Sammanfattning

Denna uppsats använder sig av SCB:s registerdata som omfattar samtliga par som gifte sig för första gången under 1998, dessa par studeras under cirka elva år fram till den 31 december 2008. Genom att använda metoder inom överlevnadsanalys beskrivs hur faktorer som individernas ursprung, utbildningsnivå, ålder samt åldersskillnad mellan individerna i paret påverkar risken för skilsmässa. Särskilt fokus riktas på kovariaten som behandlar individernas ursprung. I undersökningen jämförs och utvärderas vanligt förekommande parametriska modeller, Cox regressionsmodell och en piecewise exponential modell i syfte att ta fram en lämplig modell. Ingen av de undersökta parametriska modellerna lämpar sig perfekt för materialet, men log-normal och generaliserade gamma har bäst passform. Cox regressionsmodell modifieras för att uppfylla antagandet om proportionalitet. På det hela taget visas att Cox regressionsmodell utan hänsyn till proportionalitetsantagandet, de mest lämpliga parametriska modellerna, piecewise exponential modell samt den modifierade Cox regressionsmodellen ger mycket liknande skattningar. Samtliga nämnda faktorer påverkar risken för skilsmässa. Sett till ursprung har par där båda individerna är födda utrikes en liknande risknivå som par där individerna samt deras föräldrar är födda i Sverige. En förhöjd risk påvisas då individerna är andra generationens invandrare eller då individerna i paret har olika ursprung.

1. Inledning

Under våren 2010 publicerade Statistiska centralbyrån (SCB) rapporten *Födda i Sverige – ändå olika? Betydelsen av föräldrarnas födelseland*. I rapporten behandlas familjebildande, barnafödande, inrikes- och utrikes flyttningar samt dödlighet efter individers bakgrund. Befolkningen är uppdelad i fyra kategorier beroende på vilket ursprung individerna har. Något som inte behandlades i rapporten är i vilken omfattning personer med olika bakgrund tenderar att skilja sig.

I Sverige började SCB registrera skilsmässor år 1831, detta år registrerades 95 stycken. Med tiden har antalet skilsmässor ökat väsentligt. Ur ett internationellt perspektiv har Sverige, med en utav de mest liberala äktenskapslagstiftningarna, en relativt hög andel skilsmässor (http://unstats.un.org och SCB demografiska rapporter 1995:1). Det är därför av intresse att studera hur personer med utländsk bakgrund anpassas till denna ”svenska” företeelse. Är det möjligt att se några skillnader mellan olika personer givet deras ursprung?

Datamaterialet som ligger till grund för denna studie omfattar alla par som gifte sig för första gången under 1998. Genom registerdata har paren följts fram till den 31 december 2008. Metoden för studien är överlevnadsanalys och det primära syftet är att se om det finns några skillnader mellan par som beror på var individerna i paret och deras föräldrar är födda. För att besvara detta används kontrollvariabler som avser individernas utbildning och ålder. Individernas ålder används dels för att se till effekten av åldern vid giftermålet, dels för att kontrollera eventuella effekter av ålderskillnader mellan personerna som ingår giftermålet.

Ursprung delas in i sex olika kategorier för att tillgodose olika konstellationer beroende på var individerna samt deras föräldrar är födda. Till skillnad mot andra liknande studier, vilka främst är amerikanska, tar denna studie hänsyn till personer som är födda i Sverige men är andra generationens invandrare. Det är rimligt att anta att individer som har föräldrar födda i Sverige integrerats mer i det svenska samhället. Genom vår kategorisering kan vi tillgodose hur nära band individerna har med ett annat land och kultur.

1.1 Problemformulering

SCB påvisade i sin rapport: *Födda i Sverige - ändå olika? Betydelsen av föräldrarnas födelseland* (2010) att det demografiska beteendet är annorlunda för dem födda i Sverige som har utrikesfödda föräldrar jämfört mot de som är födda i Sverige med föräldrar födda i Sverige. De områden som undersöktes var utbildning, arbetsmarknad, familjebildning, barnafödande, utrikes och inrikes flyttningar samt dödlighet. Samtliga av dessa områden visade på mer eller mindre skillnader.

