Stockholms universitet

Analys av sjuktider med en parametrisk överlevnadsmodell

Lily Melmoth

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2014:4
Matematisk statistik
Maj 2014

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2014:4**
http://www.math.su.se/matstat

Analys av sjuktider med en parametrisk överlevnadsmodell

Lily Melmoth*

Maj 2014

Sammanfattning

I denna uppsats modelleras sjuktider med statistiska metoder inom överlevnadsanalys. Vi undersöker sjukförsäkringar från fyra försäkringsbolag inom Folksamkoncernen för att hitta en modell som beskriver individers förväntade sjuklängder och sannolikheten att fortsätta vara sjuk vid olika tidpunkter. Resultatet blir en parametrisk modell som ger en bra anpassning till data, men datamaterialet innehåller sjukfall som ännu inte har avslutats, vilket gör det svårt att bedöma modellens prediktionsförmåga.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: lily.melmoth@gmail.com. Handledare: Jan-Olov Persson.

Abstract

This thesis models sickness length using survival analysis. We use data from the disability insurance of four insurance companies to find a model for expected sickness length and the probability of remaining sick beyond a specified time. The resulting parametric model fits data well, but the presence of incomplete observations makes it difficult to evaluate the predictive ability of the model.

Contents

1 Inledning

En privat sjukförsäkring som tecknas hos ett försäkringsbolag ger försäkringstagaren utbetalning under tiden denne är sjuk. Personer med privat sjukförsäkring får alltså pengar utöver den allmänna sjukförsäkringen från Försäkringskassan så länge de är sjukanmälda hos bolaget. Försäkringsbolaget som betalar ut ersättningen behöver veta hur länge kunden förväntas vara sjuk för att kunna beräkna premier och hur mycket pengar som behöver sättas av för att kunna betala ut framtida ersättningar. Syftet med denna uppsats är att undersöka hur individers sjuktider påverkas av olika förklarande variabler. Vi modellerar förväntad sjuktid och den så kallade avvecklingsfunktionen, som beskriver sannolikheten att sjuktiden T för en individ är längre än t.

Vi har data från fyra försäkringsbolag och undersöker tillgängliga variabler som kan tänkas påverka längden av ett sjukfall. En svårighet med att i förväg bedöma förväntade sjuktider är att de påverkas av externa faktorer som är svåra att förutsäga. Försäkringskassan bedömer vem som kan bli sjukskriven och när de genomför ändringar som gör det enklare eller svårare att bli sjukskriven så kommer sjukfallslängden påverkas och det är av speciellt intresse att försöka fånga in sådana externa effekter. En variabel som kan tänkas göra detta är andel långtidssjukskrivna i landet och undersöks i uppsatsen.

Eftersom tiden som sjuk är en form av livstid så görs en överlevnadsanalys - ett ämnesområde som modellerar tid till en viss händelse, i vårt fall tid till avveckling. Om vi låter T beteckna antal sjukdagar, med tillhörande fördelningsfunktion F, så ges den så kallade överlevnadsfunktionen $S(t)$ av:

$$S(t) = P(T > t) = 1 - F(t)$$

där tiden t anges i dagar. Överlevnadsfunktionen anger alltså sannolikheten för att sjuktiden är längre än tiden t och är vår sökta avvecklingsfunktion.

1.1 Allmänt

Arbetet är avgränsat till att undersöka hur sjuklängden påverkas av olika variabler, för personer som redan är sjuka. Vi undersöker alltså inte insjuknandet.

Alla förklarande variabler behandlas som konstanta över tiden, det vill säga man ändrar inte kön, bostadsort med mera. En person kan självklart flytta, men vi har endast tillgång till den senaste adressen.

All beräkning och modellering har gjorts i programmeringsspråket SAS genom inbyggda funktioner för överlevnadsanalys som beskrivs i boken [2].

2 Data

Data kommer från bolagen Folksam Liv, Förenade Liv, Salus Ansvar samt KP Pension & Försäkring och innehåller 58 453 observationer från bolagens

sjukförsäkringar. Observationsperioden för dessa är mellan 1 januari 1978 till 18 november 2013 och eftersom bolagen använder olika system för datahantering så har variablerna nedan behandlats så att de har samma betydelse:

Kön
Individens kön.

Insjuknandeålder
Individens ålder vid insjuknande, 18-65.

Bolag
Antingen Förenade Liv, Folksam Liv, Salus Ansvar eller KP. Bolagen tilldelas slumpmässigt värdet 1, 2, 3 eller 4 på grund av sekretesskäl.

Kommungrupp
Individens bostadsort indelad i 10 grupper enligt Sveriges kommuner och landstings kommunindelning år 2011. Denna variabel är den mest lättillgängliga i vårt fall när vi vill undersöka om bostadsorten har en effekt på sjuktiden. En detaljerad beskrivning av grupperna finns i bilaga A. Indelningen är till stor del baserad på folkmängd och ger även en idé om sysselsättning, faktorer som skulle kunna tänkas påverka sjuktiden.

Antal tidigare sjuktillfällen
Antal gånger individen tidigare har registrerats som sjuk hos bolaget.

Sjukersättningsfaktor
Denna variabel beskrivs mer detaljerat i bilaga B och anger andel av befolkningen som nybeviljats sjukersättning och aktivitetsersättning vid tillfället då kunden insjuknade. Försäkringskassan beviljar sjukersättning till individer i åldern 30-64 år som troligtvis aldrig kommer arbeta heltid igen på grund av sjukdom, skada eller funktionsnedsättning. Aktivitetsersättning är motsvarigheten för individer i åldern 19-29, och ges till personer som inte kommer arbeta heltid under minst ett år på grund av sjukdom, skada eller funktionsnedsättning. Både sjuk- och aktivitetsersättning kräver att man har nedsatt arbetsförmåga till minst 25%.

Variabeln är alltså ett mått på andelen långtidssjukskrivna i landet vid tidpunkten då vår individ blev sjuk. Uppgiften om nybeviljade sjukersättningar kommer från Försäkringskassan och publiceras årligen, efter varje avslutat år, vilket innebär att 2012 års siffra blir tillgänglig under 2013. Försäkringsbolagens sjukdata uppdateras löpande, och därför förskjuts sjukersättningsåret ett år framåt, så att sjukersättningsgraden år 2012 används tillsammans med bolagets data för år 2013.

Orsaken till att denna externa variabel undersöks är att Försäkringskassan bedömer vem som räknas som sjuk, vilket har en direkt påverkan på sjukfall. Det finns skäl att tro att när antal långa sjukfall ökar i hela landet, så borde de öka även i försäkringsbeståndet. När nybeviljade ersättningar går upp på grund av att Försäkringskassan har gjort det enklare att bli klassad som sjuk så skulle det vara rimligt att både antal sjukfall och sjuklängd ökar

i det försäkrade beståndet. Höga värden på variabeln sjukersättning skulle då motsvaras av längre observerade sjuktider T.

Karenstid
Karenstiden anger den tid som en kund måste vara sjuk innan denne kan börja få ersättning. Det finns karenser som är fastställda och anges i månader och R-karens, som är rörlig och innebär tid fram till att man får sjukersättning.

Sjuktid
Denna är vår responsvariabel som anger antal dagar kunden varit sjuk och beräknas som antal dagar mellan insjuknande och avveckling. Sjuktiden innehåller karenstiden och representerar alltså total sjuklängd.

Censor
Anger om observationen är högercensurerad eller ej. 0 innebär en censurerad observation och 1 en faktisk avveckling enligt figur 1.

För ett sjukfall som både har påbörjats och avslutats under observationsperioden kan den totala sjuktiden observeras, vi vet exakt hur länge personen varit sjuk. Däremot känner vi inte till hur lång sjukperioden faktiskt är om ett sjukfall fortfarande pågår vid sista observationstillfället, 18 november 2013, eller om det avslutas av en annan anledning än att personen tillfrisknar eller dör. Dessa fall kallas högercensurerade och för dem vet vi bara att personen har varit sjuk minst så många dagar som observerats och antagligen fler. 47% av observationerna är högercensurerade och de är ofta långa sjukfall som inte hunnit avslutas, vilket innebär att antalet sjukdagar skulle underskattas om de inte inkluderades i analysen. Ett lämpligt sätt att få med dem beskrivs i avsnitt 3.

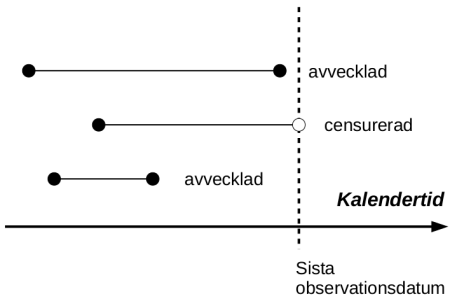

Figure 1: Sjuklängd efter kalendertid

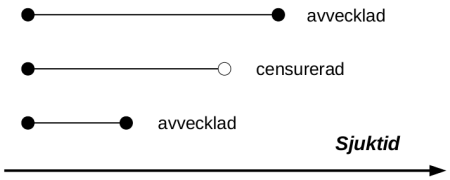

Figure 2: Sjuklängd efter sjuktid

Figur 1 och 2 visar två sätt att se på samma observationer, sjuktiderna i förhållande till kalendertiden i den första bilden och längden för sjukfallen i förhållande till varandra, som modelleras i uppsatsen, i den andra. I vårt data känner vi alltid till när sjukfallet började, men vid högercensurering saknar vi information om när sjukfallet avslutades. Censurorsakerna som finns i datama-

terialet är avveckling på grund av observationsperiodens slut, ersättningsperiodens slut eller okänd orsak. Ersättningsperioden i ett försäkringsavtal anger hur länge kunden längst kan få utbetalning och kan exempelvis vara 6 månader, 5 år eller fram till avveckling. Den är dock inte med som en förklarande variabel på grund av att den inte går att ta fram för alla avtal.

3 Teori

3.1 Sannolikhetsfördelning, överlevnadsfunktion, hazardfunktion

Vi modellerar sjuktiden T med överlevnadsanalys, som har både parametriska och icke-parametriska (Cox) modeller. Vi är intresserade av att prediktera förväntad sjuktid för individer med en viss uppsättning av förklarande variabler. Detta kan endast göras med hjälp av parametriska modeller där en fördelning antas för responsvariabeln, så vi utesluter de icke-parametriska modellerna.

Innan vi går in djupare på modellerna gör vi följande definitioner:

Sjuktiden T i denna uppsats har namnet livstid inom överlevnadsanalysen. Till T hör fördelningsfunktionen F och täthetsfunktionen f

$$F(t) = P(T < t) = \int_0^t f(u)du$$

Överlevnadsfunktionen $S(t)$ är vår avvecklingsfunktion och anger sannolikheten att inte ha avvecklats vid tiden t. Den ges av

$$S(t) = P(T \geq t) = 1 - F(t)$$

Det finns även en tillhörande hazardfunktion $h(t)$ som anger den momentana risken, eller hazarden, för att individen avvecklas i det korta tidsintervallet $(t, t + \delta t)$ givet att denne var sjuk vid tiden t.

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

Vi har

$$P(t \leq T < t + \delta t | T \geq t) = \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)}$$

Detta ger

$$h(t) = \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \cdot \frac{1}{S(t)} = \frac{f(t)}{S(t)}$$

Från detta kan vi få

$$
\begin{aligned}
h(t) &= \frac{\frac{d}{dt}F(t)}{S(t)} \\
&= \frac{\frac{d}{dt}(1 - S(t))}{S(t)} \\
&= \frac{S'(t)}{S(t)}
\end{aligned}
$$

Alltså får vi sambandet

$$h(t) = -\frac{d}{dt} log S(t) \tag{1}$$

samt

$$S(t) = e^{-\int_0^t h(u)du}$$

Så fort någon av funktionerna $S(t)$, $f(t)$ eller $h(t)$ är känd kan de andra två härledas.

3.2 Parametrisk proportionell hazard (PH) modell

De enklaste parametriska modellerna har namnet parametric proportional hazards model och definieras genom sambandet

$$h_i(t) = h_0(t) exp(\eta_i)$$

$h_0(t)$ kallas baseline hazard och är hazarden för individen med alla förklarande variabler $x = 0$. $h_0(t)$ antas komma från en viss sannolikhetsfördelning och som namnet proportional hazard antyder innebär modellen att kvoten mellan baselinehazarden och hazarden för individ i, med de förklarande variablerna $x_{1i}, ..., x_{pi}$, är proportionell vid varje tidpunkt.

$$\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

Koefficienterna β_j anger effekten som de förklarande variablerna $x_{1i}, ..., x_{pi}$ har på hazarden. $j = 1, ..., p$.

De tillgängliga fördelningarna i SAS som kan anpassas som PH modeller är Weibull och exponential. Överlevnads- och hazardfunktioner för dessa finns i appendix C.

I våra två PH modeller kan hazardfunktionen endast öka eller minska monotont. När det gäller sannolikheten att avvecklas från tillståndet sjuk kan man tänka sig att den inte är monoton, utan istället är låg precis efter insjuknandet för att sedan öka eftersom man inte brukar tillfriskna direkt efter att ha blivit sjuk. En person som har varit sjuk länge kan tänkas fortsätta vara sjuk med större sannolikhet än någon som varit sjuk en vecka, exempelvis i influensa, och avvecklingssannolikheten skulle då kunna minska ju längre man är sjuk. Den skulle sedan kunna öka igen då man har varit sjuk länge om dödligheten hinner slå igenom. Accelerated Failure Time (AFT) modeller tillåter fördelningar med icke-monotona hazardfunktioner.

3.3 Accelerated Failure Time (AFT) modell

AFT modeller anger hur variabler påverkar sjuktiden istället för hazarden och är därför mer lättolkad i vårt fall. För att predikterade värden för T alltid ska vara positiva används en log transformation. Den loglinjära AFT modellen har utseendet

$$log(T_i) = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\epsilon_i$$

Sjuktiden är $T_i = exp(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\epsilon_i)$. α_i-parametrarna mäter effekten som motsvarande förklarande variabler x_i har på sjuktiden. μ är interceptet och σ kallas skalparametern. ϵ_i är stokastiska variabler som mäter avvikelser från modellens linjära del av $log(T_i)$. ϵ_i antas vara oberoende och ha en viss sannolikhetsfördelning som ger T_i sin fördelning, men oftast pratar man direkt om fördelningar för T_i. Tabellen nedan summerar de tillgängliga fördelningarna för T_i i SAS och de motsvarande fördelningarna för ϵ_i.

