery C. Wayman, Ph.D. **Multiple Imputation For Missing Data: What Is It And How Can I Use It?** Center of Social Organization of Schools John Hopkins University

[2] http://www.tillvaxtverket.se

[3] Gunnar Andersson **Livförsäkringsmatematik** 2005

[4] Germán Rodríguez **Non-Parametric Estimation in Survival Models** Spring, 2001 (revised 2005)

[5] John P.Klein, Melvin L Moeschberger **Survival Analysis - Techniques for Censored and Truncated Data** Second edition, Springer

[6] Nihal Ata, M. Tekin Sözer **Cox regression models with nonproportional hazards applied to lung cancer survival data** Hacettepe Journal of Mathematics and Statistics Volume 36(2) (2007), 157-167.

[7] Anika Buchholz, Willi Sauerbrei **Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data** Biometrical Journal Volume 53 (2011) 2, 308-331.

[8] **Wikipedia - Logrank test** http://en.wikipedia.org/wiki/Logrank_test

[9] Pranjneshu **Non-linear Regression Models and Their Applications** Indian Agricultural Statistics Research Institute

[10] Peter Lundborg m fl **Hur påverkar socioekonomisk status och ålder arbetsmarknadseffekterna av olika hälsoproblem?** Institutet för arbetsmarknads- och utbildningspolitisk utvärdering 2011

[11] Johanna Eriksson **Avvecklingsfunktionen i sjukförsäkring** Matematiska Institutionen, Stockholms Universitet 2009

[12] Marija Miličević **Long-term Health Insurance** Matematiska Institutionen, Stockholms Universitet 2003