Det skulle vara av särskilt intresse att studera om det finns några betydande skillnader vad avser skilsmässa, något som sällan undersökts i Sverige. För att utreda detta behövs en lämplig statistisk modell. SCB använder i stor utsträckning Cox regressionsmodell för demografiska analyser. Det skulle vara intressant att testa om vi kan finna en modell som är bättre lämpad.

Förutom individers ursprung kan det vara intressant att utröna huruvida andra faktorer kan förväntas influera risken för skilsmässa. De frågeställningar som uppsatsen ska försöka att besvara är följande:

- Är någon parametrisk modell, Cox regressionsmodell eller någon annan modell lämplig för datamaterialet?
- Hur påverkar ålder, utbildning och ålderskillnad mellan individerna risken för skilsmässa?
- Har individernas samt deras föräldrars härkomst någon effekt på risken för skilsmässa? Går det att uttyda några skillnader beroende på om en eller båda individerna i paret är födda i Sverige, är födda utrikes eller är andra generationens invandrare?

1.2 Syfte

Syftet med denna uppsats är att undersöka om det finns några skillnader vad avser risk för skilsmässa mellan olika par där individerna i paret är andra generationens invandrare, födda utrikes eller födda i Sverige med föräldrar födda i Sverige. Även hur utbildningsnivå och individernas ålder inverkar på risken för skilsmässa. Som ett delsyfte ämnar uppsatsen att identifiera en passande och tillförlitlig modell inom överlevnadsanalys.

1.3 Avgränsningar

I denna uppsats studeras alla par i Sverige som gift sig för första gången under 1998. Dessa par följs till och med 31 december 2008. Slutsatser kan således endast göras beträffande nämnda tidsperiod. Personer som gift sig utomlands eller genomgått tidigare giftermål kommer inte att beaktas.

Skilsmässodatumet mäts från den dag som domen vinner laga kraft i domstol. Detta innebär att betänketider som i vissa falls krävs inte tas hänsyn till. Även de individer i ett äktenskap som är bosatta på olika fastigheter eller i olika länder kommer inte beaktas som skilda.

De par där antingen båda individerna flyttat utomlands, eller en individ avlidit under den studerade tiden kommer att behandlas som censurerade. Dessa observationer antas inte vara informativa om risken för skilsmässa.

Denna uppsats ämnar svara på om det går att påvisa skillnader av risken för skilsmässa. Däremot behandlas inte varför dessa eventuella skillnader uppstår. I avsnittet *tidigare forskning* ges förslag på teorier till orsaker för tidigare studiers utfall, dessa utvecklas dock inte.

2. Tidigare studier

I detta kapitel redogörs för liknande studier som tidigare utförts i Sverige och andra länder. Statistiska metoder, resultat och variabler för tidigare forskning presenteras.

Svenska kvantitativa studier som jämför skilsmässorisken bland individer med olika härkomst är relativt sparsamma. 1995 släppte SCB rapporten *skilsmässor och separationer: bakgrund och utveckling*, vilken kort behandlar invandrares separationsrisk under perioden 1991-1992. Grupperna som jämförs delas in efter den världsdel där mannen är född (Sverige och Norden behandlas dock separat). Utan att kontrollera andra faktorer än mannens härkomst bedöms alla grupper ha mellan cirka 30 (Europa utanför Norden) – 200 (Sydamerika) procents ökad risk jämfört med då mannens födelseland var Sverige. Analysen bygger på Cox-regression, där censurering till följd av exempelvis emmigration eller dödsfall inte har beaktats.

Mehrdad Darvishpour har i ett flertal rapporter (bl.a. 2004) studerat skilsmässor bland olika grupper av invandrare i Sverige. För studieåret 1991 fastslår han att skilsmässorisken för chilenska och iranska barnfamiljer var högre än för andra invandrargrupper, detta när risken endast beaktas med avseende på nationalitet. Risken minskar när hänsyn tas till fler riskfaktorer (ålder, vistelsetid i Sverige, äktenskapets varaktighet, barnantal, barn före giftermålet, mannens respektive kvinnans utbildning, socioekonomisk grupp, inkomst, boendesituation och region).