Table 1: Fördelningar för T och ϵ

Fördelning för T	Fördelning för ϵ
Exponential	Extreme value (1 parameter)
Weibull	Extreme value (2 parameters)
Log-logistic	Logistic
Log-normal	Normal
Gamma	Log-Gamma

Sambandet mellan ϵ_i och T_i är

$$\begin{aligned} S_i(t) &= P(T_i \geq t) \\ &= P(log(T_i) \geq log(t)) \\ &= P(\mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi} + \sigma\epsilon_i \geq log(t)) \\ &= P(\epsilon_i \geq \frac{log(t) - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}) \end{aligned}$$

Om $S_{\epsilon_i}(\epsilon)$ betecknar överlevnadsfunktionen för ϵ_i så ges överlevnadsfunktionen för T_i av

$$S_i(t) = S_{\epsilon_i}(\frac{log(t) - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}) \tag{2}$$

Om baselinefunktionerna för överlevnads-, hazard- och täthetsfunktionen betecknas $S_0(t)$, $h_0(t)$ och $f_0(t)$ får vi följande relationer till en individ med uppsättningen $x_1, ..., x_p$ av förklarande variabler:

$$S_i(t) = S_0(\frac{t}{e^{\eta_i}}) \tag{3}$$

$$h_i(t) = \frac{1}{e^{\eta_i}} h_0(\frac{t}{e^{\eta_i}})$$

$$f_i(t) = \frac{1}{e^{\eta_i}} f_0(\frac{t}{e^{\eta_i}})$$

$\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi}$ och $\frac{1}{e^{\eta_i}}$ kallas accelerationsfaktorn på grund av att livstiden för individ i är e^{η_i} gånger livstiden för en individ med alla förklarande variabler lika med 0.

Hazardfunktionen $h_i(t)$ fås på följande sätt. Den kumulativa hazardfunktionen $H_i(t)$ är definierad som $-logS_i(t)$. Formel (2) ger

$$\begin{aligned} H_i(t) &= -logS_{\epsilon_i}(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}) \\ &= H_{\epsilon_i}(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}) \end{aligned}$$

Differentiering av $H_i(t)$ med avseende på t ger hazardfunktionen

$$h_i(t) = \frac{1}{\sigma t} h_{\epsilon_i}(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}) \tag{4}$$

3.3.1 Residualer

Residualerna nedan anger skillnaden mellan skattat och observerat värde och används därför till att bedöma modellanpassningen.

Standardiserade residualer
I AFT modellen

$$log(T_i) = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_p x_{pi} + \sigma\epsilon_i$$

ges de standardiserade residualerna av

$$r_{S_i} = \frac{logt_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \ldots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}}$$

Residualerna har samma fördelning som ϵ_i om den valda modellen är korrekt och den skattade överlevnadsfunktionen för residualerna borde då vara mycket lik överlevnadsfunktionen för ϵ_i. Collett bevisar i [1, s.112-113] att $-logS(t)$ är exponentialfördelad med väntevärde 1, oavsett vilken underliggande fördelning $S(t)$ har. Alltså är $-logS_{\epsilon_i}(t)$ exponentialfördelad med väntevärde 1 om den underliggande modellen är riktig. Denna egenskap kommer väl till hands vid modellval som vi kommer se nedan.

Cox-Snell residualer

$$r_{C_i} = \hat{H}_i(t_i) = -log\hat{S}_i(t_i)$$

där

$$\hat{S}_i(t_i) = S_{\epsilon_i}(\frac{logt_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \ldots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}})$$

Vi ser att sambandet mellan Cox-Snell residualerna och de standardiserade residualerna är $r_{C_i} = -logS_{\epsilon_i}(r_{S_i})$. Att avgöra om de standardiserade residualerna har en viss fördelning är alltså detsamma som att avgöra om Cox-Snell residualerna är exponentialfördelade med väntevärde 1. En plot av $\hat{H}_i(t_i) = -log\hat{S}_i(t_i)$ mot r_{C_i} borde ge en rak linje med lutning 1 och intercept 0 om den anpassade modellen är korrekt [1, s.236].

3.3.2 Prediktion

Den p:te percentilen för individ i, $t_i(p)$ är sådan att

$$S_i(t_i(p)) = \frac{100 - p}{100}$$

Percentilen blir då

$$t_i(p) = S_i^{-1}(\frac{100 - p}{100}) \tag{5}$$

Vi använder 50:e percentilen $t_i(50)$, som är medianen, för att prediktera den förväntade sjuklängden för individ i.

Formler för percentiler, överlevnadsfunktioner och hazardfunktioner för specifika AFT fördelningar finns i appendix C.

3.4 Parameterskattningar

För både PH och AFT modellerna gäller att alla våra förklarande variabler till en början är indelade i grupper och när deras parametrar ska skattas väljer vi ett referensvärde per variabel. Om det finns p stycken grupper inom en variabel kommer vi få $p - 1$ stycken parameterskattningar för den variabeln, som anger påverkan på $log(T_i)$ jämfört med referensuppsättningen. Varje variabels referenscell väljs som den cell med flest observationer för att få så säkra skattningar som möjligt.

I de parametriska modellerna skattas parametrarna α_i genom maximum likelihood metoden med hänsyn till censurerade data.

Låt t_i representera tiden för avveckling alternativt censurering och låt indikatorvariabeln δ_i vara 0 om observationen är censurerad och 1 annars. För ocensurerade data är likelihoodfunktionen helt enkelt

$$L = \prod_{i=1}^{n} f_i(t_i)$$

där indexeringen för f indikerar att varje individ har olika sannolikhetsfunktioner som beror på de förklarande variablerna. För observationer som har censurerats vid tiden t_i vet vi att den individens sjuktid är längre än t_i. Sannolikheten för detta är ju $P(T \geq t_i) = S(t_i)$, vid tiden t_i, så likelihoodfunktionen

blir:

$$L = \prod_{i=1}^{n} = [f_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i}$$

Koefficientskattningarna fås genom att maximera denna funktion med avseende på parametrarna α, vilket är ekvivalent med att maximera loglikelihoodfunktionen. Med andra ord löser vi funktionen

$$\frac{\partial logL}{\partial \alpha} = 0$$

α-värdena fås genom Newton Raphson metoden. Till vår hjälp har vi

$$U(\alpha) = \frac{\partial logL}{\partial \alpha}$$

$$I(\alpha) = \frac{\partial^2 logL}{\partial \alpha \partial \alpha'}$$

Newton Raphson algoritmen blir

$$\alpha_{j+1} = \alpha_j - I^{-1}(\alpha_j)U(\alpha_j)$$

där I^{-1} är inversen till I. Vi börjar med en samling startvärden α_0, som fås genom minsta kvadratmetoden, där alla observationer behandlas som om de vore ocensurerade. Dessa α_0 sätts in i högerledet i ekvationen och ger första iterationen α_1. α_1 sätts sedan in i högerledet, första- och andraderivatorna U och I beräknas om och ger α_2. Processen upprepas tills ändringen i parameterskattningarna från ett steg till nästa är mindre än 0.00000001.

Koefficienternas kovariansmatris ges av $-I^{-1}(\hat{\alpha}_j)$ och standardavvikelserna fås som roten ur diagonalelementen i denna matris.

3.5 Mått för modellbedömning

När vi ska försöka hitta en modell som ger bäst anpassning använder vi följande mått för att jämföra olika modeller.

3.5.1 Log-likelihood ratio (LR) test

LR testet kan användas vid jämförelse av nästlade modeller, där den ena är ett specialfall av den andra. Anta att modell 1 innehåller de förklarande variablerna $x_1, ..., x_p$ och modell 2 innehåller $x_1, \ldots, x_p, x_{p+1}, \ldots, x_{p+q}$. Modell 1 ingår i modell 2 och de kan jämföras genom skillnaden i loglikelihood med de skattade parametrarna insatta.

$$-2log\hat{L}_1 - (-2log\hat{L}_2) = -2(log\hat{L}_1 - log\hat{L}_2) = -2log\frac{\hat{L}_1}{\hat{L}_2}$$

Detta är LR statistikan för att testa nollhypotesen att de q parametrarna $\alpha_{p+1}, ..., \alpha_{p+q}$ i modell 2 är noll. Om nollhypotesen stämmer så är $-2log\frac{\hat{L}_1}{\hat{L}_2}$

asymptotiskt χ^2-fördelad med f frihetsgrader, där f är skillnaden mellan antal fria parametrar i de två modellerna. Nollhypotesen förkastas om

$$-2log\frac{\hat{L}_1}{\hat{L}_2} > \chi^2(f)$$

Vid modellval använder vi följande procedur som bygger på denna statistika och rekommenderas av Collett [1, s. 83].

1. Först anpassas modeller som innehåller en förklarande variabel var. Sedan anpassas en modell helt utan förklarande variabler och jämförs mot var och en av envariabelsmodellerna genom LR testet.

2. De variabler som är signifikanta undersöks sedan tillsammans i en modell. Här kan det hända att variabler som är signifikanta för sig själva blir osignifikanta när andra finns med och vi tar bort de som inte ger en signifikant ökning av $-2log\hat{L}$ när den exkluderas.

3. De variabler som inte visade sig vara signifikanta i steg 1 skulle kunna bli det i andras närvaro. Dessa läggs till en i taget till modellen i steg 2 och de som ger en signifikant minskning av $-2log\hat{L}$ behålls. Detta kan leda till att vissa termer från den valda modellen i steg 2 slutar vara signifikanta.

4. En sista kontroll görs så att inga termer kan exkluderas utan att signifikant öka $-2log\hat{L}$, och inga termer kan inkluderas som signifikant minskar $-2log\hat{L}$.

3.5.2 AIC

När vi jämför modeller där den ena inte ingår i den andra kan Akaikes informationskriterium, AIC, användas.

$$AIC = -2logL + 2k$$

$logL$ är modellens loglikelihood och k är antal parametrar. Eftersom loglikelihood mäter hur bra modellen beskriver data så är modellen bättre ju högre värde på loglikelihood. $2k$ finns med som en straffterm mot en för komplex modell och modellen är därmed bättre för lägre värden på AIC. AIC i sig säger inget om hur bra modellen är, värdet måste jämföras mellan olika modeller.

3.5.3 Wald χ^2 test

Vi använder ett Wald χ^2 test när vi vill avgöra om två koefficienter inom samma förklarande variabel skiljer sig åt. Nollhypotesen $\hat{\alpha}_1 = \hat{\alpha}_2$ ställs upp och förkastas om

$$\frac{(\hat{\alpha_1} - \hat{\alpha_2})^2}{Var(\hat{\alpha_1}) + Var(\hat{\alpha_2}) - 2Cov(\hat{\alpha_1}, \hat{\alpha_2})} > \chi^2(f)$$

Varians och kovarians hämtas från kovariansmatrisen I.

3.5.4 AFT egenskapen

För att bedöma om AFT modellen är lämplig kan en percentil-percentil plot användas.

Om vi låter $t_0(p)$ och $t_1(p)$ vara de p:e percentilerna för grupp 0 och grupp 1 så ger formel (5)

$$t_0(p) = S_0^{-1}(\frac{100-p}{100}), t_1(p) = S_1^{-1}(\frac{100-p}{100})$$

Från dessa två ekvationer ser vi att

$$S_1(t_1(p)) = S_0(t_0(p))$$

Eftersom $S_1(t) = S_0(\frac{t}{e^{\eta_i}})$ enligt formel (3) så får vi

$$S_1(t_1(p)) = S_0(\frac{t_1(p)}{e^{\eta_i}})$$

Vi får alltså

$$t_0(p) = S_0^{-1}(S_1(t_1(p))) = \frac{t_1(p)}{e^{\eta_i}}$$

När gruppernas percentiler plottas mot varandra borde vi se en rak linje genom origo om AFT modellen är korrekt. Lutningen för linjen blir en skattning av accelerationsfaktorn $\frac{1}{e^{\eta_i}}$. Percentilerna från de två grupperna kan fås genom Kaplan-Meier skattningen av överlevnadsfunktionerna för grupperna.

4 Statistisk analys

Vi har 58 453 observationer och sparar undan de 8 344 (cirka en sjundedel) nyaste sjukfallen i ett valideringsdataset för att kunna bedöma hur bra modellen kan prediktera nya sjuktider. Det återstår 50 109 stycken observationer att bygga modellen på.

Efter att ha valt en modell och skattat de okända parametrarna i modelldatasetet sätter vi in de kända variabelvärdena x_i från vårt undansparade valideringsdataset och jämför de faktiska utfallen för sjuktiden T_i med modellens prediktioner. Prediktionerna är 50:e percentilen (medianen) som beräknas genom formel (5).

4.1 Inledande undersökning av data

Som ett första steg undersöker vi variablerna grafiskt genom att titta på överlevnadsfunktionen $S(t)$ i modelldatasetet med 50 109 observationer. Här skattas $S(t)$ med den icke-parametriska Kaplan-Meier metoden för en variabel i taget utan hänsyn till de andra:

$$S(t) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i})$$

Detta är produkten av andel sjuka individer som fortfarande befinner sig under risk och inte har avvecklats. d_i är antal händelser, i vårt fall avvecklingar, som inträffat vid tiden t_i, n_i är antal individer under risk precis innan tiden t_i. En individ under risk är någon som fortfarande finns kvar i beståndet och inte har avvecklats eller censurerats. Varje tidpunkt t motsvarar en viss sjukdag.

Graferna 3-9 nedan visar $S(t)$ för olika grupper inom varje variabel. På x-axeln finns sjuktiden i dagar och på y-axeln finns $S(t)$. Högercensurerade observationer markeras med cirklar.