Darvishpour använder logistisk regression och studerar vissa invandrargrupper där båda i paret är födda utomlands med föräldrar födda utomlands. Rapporten fokuserar inte på statistiska metoder utan ämnar snarare förklara varför skillnaderna uppstår ur ett klass-, etnicitets- och genusperspektiv. Slutsatsen är att främst tre faktorer kan förklara varför vissa invandrargrupper har högre skilsmässofrekvens än svenskar. 1) Sämre socioekonomisk situation, 2) psykoligiska och kulturella påfrestningar, 3) konflikter mellan maken och makan intensifieras efter invandringen.

Utanför Sverige har flertalet studier genomförts. Lejonparten av dessa är amerikanska och fokuserar på skilsmässorisken bland olika etniska grupper (ofta svarta, vita och latinamerikaner). Dessa tar även ofta hänsyn till olika konstellationer i paret till exempel giftermål mellan vit man och latinamerikansk kvinna.

I artikeln *Marital Dissolution Among Interracial Couples* från 2009 (Zhang och Hook) undersöks hur stabila giftermål mellan olika etniska grupper i USA är. Studien grundar sig på ett stratifierat urval av drygt 23 000 par som intervjuats under sex studieperioder mellan 1990-2001. Varje studieperiod inkluderar upp till åtta intervjuer med paren. En Cox proportional hasard modell användes, och rättfärdigas med att den anses som robust samt inget antagande om baseline överlevnadsfunktionen behövs. Variabeln av särskilt intresse var den kategoriska variabeln etnicitet (delas in i sju kategorier) för parterna i paret. Fyra varianter av Cox modellen användes, där olika många oberoende variabler inkluderats. Modellen med flest variabler använde förutom etnicitet: antal barn, tidsperiod för giftermålet, region, kvinnans ålder vid

giftermålet, ålderskillnad mellan make och maka, utbildning, logaritmen av inkomst och var individerna var födda. Ingen av dessa var tidsberoende och alla variabler förutom barn och inkomst var kategoriska. Slutsatsen i studien är att separationer är starkt associerat med etnicitet, men att olika etnicitet i ett par inte nödvändigtvis innebär ökad risk för separation.

En annan amerikansk studie har utförts av Sweeney och Phillips (2004), som undersökt trender av skilsmässor bland svarta och vita kvinnor. Datamaterial var baserat på intervjuer av kvinnor som någon gång varit gifta, totalt cirka 40 000. För analysen användes en logistisk regressionsmodell med diskret data, vilket innebär att tiden delas in i perioder. Författarna finner bevis för att utvecklingen i skilsmässor skiljer sig mellan svarta och vita kvinnor. De vita kvinnorna skilde sig i en ökande omfattning till mitten av 1970-talet, då det stabiliserades. Medan trenden för de svarta kvinnorna var något tilltagande från slutet av 1980-talet.

Ingen av de påträffade studierna undersöker skillnader med hänsyn till om personerna är andra generationens invandrare.

3. Metod

Kapitel 3 inleder med att allmänt beskriva överlevnadsanalys. Olika metoder bland annat parametriska och Cox regressionsmodell förklaras. Deras antaganden och vad man kan göra om dessa inte är uppfyllda beskrivs.

3.1 Överlevnadsanalys

Överlevnadsanalys lämpar sig väl för att modellera ett väldefinierat tidsförlopp. Det studerade tidsförloppet sträcker sig mellan en startpunkt då observationer börjar studeras tills en slutpunkt då en händelse av intresse inträffar.

En händelse kan vara någon angiven upplevelse av intresse som möjligen kan hända en individ. Det är även möjligt att studera fler än en händelse i samma analys. Tidsförloppet kan vara förlupen tid i dagar, veckor, månader eller år. Valet av tidsmått bör grunda sig på det som gör analysen mest relevant och tydlig med avseende på ämnet man undersöker.

Ett typexempel av överlevnadsanalys inom medicin är en studie där ett antal individer observeras mellan två tidpunkter från studiens början till slut, till exempel för att utvärdera två olika behandlingssätt. Slutpunkten för observation av en individ inträffar när denna avlider. I exemplet ses dödsfall som händelsen av intresse och tiden till detta inträffar, vilket benämns som överlevnadstiden, den beroende variabeln.

Ovan exempel ger en innebörd åt namnet överlevnadsanalys, men händelsen av intresse behöver inte karaktäriseras av ett dödsfall utan kan likväl vara någon annan distinkt händelse som inträffar. Denna förändring kan betecknas som en kvalitativ förändring eftersom det är möjligt att fastställa en tidpunkt för händelsen som är en övergång från ett diskret tillstånd till ett annat. (Collett 2003)

Överlevnadsanalys är således mycket tillämpbart när man studerar olika typer av händelser inom samhälls- och naturvetenskapliga områden. Begreppet överlevnadsanalys används i störst utsträckning för metoden men har dock olika begreppsdefinitioner beroende på ämnesområde och benämns exempelvis livsförloppsanalys inom sociologi, ”duration analysis” inom ekonomi och ”failure time analysis” inom ingenjörskonst(Allison 2010). Det råder dock inte några betydande skillnader i tillämpningen av metoden fastän olika ämnesområden kan använda något annorlunda ansatser. Fortsättningsvis används begreppet överlevnadsanalys i denna studie.

För att kunna applicera en överlevnadsanalys i studien om skilsmässor är det inte tillräckligt att ta reda på vilka som ingått i äktenskap vid studiens begynnelse utan vi måste framförallt få vetskap om när ändringen sker, det vill säga skilsmässa som händelse. Det bör följaktligen fastställas en tid för händelsen och det är angeläget att tydligt redogöra för denna händelse av intresse.

Jämfört med andra metoder exempelvis vanliga regressionsmodeller kan modeller inom överlevnadsanalys hantera censurerade observationer det vill säga observationer för vilka händelsen av intresse inte inträffade för under den studerade perioden.

3.2 Censurering

Det förekommer olika former av censurering men vi avgränsar oss till att enbart beskriva den typ av censurering som förekommer i vår studie, det vill säga högercensurering.

Det är lätt att bilda sig en uppfattning om att en analys av överlevnadstid för ett urval kan hanteras med hjälp av välutvecklade statistiska metoder för att analysera kontinuerlig eller diskret data. Orsaken till att det inte är optimalt att använda sig av vanligt förekommande statistiska metoder såsom linjära regressions modeller när man analyserar överlevnadsdata är att man i själva verket måste vänta till händelsen inträffar (Aalen m.fl. 2008).

Till studien avslutas och från att den påbörjas kommer man vanligtvis observera att händelsen av intresse inte inträffar för en del av individerna i studien. I vår studie om skilsmässor kommer inte alla par att skilja sig under tiden studien varar och vi kommer därmed inte få vetskap om när och hur många som skiljer sig efter det att studien avslutas.

I metoder inom överlevnadsanalys behöver inte händelsen av intresse inträffa för samtliga observationer. Informationen från dessa censurerade observationer kan tas till vara på för att bland annat beräkna estimat.

Datamaterialet vi kommer att behandla består således av både ”fullbordade” samt ”ofullbordade” observationer. De ofullbordade observationerna i vår studie benämns som högercensurerade observationer.

3.3 Överlevnadsfunktionen och hasardfunktionen

För att analysera censurerad data behövs två anpassade funktioner, dessa är överlevnads- och hasardfunktionen. Genom att funktionerna estimeras, givet observerade överlevnadstider, kan fördelningen av överlevnadsdata beskrivas.

Tiden till en händelse (överlevnadstiden) för en individ är ett icke-negativt värde på den slumpmässiga variabeln T. Fördelningsfunktionen för *T* ges av:

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

där *f(u)* är den underliggande täthetsfunktionen för *T*. Funktionen representerar sannolikheten att överlevnadstiden är mindre eller lika med ett visst värde *t*.

Överlevnadsfunktionen tolkas däremot som sannolikheten att överlevnadstiden är större än *t*, eller annorlunda uttryckt sannolikheten att en händelse av intresse inte inträffat från startpunkten (*t*=0) till tidpunkten *t*, och ges av:

S(t) = P(T>t) = 1 – F(t), där *S(t)=1* då *t=0* och *S(t)=0* då *t*=∞.

Hasardfunktionen används för att beskriva risken eller hasarden för att händelsen inträffar momentant, vid *t*, för en individ som ännu inte upplevt händelsen. Hasardfunktionen definieras som gränsvärdet när δt går mot noll på följande sätt: $h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}$, och tolkas som sannolikheten att *T* ligger mellan tidpunkterna *t* och *t*+ *δt* givet att *T* är större eller lika med *t*. Eftersom sannolikheten divideras med ett tidsintervall, *δt*, uttrycks hasarden inte som en sannolikhet utan som en intensitet. För att få fram sannolikheten för att händelsen inträffar vid *t* multipliceras således *δt* med *h(t)*.