Kön
Det finns 31 394 (63%) kvinnor och 18 715 (37%) män i datamaterialet. Kaplan-Meier skattningar av överlevnadsfunktionen $S(t)$ i figur 3 för kvinnor respektive män ges av:

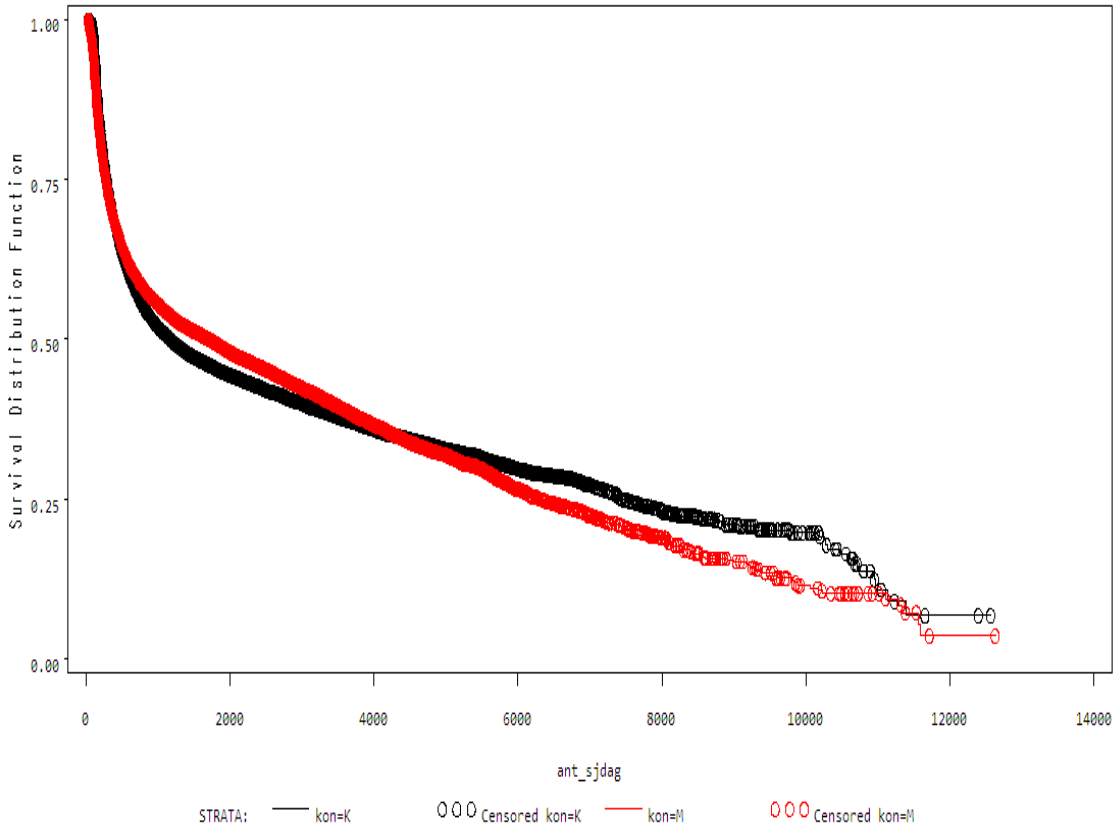

Figure 3: Överlevnadsfunktion per kön

Överlevnadsfunktionen ser ut att vara olika för könen, trots att gaferna korsar varandra två gånger. Här finns dock inte effekten av de andra variablerna med och under deras inflytande kan det se annorlunda ut.

Insjuknandeålder
Vi vill undersöka om denna variabel kan anpassas som linjär och kontinuerlig och börjar därför med att dela upp åldern i grupper om femårsintervall. Dessa modelleras senare tillsammans med andra variabler och vi kan då se om åldersgruppernas påverkan på sjuktiden visar ett linjärt mönster genom gruppnivåerna. Är sambandet icke-linjärt behåller vi gruppindelningen och får då en skattad koefficient för varje åldersgrupp. Åldersgrupp 1 representerar 18-25 år och har fått fler åldrar på grund av lite data och vi får 9 åldersgrupper enligt tabellen nedan.

Table 2: Frekvenser för åldersgrupp vid insjuknande

Gruppnivå (åldersintervall)	Antal observationer	Procent
1 (-25)	499	1%
2 (26-30)	1637	3%
3 (31-35)	3245	6%
4 (36-40)	4107	8%
5 (41-45)	4993	10%
6 (46-50)	7016	14%
7 (51-55)	10271	21%
8 (56-60)	12545	25%
9 (61-65)	5796	12%

Figur 4 visar en tydlig trend till att sannolikheten att fortsätta vara sjuk ökar med åldern, samtidigt som den avtar tidigare med högre ålder. Orsaken till att avvecklingstakten ökar tidigare för högre åldrar skulle kunna bero på att dödligheten börjar få en inverkan. För åldersgrupp 9, 61-65 år, följer $S(t)$ samma form som de övriga avvecklingsfunktionerna fram till tidpunkten 1000. 1000 sjukdagar motsvarar ungefär 2.7 år och individen som insjuknade vid åldern 65 är då nära 68 år. Eftersom sjukdom bidrar till högre dödsrisk kan det vara rimligt att avvecklingstakten ökar genom ökad dödsrisk.

Bolag
Tabell 3 nedan visar hur många observationer som finns för respektive bolag.

De fyra bolagen försäkrar olika kunder, som skulle kunna tillhöra olika riskgrupper. Detta är den enda variabeln i vårt data som kan fånga in eventuella riskskillnader mellan bolagen.

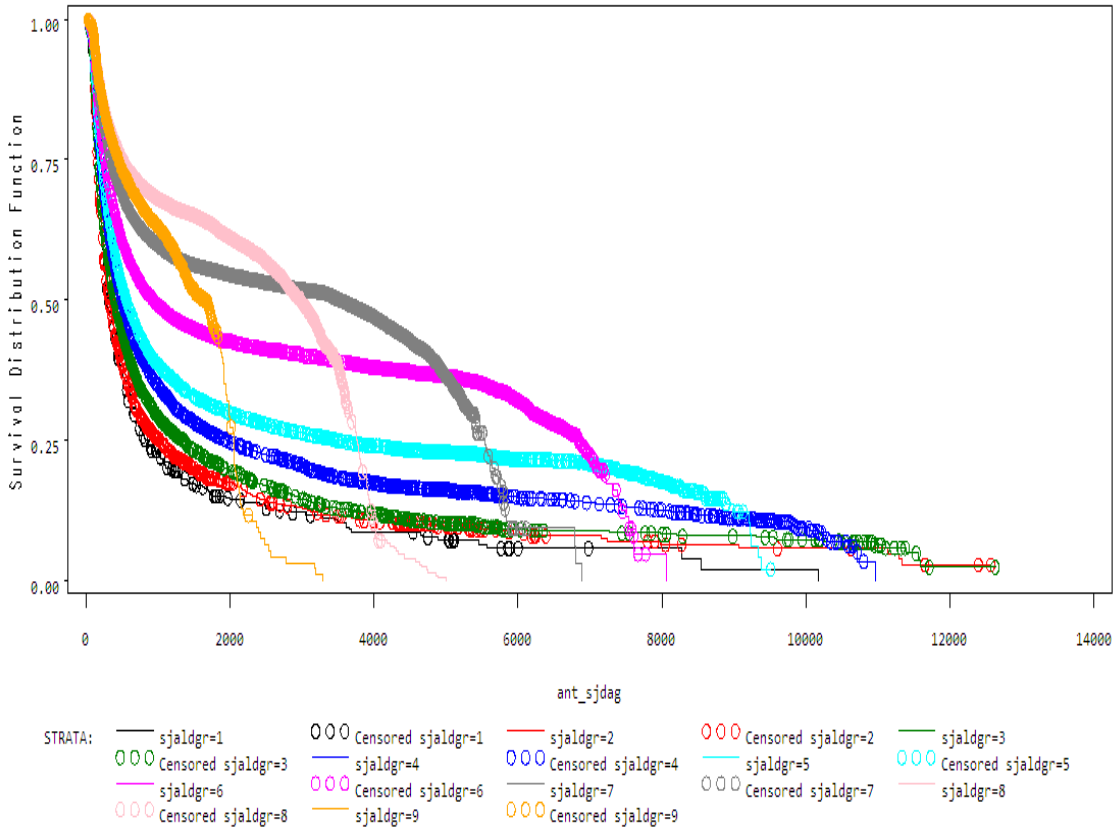

Figure 4: Överlevnadsfunktion per sjukåldersgrupp

Table 3: Frekvenser för bolag

Bolag	Antal observationer	Procent
1	4306	9%
2	6419	13%
3	16901	34%
4	22483	45%

I figur 5 verkar bolag 4:s kunder ha högre sannolikhet att kvarstå som sjuka vid varje tidpunkt. $S(t)$ för bolag 3 har samma form som bolag 4, men är lägre än bolag 1 och 2 i början för att bli högre efter cirka 2000 dagar. Detta är en antydan till att hazarden inte är konstant.

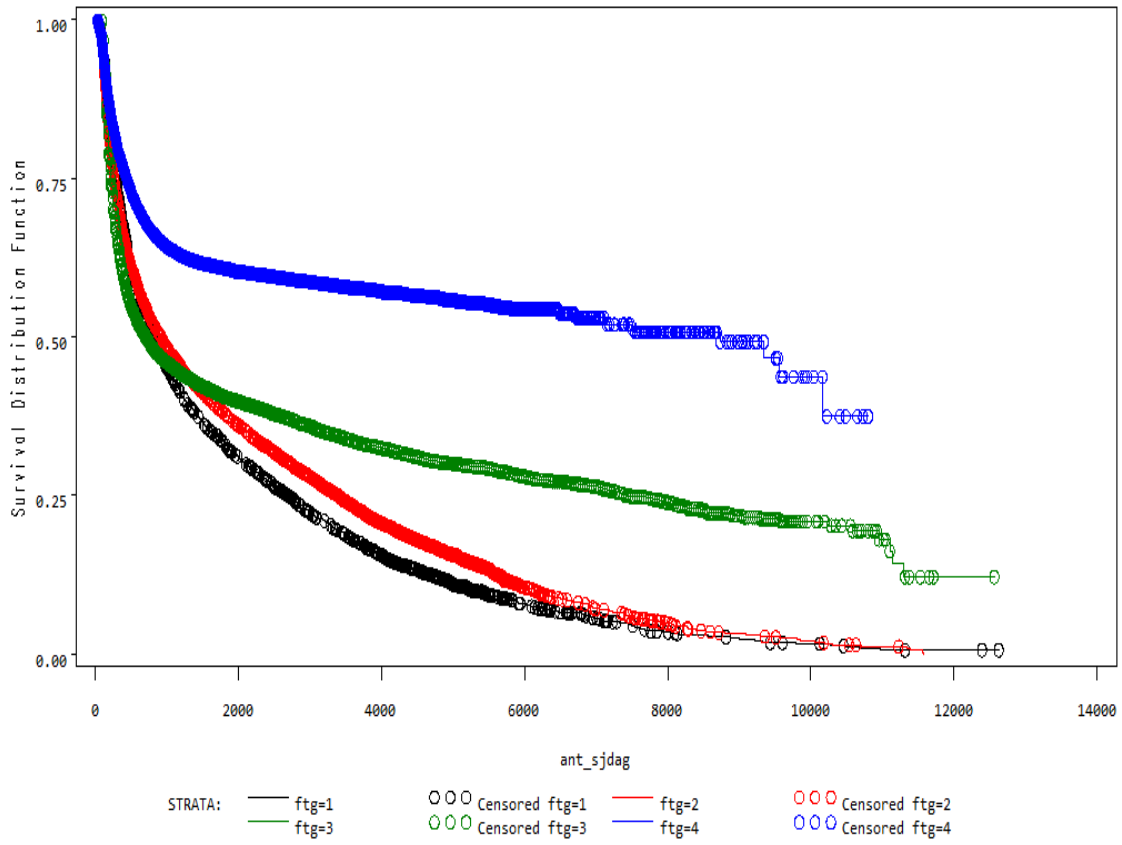

Figure 5: Överlevnadsfunktion per bolag

Karenstid

Karenstyperna 1, 3, 12 och R har tillräckligt många observationer för att ge säkra skattningar och dessa visar i figur 6 att överlevnadsfunktionen ligger högre för längre karenser, högst för den rörliga R-karensen. En möjlig orsak skulle kunna vara att personer med längre karenser har varit utan betalning under en lång tid och har lägre motivation att tillfriskna eftersom de vill få ut mer.

Det kan verka konstigt att $S(t)$ för karensen 8 månader ligger under 1 och 3 månader, men det finns endast 34 observationer i denna grupp och det skulle kunna se annorlunda ut med fler observationer. $S(t)$ bygger på andelen forsatt

Table 4: Frekvenser för karenstyper

Karenstyp (månad)	Antal observationer	Procent
1	5078	10%
3	41836	83%
5	5	0.0%
6	6	0.0%
8	18	0.0%
9	1	0.0%
12	1897	3.8%
24	5	0.0%
36	26	0.1%
48	33	0.1%
R	841	1.7%
Information saknas	363	0.7%

sjuka individer, och skulle vi till exempel ha fler observationer men samma antal avvecklingar så skulle resultatet bli att $S(t)$ ligger högre.

En svårighet med att undersöka olika karenser är just att det finns väldigt få observationer för många av karenstyperna. För R-karenser vet man inte på förhand hur lång tid efter insjuknandet som pengar börjar betalas ut och det skulle vara intressant att undersöka skillnaden mellan den rörliga R-karensen och de fastställda karenserna. R-karensen gäller fram tills dess att kunden beviljats sjukersättning. Denna period kan vara kort, men är normalt ganska lång, vilket också innebär en längre sjuktid. Vi låter variabeln vara kategorisk med värde 1 för R-karens och 2 för fastställda karenser. Trots att endast 1.7% av observationerna har R-karens så borde 841 observationer räcka för att bygga en skattning på.

Kommungrupp
Tabell 5 visar hur många observationer som finns i varje kommungrupp.

Table 5: Frekvenser för kommungrupper

Kommungrupp	Antal observationer	Procent
1	7361	15%
2	8019	16%
3	16082	32%
4	1458	3%
5	3466	7%
6	1661	3%
7	3756	8%
8	1118	2%
9	4271	9%
10	2917	6%

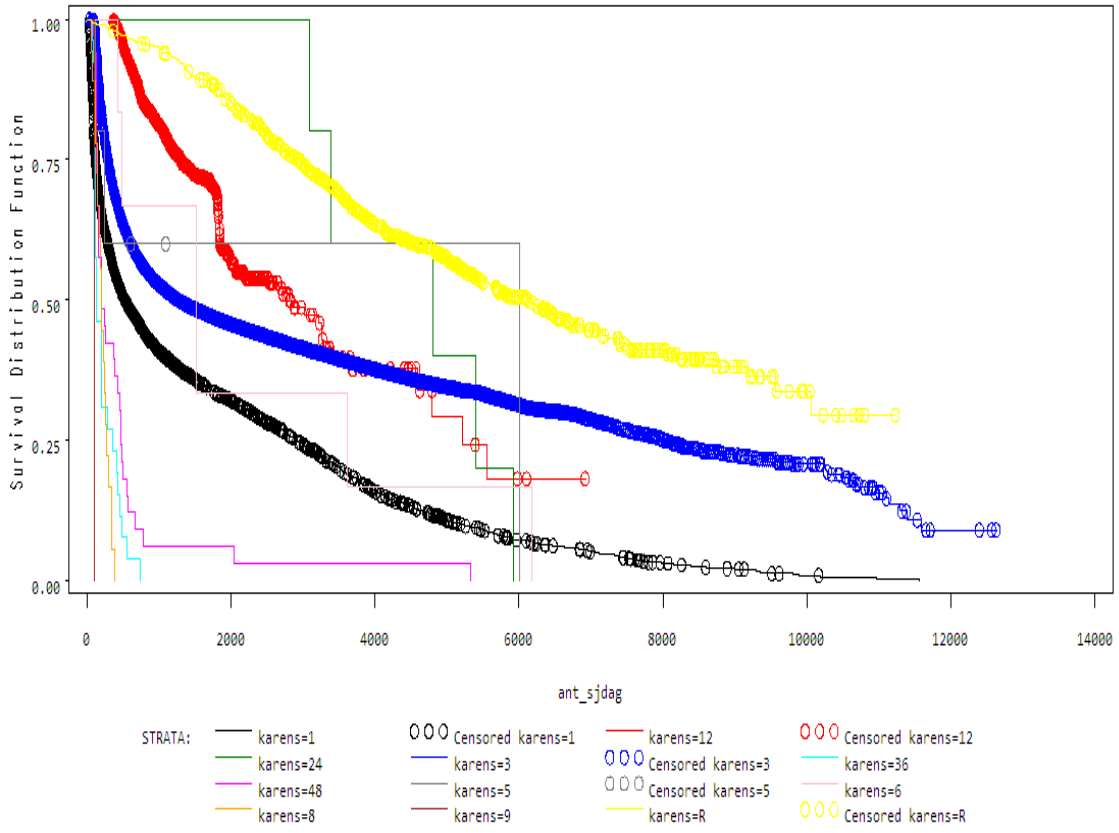

Figure 6: Överlevnadsfunktion per karenstyp

Figur 7 tyder på att det finns skillnader mellan kommungrupperna, men det är inget entydigt mönster.

Tidigare sjuktillfällen

I datasetet finns varje individ med en gång och den observation som finns i tabell 6 för varje enskild individ är den med individens högsta observerade sjuktillfälle. Orsaken till att varje individ endast finns med en gång är att observationerna inte skulle vara helt oberoende annars, vilket kan innebära att parameterskattningarna får en bias.

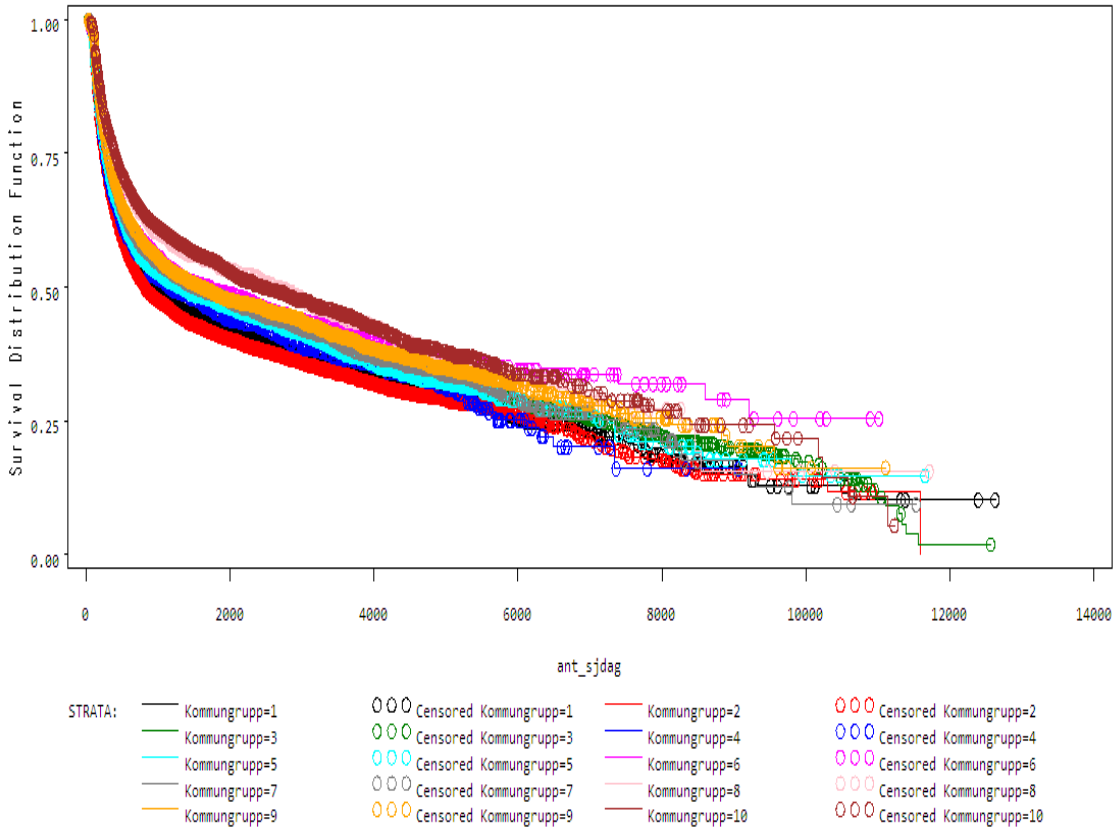

Figure 7: Överlevnadsfunktion per kommungrupp

Table 6: Frekvenser för tidigare sjuktillfällen

Tidigare sjuktillfällen	Antal observationer	Procent
0	44061	87.9%
1	4909	9.8%
2	858	1.7%
3	194	0.4%
4	53	0.1%
5	25	0.1%
6	3	0.0%
7	3	0.0%
8	3	0.0%

Eftersom sjuktillfällena 6, 7 och 8 endast har tre observationer var slås de ihop med grupp 5.

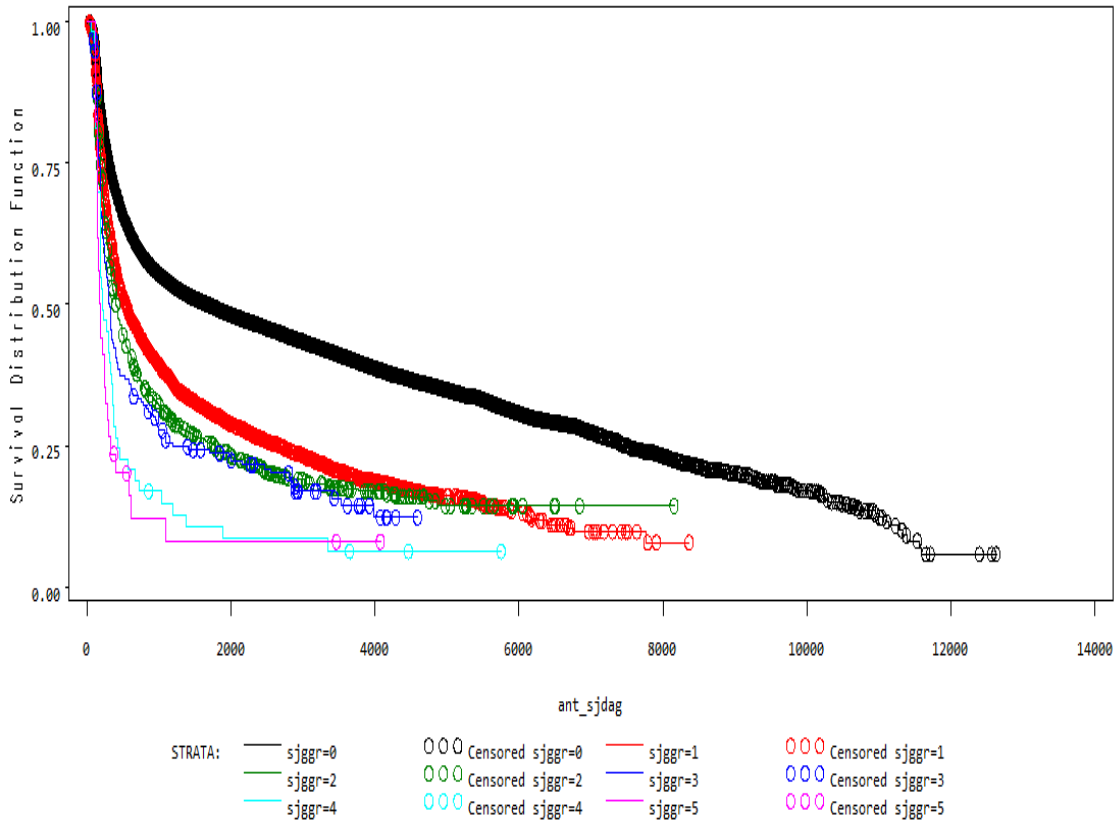

Figure 8: Överlevnadsfunktion per tidigare sjuktillfälle

Figur 8 visar att sannolikheten att kvarstå som sjuk är lägre ju fler gånger man varit sjuk tidigare. En möjlig förklaring till detta skulle kunna vara att individer insjuknar i samma sjukdom, vänjer sig fysiskt och psykiskt vid följderna och därför tillfrisknar snabbare. Det finns dock ingen självklar förklaring och istället för att spekulera vidare konstaterar vi bara att överlevnadsfunktionerna är olika för sjuktillfällena.

Sjukersättningsfaktor
Variabeln är en uppskattning av andel långtidssjuka i landet och antar värdena

Table 7: Frekvenser för sjukersättningsfaktor

Grupp (ersättningsandel %)	Antal observationer	Procent
1 (<0.4%)	6792	14%
2 (0.4-0.45%)	8251	16%
3 (0.45-0.5%)	8025	16%
4 (0.5-0.55%)	4790	10%
5 (0.55-0.6%)	3408	7%
6 (0.6-0.65%)	9141	18%
7 (0.65-0.7%)	735	1%
8 (0.7-0.75%)	6458	13%
9 (>0.75%)	2509	5%

0.15% - 0.82%. Vi undersöker om variabeln påverkar sjuktiden linjärt genom att dela in den i kategorier med 0.05%-intervaller. Eftersom det bara finns ett värde per år blir vissa av kategorierna tomma och vi måste slå ihop dem. De två lägsta kategorierna är från de senaste fyra åren och i vårt modelldataset finns dessa inte med, vilket innebär att maximum-likelihood inte kan skattas. Därför slås de ihop med grupp 3 som ligger närmast i värde. Detta ger 9 kategorier enligt tabell 7.

Överlevnadsfunktionerna ser ut att vara olika för olika faktorer i figur 9, men många av dem korsar varandra och det finns inte någon tydlig trend genom grupperna.

Hazardfunktionen
Vi skattar även en icke-parametrisk Hazardfunktion för sjuktiderna, som visas i figur 10.

Figur 10 visar att momentanrisken för avveckling börjar från 0 vid dag 0, når ett maximum efter cirka 120 dagar och minskar igen. $h(t)$ är den betingade sannolikheten att avvecklas vid tiden t givet att man har varit sjuk fram tills t och det verkar rimligt att denna ökar under ett tidigt skede då personen precis har insjuknat, för att sedan avta. En individ som har varit sjuk under en väldigt lång tid borde rimligtvis ha stor sannolikhet att fortsätta vara det och därmed ha en lägre hazard.

Hittills har vi tittat på varje variabel för sig, så resultaten kan se annorlunda ut när flera variabler modelleras tillsammans. Detta undersöks i nästa steg.