Man kan visa att *h(t)=f(t)/S(t)* gäller. Sambandet mellan hasardfunktionen, överlevnadsfunktionen och täthetsfunktionen innebär att det är möjligt att specificera alla funktioner givet att någon av dem är känd. (Kleinbaum och Klein 2005)

En annan användbar funktion är den kumulativa hasardfunktionen som ges av:

$$H(t) = \int_0^t h(u)\mathrm{d}u$$

och kopplas ihop med överlevnadsfunktionen på följande sätt: $H(t) = -logS(t)$ (Collett 2003). Funktionens användningsområde beskrivs i avsnitt 3.7.5.

3.3.1 Olika sätt att estimera överlevnadsfunktionen

Life-table estimatorn är en metod som är användbar när den exakta tiden för händelsen av intresse inte är känd. Den undersökta perioden delas in i relativt få tidsintervall. Antal intervall och deras längd beror på antalet observationer. För censurerad data ges life-table estimatet för överlevnadsfunktionen av $S(t) = \prod_{j=1}^{k}\left(\frac{n'_j - d_j}{n'_j}\right)$, där $n'_j = n_j - c_j/2$ och representerar det

genomsnittliga antalet individer, under intervall *j*, som riskerar att händelsen inträffar. d_j representerar antalet individer för vilka händelsen inträffar för under intervallet. Enligt formeln görs antagandet att de censurerade överlevnadstider, c_j, sker likformigt under intervallet. *S(t)* ger således den estimerade sannolikheten att händelsen för en individ inte inträffar till och med starten av det *k*:te tidsintervallet. (Collett 2003)

Kaplan-Meier estimatorn delar, liksom life-table estimatorn, in datamaterialet i intervall. Skillnaden mellan estimatorerna är att Kaplan-Meier delar in materialet i betydligt fler intervall, då intervallen bestäms av individernas överlevnadstid. Starten för varje intervall representerar att händelsen av intresse inträffar för en individ. Dock behöver inte antalet intervall vara lika stort som antalet observationer, då somliga observationer är högercensurerade och på grund av att händelsen kan inträffa samtidigt för flera personer.

Kaplan Meier estimatet för överlevnadsfunktionen ges av: $S(t) = \prod_{j=1}^{k}\left(\frac{n_j - d_j}{n_j}\right)$, där n_j

representerar antalet individer för vilka händelsen inte inträffat för precis innan tidpunkt *j*. d_j visar antalet individer för vilka händelsen inträffar för vid *j* (Collett 2003). En graf över Kaplan-Meier estimaten ger en steg-funktion där den estimerade sannolikheten är konstant mellan två intervall och minskar för varje intervall.

3.3.2 Olika sätt att estimera hasardfunktionen

Life-table estimatorn för hasardfunktionen estimerar den genomsnittliga risken för att en händelse inträffar per tidsenhet för varje intervall. $h(t) = \frac{d_j}{(n'_j - d_j/2)\tau_j}$. Täljaren i funktionen

representerar antalet inträffade händelser under tidsintervallet och nämnaren den genomsnittliga överlevnadstiden i intervallet. Längden på intervallet ges av τ_j. I funktionen antas att frekvensen av händelser är konstant under intervallet.

Kaplan Meier estimatorn för hasardfunktionen ges av: $h(t) = \frac{d_j}{n_j \tau_j}$ och ger kvoten av antalet

händelser för en given tidpunkt och antalet individer som riskerar att utsättas för händelsen. τ_j anger tidsintervallet mellan två händelser, enligt $\tau_j = t_{(j+1)} - t_j$. För att få fram plottar med tydliga mönster av funktionen behöver den oftast först jämnas ut då den tenderar att bli irreguljär (Collett 2003). I SAS finns en metod som heter *Epanechnikov Kernel-Smoothed hazard function* som kan användas för att uppnå detta (Allison 2010).

3.3.3 Test av överlevnadsfunktionen för olika grupper

Det är ofta av intresse att utvärdera om olika grupper har samma överlevnadsfunktion. För att statistiskt testa detta antagande kan log-rank test användas. Testet ger en genomsnittlig jämförelse av gruppernas överlevnadsfunktioner estimerade med Kaplan-Meier. Nollhypotesen är att grupperna har samma överlevnadsfunktion. Om nollhypotesen förkastas har vi således stöd för att grupperna skiljer sig vad avser täthetsfunktionerna för tiden till en händelse.