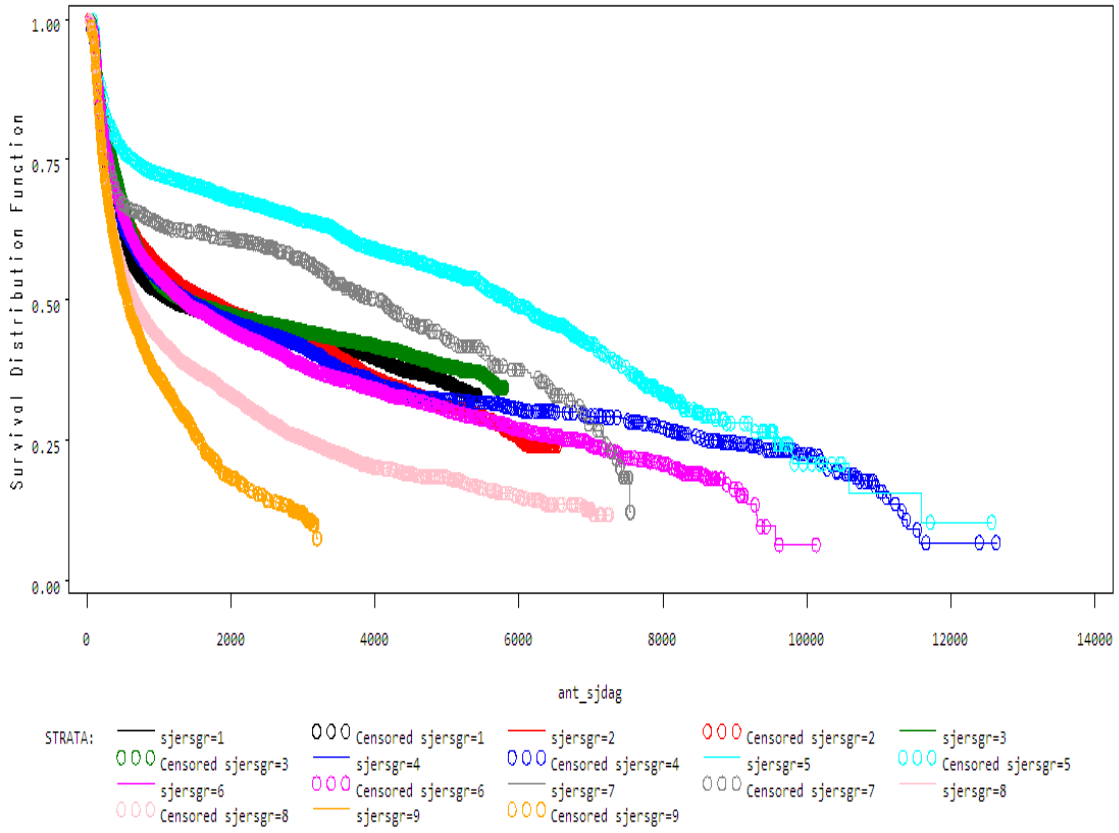

Figure 9: Överlevnadsfunktion per sjukersättningsfaktor

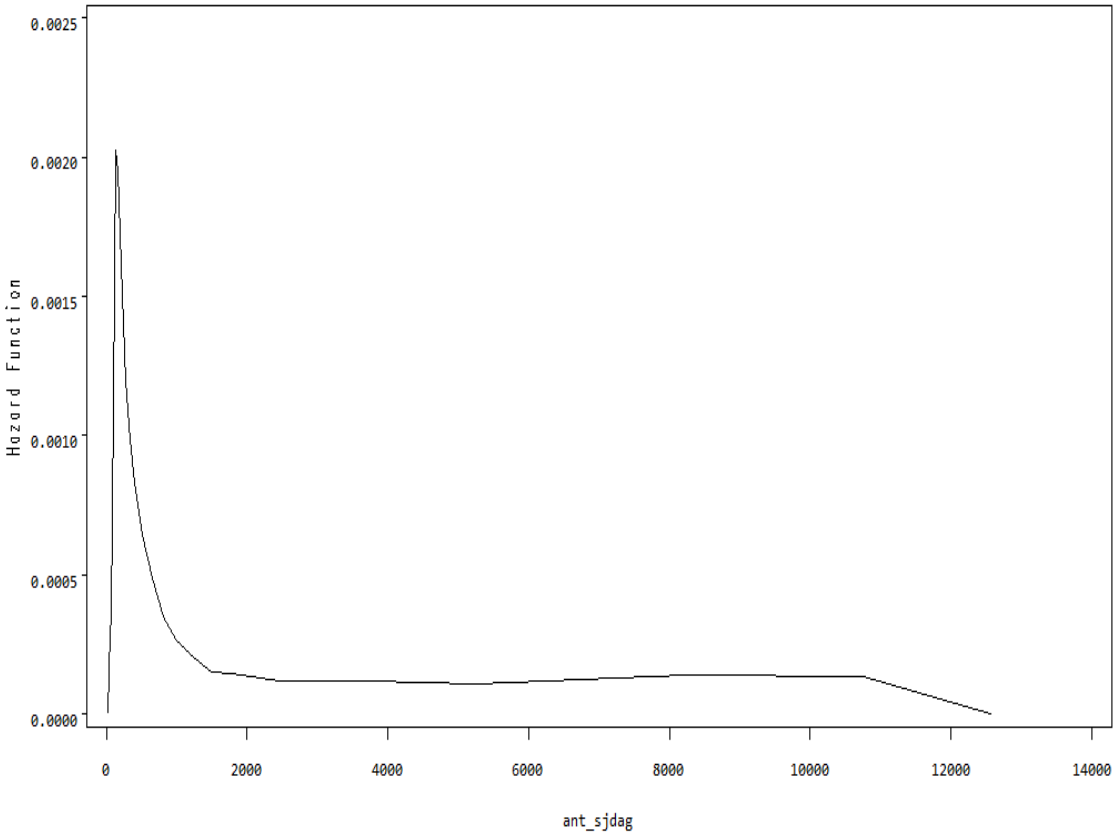

Figure 10: Hazardfunktion för sjuktid i dagar

4.2 Modellval

Vi har 50 109 observationer att modellera och 8 344 observationer undansparade i ett valideringsdataset för att kunna jämföra modellens prediktioner mot faktiskt utfall. Alla observationer är ordnade efter datum så att valideringssetet innehåller nyare sjukfall än modellsetet, på grund av att vi vill kunna använda modellen på nyare data. Följande förklarande variabler finns att undersöka:

I SAS kan följande AFT fördelningar anpassas till T_i: Exponential, generaliserad Gamma, Loglogistic, Logistic, Lognormal, Normal och Weibull. AFT Exponential och AFT Weibull fördelningarna kan göras om till PH motsvarigheterna genom formlerna i appendix C.

Table 8: Förklarande variabler

Variabel	Värde
Insjuknandeålder	1-9
Kön	0 (man), 1 (kvinna)
Kommungrupp	1-10
Sjukersättningsfaktor	1-9
Karensgrupp	1 (rörlig), 2 (fastställd)
Sjuktillfällen	0-8
Bolag	1-4

När den tomma modellen jämförs mot modellerna med en förklarande variabel var visar LR statistikan att alla sju förklarande variabler är signifikanta med p-värde < 0.0001. Detta gäller för alla fördelningar.

I vårt fall passar fördelningarna Gamma, Lognormal och Loglogistic bäst. Exempel på detta visas i tabell 9 nedan genom jämförelsen av AIC för modellerna utan förklarande variabler och modellerna med alla sju variabler.

Table 9: Jämförelse av AIC mellan olika modeller

Modell	AIC (inga variabler)	AIC (alla sju variabler)
Exponential	164 843	150 973
Weibull	155 063	144 236
Normal	557 702	543 977
Logistic	558 561	543 989
Gamma	140 722	136 155
Loglogistic	150 642	140 801
Lognormal	148 092	139 142

Här ser vi att den generaliserade Gamma modellen passar data bäst medans Lognormal och Loglogistic ger de näst och tredje bästa anpassningarna. Fördelningarna Normal och Logistic passar mycket dåligt.

Ett annat verktyg för modellbedömning är Cox-Snell residualer, r_{Ci}. En plot av $-log\hat{S}(r_{Ci})$ mot r_{Ci} borde visa en rak linje med lutning 1 om den valda modellen är korrekt, se avsnitt 3.3.1 på sida 10.

Residualerna i figur 11 tyder på att Lognormal och Loglogistic passar bäst och visar en rak linje med lutning 1. Även här ger Normal, Logistic och Exponential fördelningarna sämst anpassning och dessa modeller undersöks inte vidare. Vi kan påpeka att Exponentialfördelningen kräver en konstant hazardfunktion,

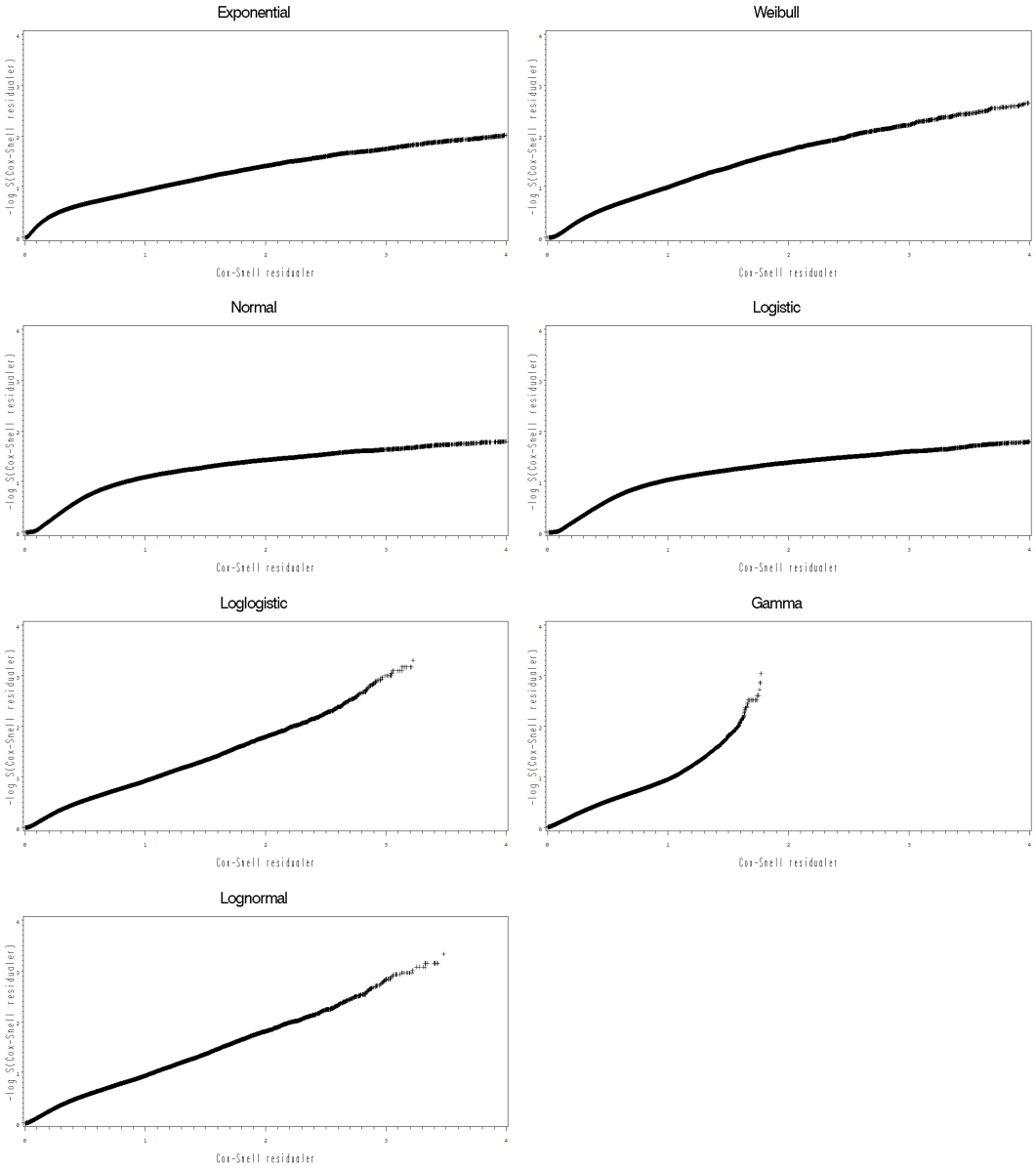

Figure 11: Jämförelse av Cox-Snell residualer, med alla sju signifikanta variabler inkluderade i modellen

vilket inte stämde med vår skattade hazardfunktion i figur 10.

AFT Weibull fördelningen ger inte en särskilt bra anpassning enligt AIC och Cox-Snell residualerna. Vi undersöker ändå om PH Weibull kan vara lämplig genom att testa om proportionell hazard verkligen gäller. Om hazarderna för olika grupper är proportionella så borde grupperna ha samma formparameter γ. Vi anpassar en Weibull modell till var och en av de g grupperna, med samma

linjära komponenter för alla grupper. Vi summerar $-2log\hat{L}$ för grupperna och kallar statistikan $-2log\hat{L}_1$. Sedan anpassas en Weibullmodell för alla grupperna tillsammans, som nu innehåller parametrar från gruppeffekten. Kalla $-2log\hat{L}$ från denna modell för $-2log\hat{L}_0$. Skillnaden mellan $-2log\hat{L}_0$ och $-2log\hat{L}_1$ beror på att vi inför begränsningen att formparametrarna är lika. Skillnaden $-2log\frac{\hat{L}_0}{\hat{L}_1}$ är χ^2 fördelad med $g-1$ frihetsgrader. Om denna är signifikant betyder det att proportionell hazard inte gäller. Nedan finns ett utdrag från resultatet:

Variabel	$-2log\frac{\hat{L}_0}{\hat{L}_1}$	Frihetsgrader
Kön	226	1
Karensgrupp	881	1
Kommungrupp	436	9

Resultaten är alla signifikanta med p-värde <0.0001 och proportionell hazard gäller inte. Vi utesluter alltså PH Weibull modellen.

Hittills har vi sett att Lognormal fördelningen ger bäst residualanpassning och näst bäst AIC värde. Den generaliserade Gammafördelningen hade bäst AIC, men ingen bra residualanpassning enligt figur 11 och eftersom den innehåller 3 parametrar (1 mer än övriga fördelningar) är den mindre komplicerade Lognormal fördelningen att föredra. I fortsättningen visas analysresultaten för denna fördelning, men vi behåller Loglogistic och gör resultatjämförelser mot denna.

Våra sju kategoriska variabler ger 38 parameterskattningar och vi vill undersöka om de istället kan anpassas som kontinuerliga med en linjär påverkan på sjuktiden. Företag, kön och karensgrupp är kategorivariabler och de resterande fyra variablerna tidigare sjuktillfällen, insjuknandeålder, sjukersättningsfaktor och kommungrupp undersöks närmare nedan.

Graferna 12-15 nedan visar variablernas effekt på sjuktiden och kommer från Lognormal fördelningen. Mönstret för Loglogistic är praktiskt taget identisk med mycket lika värden på koefficientskattningarna. Även Gammafördelningen är mycket lik men dess koefficientskattningar är något mindre. y-axeln visar koefficientskattningar och x-axeln variabelns nivåer.

Tidigare sjuktillfällen

Figur 12 visar en ganska linjärt nedgående trend över grupperna och vi väljer att anpassa variabeln som kontinuerlig.

Insjuknandeålder

Figur 13 visar att sjuktiden ökar linjärt med ökande åldersgrupper jämfört med referensgruppen 8. De högsta åldrarna i grupp 9 bryter dock mot trenden och har kortare sjuktider än grupp 8, vilket skulle kunna bero på ökad avveckling genom att dödligheten börjar slå igenom. Vi väljer därför att anpassa ålder som en kontinuerlig variabel och åldersgrupp 9 som en dummy variabel. Resultatet

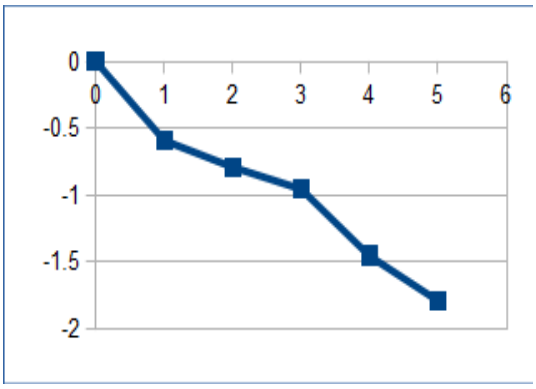

Figure 12: Tidigare sjuktillfällens effekt på sjuktiden

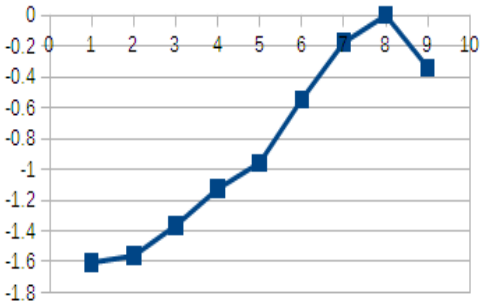

Figure 13: Insjuknandeålderns effekt på sjuktiden

blir då en koefficient som representerar förändringen av sjuktiden för varje enhetsökning av insjuknandeåldern mellan 18 och 65 år, och en koefficient som läggs till sjuktiden om insjuknandeåldern ligger i intervallet 61-65.

Sjukersättningsfaktor

Figur 14 visar att effekten av denna variabel är något överraskande. Det skulle vara rimligt att grupp 1, som har lägst andel nybeviljade sjukersättningar, även har kortast sjuktid och att sjuktiden ökar med ökad sjukersättning. Figur 14 visar istället att motsatsen gäller för några av grupperna. De två högsta ersättningsfaktorerna ger kortare sjuktid än den lägsta och för grupp 2-3, 5-6 och 7-9 gäller att sjuktiden minskar när sjukersättningen ökar. Variabeln är mer komplicerad än vi trott. Försäkringskassan försäkrar alla individer, vilket inte försäkringsbolag måste göra, och deras bestånd kan bestå av en helt annorlunda sammansättning individer än bolagens som gör att variabelns effekt blir skev. På grund av ändrade regler ett visst år kan många fler personer ur befolkningen med sämre risker och längre sjuktider få sjukersättning från Försäkringskas-

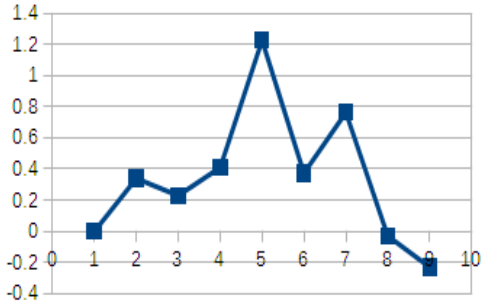

Figure 14: Sjukersättningsfaktorns effekt på sjuktiden

san. Försäkringsbolagen kan välja att inte försäkra personer med sämre risker och skulle då kunna ha ett bestånd med kortare sjuktider. Då skulle effekten av fler sjukersättningar visa kortare sjuktider enligt modellen. Variabeln sjukersättningsfaktor innehåller då inte all information som behövs för att kunna dra korrekta slutsatser.

Under analysens gång märker vi också att effekten från sjukersättningsfaktorn är skakig. När vi modellerar olika delar från datasetet, exempelvis en fjärdedel i taget, är effekten från sjukersättningsfaktorn väldigt olika för deldataseten, trots att koefficienterna är signifikant skilda enligt Wald χ^2 testet. Relationer mellan faktornivåerna varierar mellan dataseten, så att sjukersättningsfaktor 3 ger kortare sjuktider än sjukersättningsfaktor 4 i ett dataset men längre sjuktider i ett annat. Mönstrena för de andra förklarande variablerna är mer stabila och lika över deldataseten.

Att ta bort variabeln från modellen ger ingen stor påverkan på de andra koefficientskattningarna. Ingen av parametrarna ändrar tecken, effekten är relativt liten på både parameterskattningarna och deras standardavvikelser. Senare i uppsatsen kommer vi se att sjukersättningsfaktorn inte förbättrar prediktionen och vi väljer att utesluta variabeln.

Kommungrupp

Gruppindelningen för denna variabel är till stor del är baserad på fallande folkmängd genom grupperna, så vi testar om den kan anpassas som kontinuerlig. Mönstret för variabeln i figur 15 är hackig men det går att urskilja en uppåtgående trend genom grupperna. Att anpassa variabeln som kontinuerlig skulle ge en koefficient som innebär att sjuktiden ökar med varje kommungrupp från 1 till 10. Vi ser att sjuktiden för en del av grupperna istället går ner och vi skulle då få en fel uppskattning för dem. Trots detta så får vi bättre prediktioner, det vill säga fler prediktioner som ligger närmare faktiska värden beräknade genom formel (5) i avsnitt 3.3.2, när kommungrupp tillåts vara kon-

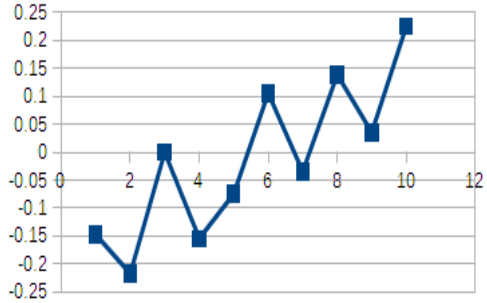

Figure 15: Kommungruppens effekt på sjuktiden

tinuerlig och eftersom en enklare modell med färre parameterskattningar alltid är att föredra väljer vi att låta kommungrupp vara kontinuerlig.

Sammanfattningsvis så exkluderas sjukersättningsfaktor. Variablerna sjuktillfällen, kommungrupp och insjuknandeålder anpassas som kontinuerliga och insjuknandeålder får också en dummyvariabel med värde 1 om åldern är 61-65.

När alla variablerna anpassas tillsammans kan det hända att någon av dem blir osignifikant och vi undersöker om detta är fallet. Resultatet av att ta bort en variabel i taget och jämföra LR statistikan från avsnitt 3.5.1 är att alla variabler fortfarande är signifikanta.

De förklarande variablerna har undersökts för Lognormal och Loglogistic fördelningarna, som ger väldigt lika resultat. Vi vill kunna skilja på dem och kan konstatera att AIC nu är 140497 för Lognormal och 142312 för Loglogistic, vilket antyder att Lognormal passar bättre. Vi undersöker också deras hazardfunktioner, som beräknas för medelvärdet av de kontinuerliga variablerna och det högsta värdet för de kategoriska.

Hazardfunktionerna för Lognormal och Loglogistic visas i figur 16 respektive 17. När dessa jämförs med den icke-parametriska hazarden i figur 10 ser vi att Lognormal är mycket lik denna medans Loglogistic inte fångar upp toppen vid de kortare sjuktiderna. Lognormal anpassningen ger en hazard som är låg i början för att nå ett maximum runt 120 dagar, samma tidpunkt som den icke-parametriska hazarden men med ett lägre värde på maximum. Efter denna topp genererar de samma hazard.

Med stöd av AIC och hazardfunktionerna väljer vi att utesluta Loglogistic fördelningen.

Sjuktiden T_i är Lognormal fördelad och modellen blir

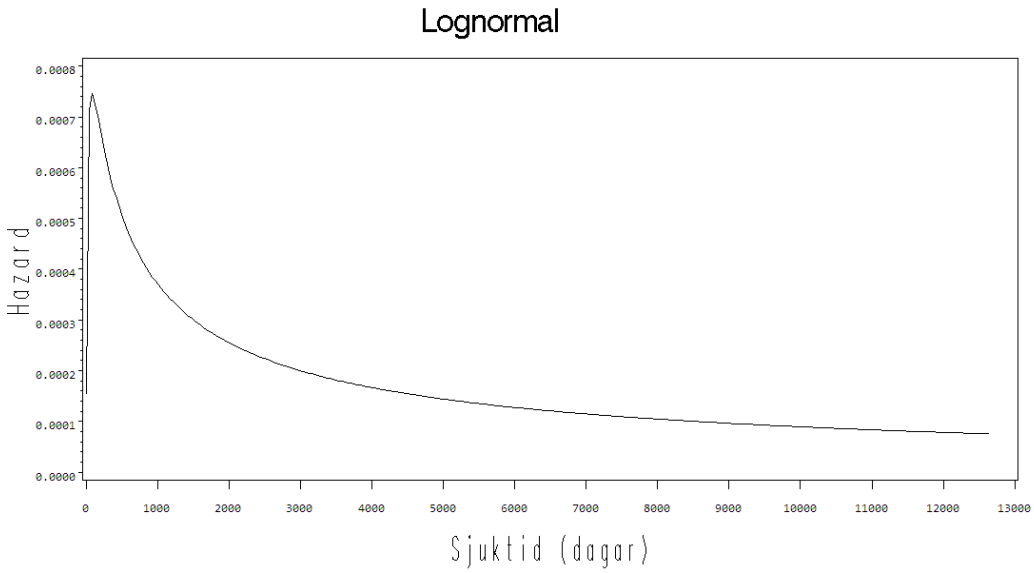

Figure 16: Lognormal hazardfunktion

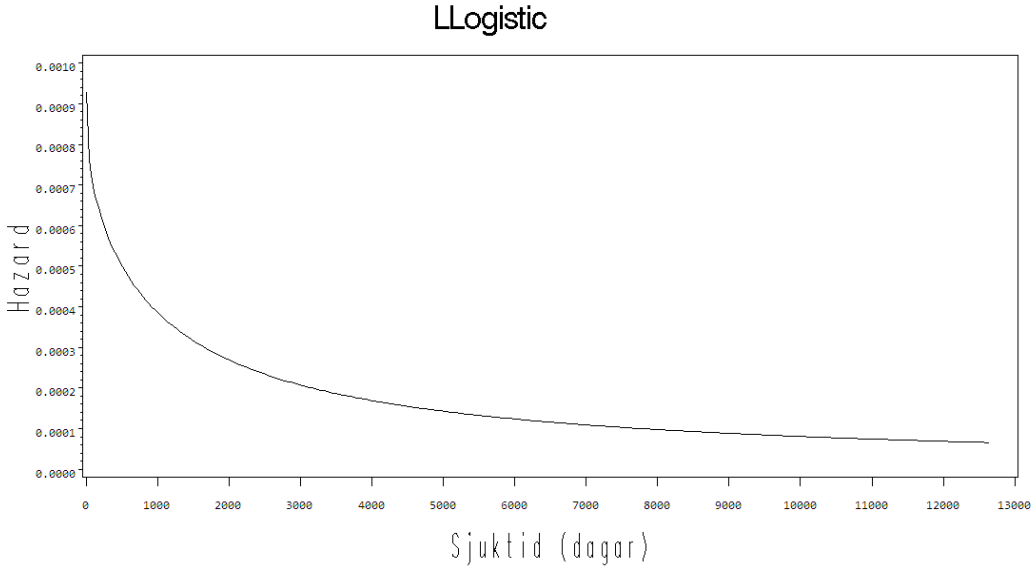

Figure 17: Loglogistic hazardfunktion

$$\begin{aligned} logT_i = 4.7261 + 0.0573 \cdot sjukålder - 0.5953 \cdot I_{sjukålder} - 0.1659 \cdot I_{kön} \\ -0.8269 \cdot I_{bolag1} - 0.5654 \cdot I_{bolag2} - 0.2568 \cdot I_{bolag3} \\ -0.4670 \cdot sjuktillfällen + 1.8908 \cdot I_{karensgrupp} + 0.0327 \cdot kommungrupp \end{aligned} \quad (6)$$

Referensvärdet för de kategoriska variablerna är fetstilsmarkerade nedan. För bolag är referensvärdet bolag 4.

$$I_{sjukålder} = \begin{cases} 1 & \text{om insjuknandeåldern är 61-65 år} \\ \mathbf{0} & \mathbf{annars} \end{cases}$$

$$I_{kön} = \begin{cases} 1 & \text{för män} \\ \mathbf{0} & \mathbf{för\ kvinnor} \end{cases}$$

$$I_{bolag1} = \begin{cases} 1 & \text{om observationen kommer från bolag 1} \\ 0 & \text{annars} \end{cases}$$

$$I_{bolag2} = \begin{cases} 1 & \text{om observationen kommer från bolag 2} \\ 0 & \text{annars} \end{cases}$$

$$I_{bolag3} = \begin{cases} 1 & \text{om observationen kommer från bolag 3} \\ 0 & \text{annars} \end{cases}$$

$$I_{karensgrupp} = \begin{cases} 1 & \text{för R-karens} \\ \mathbf{0} & \mathbf{för\ fastställda\ karenser} \end{cases}$$

Den procentuella effekten som de förklarande variablerna har på sjuktiden fås genom transformationen $100(e^{\hat{\alpha}_i} - 1)$. För en kontinuerlig variabel innebär $\hat{\alpha_i} = 0.5$ att en ökning av variabeln x_i med en enhet leder till att sjuktiden ökar med $100(e^{0.5} - 1) = 65\%$. För en kategorisk variabel innebär $\hat{\alpha_i} = 0.5$ att sjuktiden ökar 65% när variabeln x_i går från sitt referensvärde till det aktuella värdet.

Tabell 10 visar variablernas effekt på sjuktiden och den största effekten på sjuktiden enligt vår modell är att R-karenser ger 5.6 gånger längre sjuktider än fastställda karenser. Detta gäller för totala sjuklängder som har avvecklats och jämförelsevis är sjuktiden i genomsnitt 2.7 gånger längre för R-karenser i originaldata, som innehåller censurerade oavslutade fall.

För den valda Lognormal modellen blir avvecklingsfunktionen

$$S_i(t) = 1 - \Phi(\frac{logt - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \ldots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}})$$

Förväntad total sjuktid ges av medianen, den 50:e percentilen

$$t_i(50) = exp(\sigma\Phi^{-1}(0.5) + \mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi})$$

Table 10: De förklarande variablernas effekt på sjuktiden

Förklarande variabel	Variabeltyp	Koefficient-skattning	Procentuell effekt på sjuktiden	95%-igt konfidensintervall (undre, övre)
$Sjukålder$	kontinuerlig	0.0573	5.9%	(5.7%, 6.1%)
$I_{sjukålder}$	kategorisk	-0.5953	-44.9%	(-48.3%, -41.2%)
$I_{kön}$	kategorisk	-0.1659	-15.3%	(-18.3%, -12.1%)
I_{bolag1}	kategorisk	-0.8269	-56.3%	(-58.9%, -53.4%)
I_{bolag2}	kategorisk	-0.5654	-43.2%	(-46.1%, -40.1%)
I_{bolag3}	kategorisk	-0.2568	-22.6%	(-25.7%, -19.5%)
$Sjuktillfällen$	kategorisk	-0.4670	-37.3%	(-39.5%, -35.0%)
$I_{karensgrupp}$	kategorisk	1.8908	562.5%	(474.1%, 664.5%)
$Kommungrupp$	kontinuerlig	0.0327	3.3%	(2.7%, 4.0)

$\hat{\mu}$, $\hat{\alpha}_1$, ..., $\hat{\alpha}_p$ är samma parameterskattningar som i ekvation (6) och $\hat{\sigma} = 1.7679$. Φ känns igen som fördelningsfunktionen för en standard normal fördelning.

4.3 Modellutvärdering

Vi har anpassat en AFT modell och för en sådan gäller att percentilerna för två grupper plottade mot varandra ger en rak linje genom origo. Percentilerna för de två gruppernas överlevnadsfunktioner skattas med Kaplan-Meier metoden. SAS ger endast skattningar av percentilerna 25, 50 och 75.

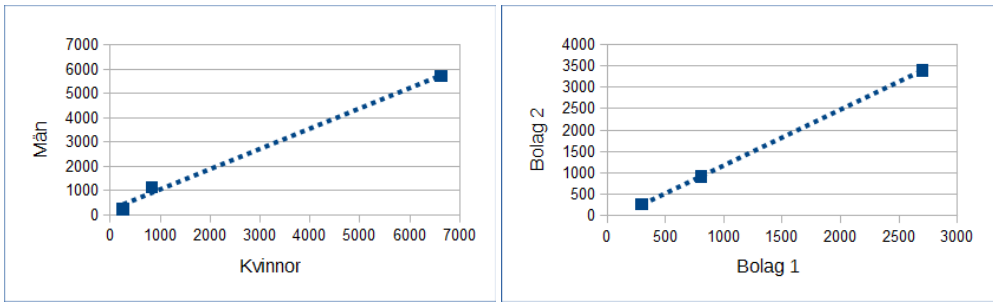

Figure 18: Percentil-percentil plot av grupperna män - kvinnor samt bolag 1 - bolag 2

Figur 18 visar hur P-P plottarna typiskt ser ut för vårt data. Tre punkter är inte mycket, och trots att den andra observationen för könsgrupperna avviker något från den raka linjen så är inga observationer extrema nog för att vi ska förkasta AFT antagandet.

Vi har tagit fram en modell som ger en bra anpassning enligt kriterierna som har undersökts och går nu vidare med att undersöka om det kan finnas vissa vari-

abelvärden eller extrema observationer som modellen passar dåligt för. Detta görs genom att plotta residualer mot predikterad sjuktid, förklarande variabler och observationer. Eftersom residualer anger skillnaden mellan observerat och skattat värde kommer plottarna kunna visa trender för eventuella avvikelser.

De standardiserade residualerna ges av

$$r_{S_i} = \frac{logt_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \ldots - \hat{\alpha}_p x_{pi}}{\hat{\sigma}}$$

Från formeln ser vi att höga värden för r_{S_i} innebär att skattningen ger ett lägre värde än observerat utfall. Tolkningen av residualplottarna 19-26 nedan kompliceras av att det finns censurerade fall. Eftersom parameterskattningarna ger residualer för helt avslutade fall borde vi se att dessa är större än de observerade värdena, som innehåller oavslutade censurerade fall. Vi kan också förvänta oss att ökad andel censurerade observationer kommer ge intrycket av att överskattningar ökar.

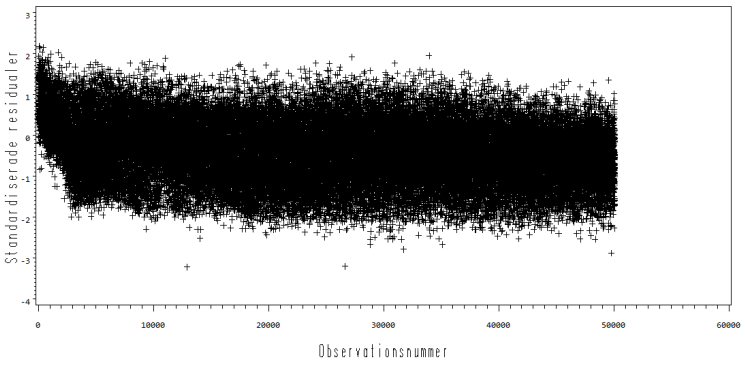

Figure 19: Standardiserade residualer plottade mot observationsnummer

Figur 19 visar standardiserade residualer plottade mot observationsnummer, som är sorterade efter insjuknandedatum, för både censurerade och ocensurerade fall. Vi ser att för de allra tidigaste datumen ligger de flesta residualerna över 0 och modellen underskattar då sjuktiden. För censurerade fall finns en svag trend till överskattning som växer med nyare sjukfall. Vi ser även två observationer som är något extrema, med residualvärden -3.31 och -3.27. Dessa individer var bara sjuka i 3 respektive 6 dagar innan de avvecklades och att exkludera dessa skulle inte ge någon effekt på parameterskattningarna. Övriga observationer är jämnt fördelade kring 0.

När residualerna plottas mot predikterad sjuktid i figur 20 ser vi en del långa predikterade sjuktider som sticker ut från resten, från cirka dag 4000 och framåt. För dessa antar de förklarande variablerna värden som ger långa sjuktider enligt vår modell. Exempelvis är karensen R för alla dessa observationer vilket ger en 5.6 gånger längre sjuktid än för fastställda karenser. För observationerna under

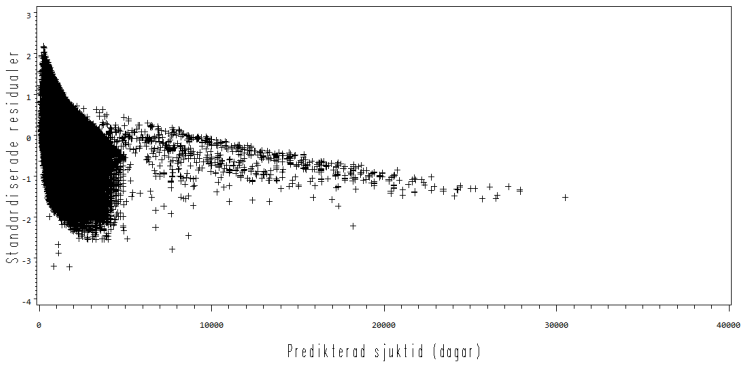

Figure 20: Standardiserade residualer plottade mot predikterad sjuktid

4000 sjukdagar syns en nedåtgående trend, korta sjuktider underskattas och långa överskattas. Orsaken till detta verkar vara censurering eftersom andelen censurerade data ökar stadigt med ökande sjuktid. Inga andra variabler visar någon sådan påverkan på prediktioner.

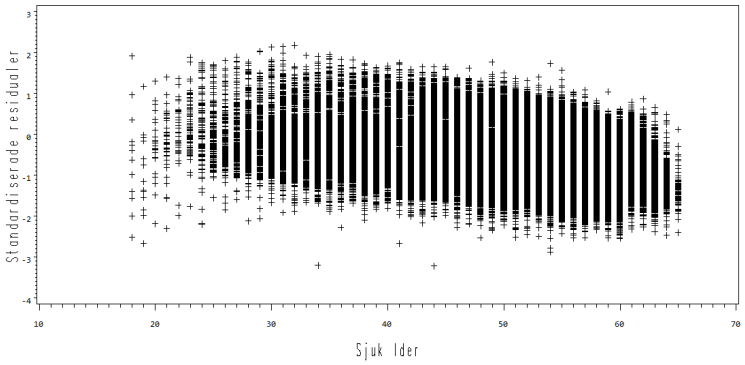

Figure 21: Standardiserade residualer plottade mot sjukålder

Figur 21 visar en trend till att överskatta sjuktiden för högre åldrar. Även andelen censurerade observationer ökar stadigt med ökande ålder, vilket återigen tyder på att censurering leder till överskattning.

Residualerna för sjuktillfällen i figur 22 ligger ganska jämnt fördelade kring 0.

Figur 23 visar att residualerna är jämnt fördelade kring 0 och har samma spridning för båda könen.

Kommungrupperna i figur 24 visar residualer som är jämnt fördelade kring 0

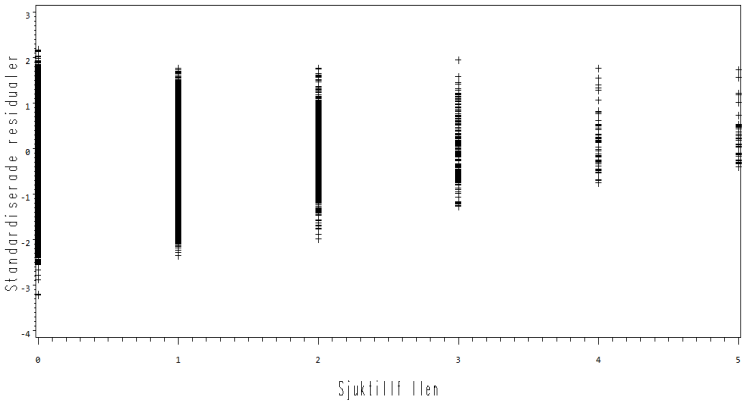

Figure 22: Standardiserade residualer plottade mot sjuktillfällen

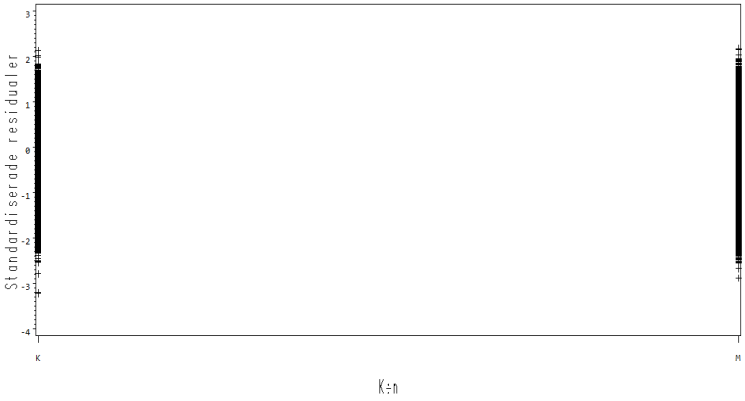

Figure 23: Standardiserade residualer plottade mot kön

utan några anmärkningsbara avvikelser.

Figur 25 visar att sjuktiden tenderar att överskattas för R-karens, som har en högre andel censurerade data på 56% jämfört med 44% för fastställd karenstid.

Vi ser i figur 26 att bolag 4 tenderar att ha överskattade sjuktider. Andelen censurerade observationer för detta bolag är 64%, nästan dubbelt så högt som för något av de andra bolagen.

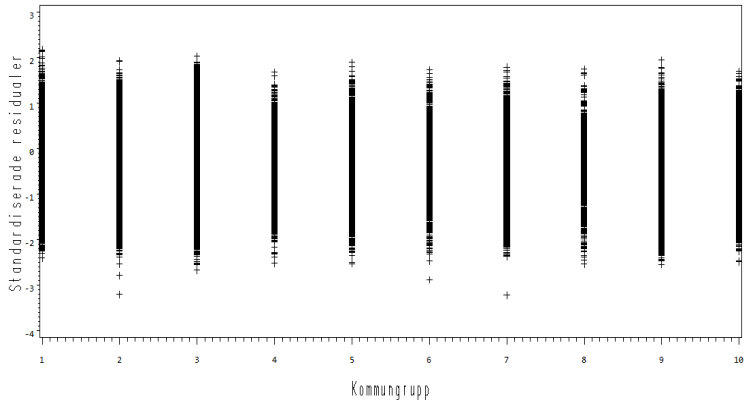

Figure 24: Standardiserade residualer plottade mot kommungrupp

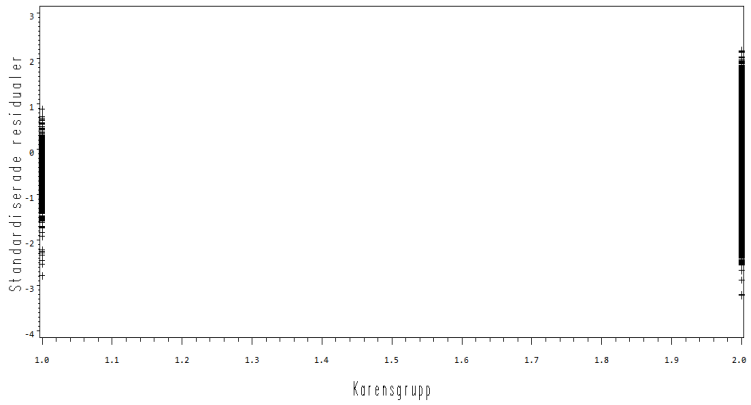

Figure 25: Standardiserade residualer plottade mot R-karens (1) och fastställd karens(2)

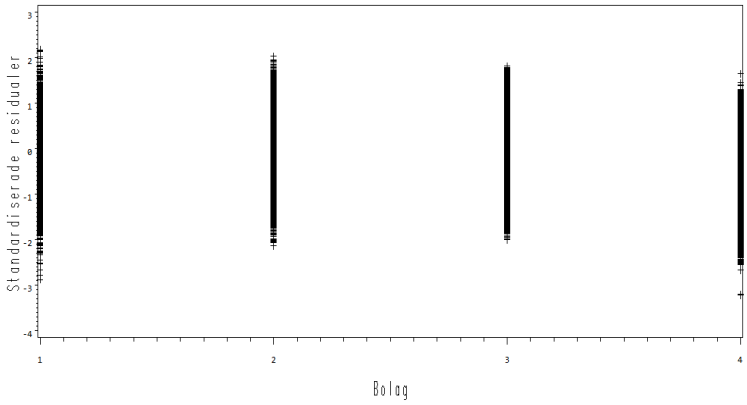

Figure 26: Standardiserade residualer plottade mot bolag

Prediktioner

Vi har sett att en del sjuktider överskattas i modellen och att dessa sammanfaller med en hög andel censurerade fall. Även andra möjliga orsaker till överskattning har undersökts, om den aktuella variabeln påverkas av någon av de andra, men det är endast censurering som visar något samband.

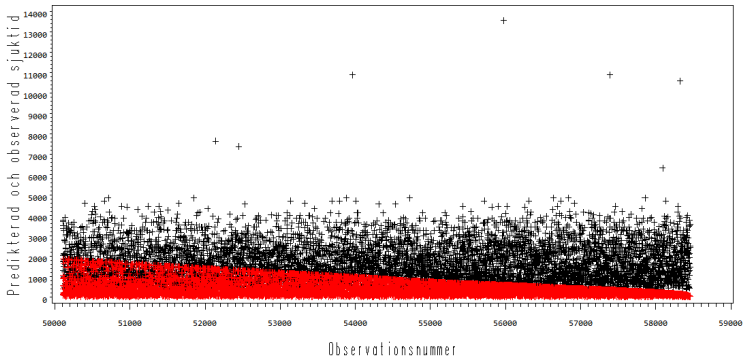

Figure 27: Predikterad median (svart) och observerat utfall (röd) för valideringsdataset

Modelldatasetets predikterade värden är också höga jämfört med valideringsdatasetets observerade utfall. Figur 27 visar de skattade 50:e percentilerna för sjuktiden (svart) mot de observerade utfallen (rött). Data är sorterat efter insjuknandedatum och vi ser en trend för nyare observerade sjukfall att vara kortare. Det ser ut som att vår modell överskattar den förväntade sjuktiden men jämförelsen av prediktion mot faktiskt utfall kompliceras av att det finns

censurerade fall. Av de nyare sjukfallen som fortfarande pågår är ju vissa inte ens nära sin totala sjuklängd och det mönster vi ser bland de faktiska (röda) sjuktiderna skulle alltså kunna innehålla fler långa sjukfall ut mot höger.

Valideringsdatasetet innehåller dock data från september 2008 och det borde finnas tillräckligt många gamla avslutade sjukfall för att det ska gå att urskilja en trend bland de äldsta. Vi vet att Försäkringskassan ändrade sjukskrivningsreglerna under 2008 som gjorde det svårare att bli klassad som sjuk, så kortare sjuklängder skulle inte vara helt överraskande. Om det är så att den faktiska totala sjuktiden har minskat över åren så fångas detta inte upp av modellen. Att inkludera variabeln sjukersättningsfaktor ger ingen märkbar skillnad.

Censurerade fall gör det alltså svårt att bedöma modellens prediktionsförmåga eftersom vi endast har kännedom om en del av den faktiska sjuklängden istället för den totala. Det går därmed inte att säga hur bra de predikterade sjuklängderna överrensstämmer med verkligheten, vilket man behöver ha i åtanke när modellen används för att prediktera förväntad total sjuklängd. Prediktion är ett område inom överlevnadsanalysen som inte är välutforskat och skulle vara intressant att undersöka vidare, men det ligger utanför ramen för denna uppsats.

5 Slutsats

Syftet med uppsatsen var att ta fram en modell för den förväntade sjuktiden och avvecklingsfunktionen. Efter att ha undersökt tillgängliga variabler från fyra bolag har vi hittat en parametrisk modell med 9 förklarande variabler som ger oss formler för avvecklingsfunktionen och predikterad sjuktid. En variabel av speciellt intresse var sjukersättningsfaktor, som vi hoppades skulle fånga upp extern påverkning på sjuktiden. Detta var inte fallet och variabeln är inte med i modellen.

I den slutliga modellen påverkas sjuktiden av individens ålder vid insjuknande, kön, bostadsort, karens, antal tidigare sjuktillfällen och vilket bolag försäkringen är tecknad hos.

Trots att den valda modellen passar bra enligt de statistiska mått som undersökts så är det svårt att avgöra hur bra prediktionsförmågan är. Förekomsten av censureringar försvårar bedömningen eftersom vi inte kan jämföra fullständiga sjuktider med predikterade värden. Slutsatsen är att modellen är tillräckligt bra för att användas i de två syften den är framtagen för, men man behöver det nämnda problemet i åtanke.

A Kommungruppsindelning

Sveriges Kommuner och Landstings 10 grupper kommer från [3] och består av:

1 - Storstäder (3 kommuner): Kommuner med en folkmängd som överstiger 200 000 invånare.
2 - Förortskommuner till storstäder (38 kommuner): Kommuner där mer än 50 procent av nattbefolkningen pendlar till arbetet i någon annan kommun. Det vanligaste utpendlingsmålet ska vara någon av storstäderna.
3 - Större städer (31 kommuner): Kommuner med 50 000-200 000 invånare samt en tätortsgrad överstigande 70 procent.
4 - Förortskommuner till större städer (22 kommuner): Kommuner där mer än 50 procent av nattbefolkningen pendlar till arbetet i en annan kommun. Det vanligaste utpendlingsmålet ska vara någon av de större städerna i grupp 3.
5 - Pendlingskommuner (51 kommuner): Kommuner där mer än 40 procent av nattbefolkningen pendlar till en annan kommun.
6 - Turism- och besöksnäringskommuner (20 kommuner): Kommuner där antalet gästnätter på hotell, vandrarhem och campingar överstiger 21 per invånare eller där antalet fritidshus överstiger 0,20 per invånare.
7 - Varuproducerande kommuner (54 kommuner): Kommun där 34 procent eller mer av nattbefolkningen mellan 16 och 64 år är sysselsatta inom tillverkning och utvinning, energi och miljö samt byggverksamhet
8 - Glesbygdskommuner (20 kommuner): Kommun med en tätortsgrad understigande 70 procent och mindre än åtta invånare per kvadratkilometer.
9 - Kommuner i tätbefolkad region (35 kommuner): Kommun med mer än 300 000 personer inom en radie på 112,5 kilometer.
10 - Kommuner i glesbefolkad region (16 kommuner): Kommun med mindre än 300 000 personer inom en radie på 112,5 km.

B Sjukersättningsfaktor

Variabeln sjukersättningsfaktor är beräknad som antal nybeviljade sjukersättningar dividerat med medelbefolkningen. Medelbefolkningen är beräknad enligt SCB:s metod som medelvärdet av ingående och utgående folkmängd för det aktuella året. Uppgiften om nybeviljade sjukersättningar är hämtad från [4] för åren 1977-1998. Från år 2003 och framåt finns uppgiften på Försäkringskassans hemsida [8]. Siffrorna mellan 1999 och 2002 saknas och är hämtade från källorna [5], [6] och [7].

Ersättningsår	Nybeviljad ersättning	Medelbefolkning	Sjukersättningsfaktor
1977	46350	8251648	.0056
1978	45144	8275777	.0055
1979	44278	8293724	.0053
1980	45289	8310474	.0054
1981	43615	8320485	.0052
1982	42286	8325259	.0051
1983	43338	8329029	.0052
1984	46792	8336597	.0056
1985	51009	8350380	.0061
1986	50106	8369827	.0060
1987	51691	8397799	.0062
1988	54135	8436486	.0064
1989	51991	8492962	.0061
1990	50493	8558833	.0059
1991	49554	8617375	.0058
1992	58382	8668066	.0067
1993	62465	8718561	.0072
1994	48531	8780745	.0055
1995	39204	8826939	.0044
1996	39245	8840998	.0044
1997	41198	8846062	.0047
1998	34487	8850974	.0039
1999	39506	8857874	.0045
2000	45093	8872109	.0051
2001	57000	8895960	.0064
2002	63738	8924958	.0071
2003	73161	8958229	.0082
2004	60308	8993531	.0067
2005	55581	9029572	.0062
2006	48176	9080505	.0053
2007	47683	9148092	.0052
2008	35864	9219637	.0039
2009	22525	9298515	.0024
2010	14121	9378126	.0015
2011	14369	9449213	.0015
2012	16720	9519374	.0018

C Specifika fördelningar för sjuktiden T

PH Weibull
Om sjuktiden är Weibullfördelad så ges baselinehazarden av

$$h_0 = \lambda\gamma t^{\gamma-1}$$

λ, kallas skalparametern och γ formparametern, $\lambda, \gamma > 0$. Den i:te individen, med uppsättningen $\eta_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$ av förklarande variabler, har hazardfunktionen

$$h_i(t) = exp(\eta_i)h_0 = exp(\eta_i)\lambda\gamma t^{\gamma-1}$$

Detta är en Weibullfördelning med skalparameter $\lambda exp(\eta_i)$ och formparameter γ.

Överlevnadsfunktionen är

$$S_i(t) = exp(-exp(\eta_i)\lambda t^{\gamma}) \tag{7}$$

PH exponential
Exponentialfördelningen är ett specialfall av Weibull, då γ=1. Baselinehazarden

$$h_0(t) = \lambda$$

är konstant och hazardfunktionen för individ i är

$$h_i(t) = \lambda exp(\eta_i)$$

Överlevnadsfunktionen ges av

$$S(t) = exp(-\lambda(t))$$

Nedan följer en sammanställning av de tillgängliga AFT fördelningarna i SAS för sjuktiden T med tillhörande överlevnadsfunktioner och percentiler, som beräknas utifrån formel (5) på sidan 11. Hazardfunktionen fås från formel (1), sida 8.

AFT Weibull
Om T_i är Weibullfördelad så har ϵ_i en extremvärdesfördelning (Gumbel).

$$S_{\epsilon_i}(\epsilon) = exp(-exp(\epsilon))$$

$$S_i(t) = exp(-exp(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}))$$

$$t_i(p) = S_i^{-1}(\frac{100-p}{100}) = exp(\sigma log(-log\frac{100-p}{100}) + \mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi})$$

$$h_i(t) = -\frac{d}{dt}logS(t) = \frac{1}{\sigma t}exp(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma})$$

Samband mellan AFT Weibull och PH Weibull
Överlevnadsfunktionen för PH Weibull från formel (7) var

$$S_i(t) = exp(-exp(\beta_1 x_{1i} + \ldots + \beta_p x_{pi})\lambda t^{\gamma})$$

Överlevnadsfunktionen för AFT Weibull från ovan kan även skrivas om som

$$S_i(t) = exp(-exp(\frac{-\mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma})t^{1/\sigma})$$

En jämförelse av dessa två ekvationer ger att parametrarna λ, γ och β_j i PH Weibull modellen kan uttryckas i termer av μ, σ och α_i i AFT Weibull modellen.

$$\lambda = exp(-\mu/\sigma), \gamma = 1/\sigma, \beta_j = -\alpha_i/\sigma$$

AFT Log-logistic

$$S_{\epsilon_i}(\epsilon) = \frac{1}{1+e^{\epsilon}}$$
$$S_i(t) = (1 + exp(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma}))^{-1}$$
$$t_i(p) = exp(\sigma log(\frac{p}{100-p}) + \mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi})$$
$$h_i(t) = \frac{1}{\sigma t}(1 + exp(-(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma})))^{-1}$$

AFT Log-normal
Om T_i är lognormalfördelade så är $logT_i$ normalfördelade. ϵ_i antas därför komma från en standardiserad normalfördelning med fördelningsfunktion $\Phi(\epsilon)$.

$$S_{\epsilon_i}(\epsilon) = 1 - \Phi(\epsilon)$$
$$S_i(t) = 1 - \Phi(\frac{logt - \mu - \alpha_1 x_{1i} - \ldots - \alpha_p x_{pi}}{\sigma})$$
$$t_i(p) = exp(\sigma \Phi^{-1}(\frac{p}{100}) + \mu + \alpha_1 x_{1i} + \ldots + \alpha_p x_{pi})$$

$S_i(t)$ uttrycks i termer av fördelningsfunktionen $\Phi(\epsilon)$, så $h_i(t)$ ges av formel (4), där

$$h_{\epsilon i}(\epsilon) = \frac{f_{\epsilon i}(\epsilon)}{S_{\epsilon i}(\epsilon)}$$

AFT Gamma
Det finns en Gamma modell med två parametrar och den generaliserade Gamma modellen med tre parametrar. Modellen vi anpassar i SAS är den generaliserade med parametrarna λ, ρ och θ. Täthetsfunktionen är

$$f(t) = \frac{\theta \lambda^{\rho\theta} t^{\rho\theta - 1} exp(-(\lambda t)^{\theta})}{\Gamma(\rho)}$$

$S(t)$ uttrycks som

$$S(t) = 1 - \Gamma_{(\lambda t)^{\theta}}(\rho) = 1 - \frac{1}{\Gamma(\rho)} \int_0^{(\lambda t)^{\theta}} u^{\rho - 1} e^{-u} du$$

References

[1] David Collett, *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, 2nd Edition, 2003.

[2] Paul D. Allison, *Survival Analysis Using SAS*. SAS Publishing, 2nd Edition, 2010.

[3] Sveriges kommuner och landsting (SKL), *Kommungruppsindelning 2011*. http://www.skl.se/kommuner_och_landsting/fakta_om_kommuner/kommungruppsindelning

[4] Riksförsäkringsverket, *Sveriges officiella statistik*. Periodisk publikation, LiberFörlag/Allmänna förlaget, 1977-1998. ISSN 0082-0075

[5] http://www.regeringen.se/content/1/c4/09/16/c96a4dd7.pdf

[6] Riksförsäkringsverket med flera, *Arbetslivsfakta* http://www.av.se/dokument/statistik/alf/alf2002_3.pdf

[7] Regeringen, *Avstämning av regeringens mål för minskad ohälsa* http://www.regeringen.se/content/1/c4/29/88/1c170651.pdf

[8] Försäkringskassan, *Nybeviljad sjuk- och aktivitetsersättning år 2003 och framåt*. http://www.forsakringskassan.se/statistik/sjuk/sjukersattningaktivitetsersattning/sjukochaktivitetsersattning/