För att illustrera hur testet beräknas, ges ett exempel med två grupper. Statistikan för log-rank skrivs: $U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j})$, där skillnaden mellan faktiskt antal händelser och förväntade antal

händelser för grupp ett summeras över antalet distinkta tider för när händelserna äger rum (*r* representerar det antal händelser som inträffar för båda grupperna). Det förväntade antalet händelser i grupp ett vid tidpunkt j beräknas genom att multiplicera antalet individer i gruppen som riskerar att utsättas för händelsen med antalet händelser (för båda grupperna) vid tidpunkt *j*, detta värde divideras med det totala antalet individer som riskerar händelsen[1]. Om grupperna är lika blir således U_L =0. För att få fram ett chi-två värde (med en frihetsgrad) divideras den kvadrerade statistikan med dess varians. (Collett 2003)

Log-rank testet kan även användas för fler än två grupper (för detaljer se Collett 2003); under nollhypotesen att alla överlevnadsfunktioner är samma och alternativhypotesen att minst en överlevnadsfunktion skiljer sig från de övriga.

I SAS redovisas även ett alternativt test som kallas Wilcoxon test. Den fundamentala skillnaden mellan log-rank och Wilcoxon test är att det senare är viktat så att händelser som inträffar tidigt får större betydelse. Logiken bakom viktningen är att senare tidpunkter ska få mindre betydelse då färre individer riskerar att utsättas för händelsen (på grund av att individer redan utsatts eller censurerats).

3.4 Test för kovariaternas influens och utvärdering av modellen

Log-rank och Wilcoxon testet som tidigare visades för att jämföra överlevnadsfunktionen för olika grupper kan även användas för att utvärdera hur olika kovariater influerar överlevnadstiden. Nollhypotesen i testet för kovariaterna är att koefficienterna är lika med noll, vilket innebär att kovariaterna inte påverkar överlevnadstiden. För att testa effekten av kovariaterna kan ett univariat test genomföras, där varje kovariat testas oberoende av de andra.

För att ta hänsyn till de andra kovariaterna kan *forward selection* användas, vilket innebär att en kovariat i taget inkluderas i modellen. Den kovariat som får högst chi-två värde i det univariata testet inkluderas först i modellen. I fallande värde inkluderas sedan resterande kovariater vilket resulterar i nya modeller där varje modell har en ”ny” kovariat. När en ny kovariat sätts in i modellen kontrolleras effekten för de som redan finns med. Genom att jämföra chi-två värdena från det univariata testet och *forward selection* kan slutsatser göras om kovariaterna är högt

[1]Formeln skrivs: $e_{1j} = n_{1j}\, d_j / n_j$

korrelerade och därmed möjligtvis överflödiga. Det ska dock observeras att testet behandlar variablerna som kontinuerliga.

För Cox proportional hasard och de parametriska modellerna redovisas ett betaestimat med tillhörande standardfel för varje kovariat. I SAS-utskriften finns även ett test för om betaestimaten är skilda från noll, detta i form av ett chi-två värde med tillhörande p-värde. Testet kallas för Wald test och beräknas genom att kvadrera kvoten av betaestimatet genom dess standardfel. Testet tar hänsyn till de andra kovariaterna i modellen, det vill säga kovariaterna testas inte isolerade av varandra. (Allison 2010)

Kovariaternas effekt på överlevnadstiden kan även utvärderas med likelihoodfunktionen och Akaike´s informations kriterium. Dessa ger en samlad bild över hur bra hela modellen förklarar den beroende variabeln, och värdena kan jämföras för olika modeller. Likelihoodfunktionen beskrivs mer i avsnitt 3.5.2. Akaike´s informations kriterium (AIC) är en modifikation av likelihoodfunktionen. För att jämföra detta värde behöver inte modellerna vara specialfall av varandra. Statistikan ges av: $AIC = -2logL + \alpha q$ (Collett 2003), där *logL* är den maximerade log-likelihoodfunktionen för modellen, *q* antal estimerade koefficienter och α en förutbestämd konstant. Modellen utvärderas genom AIC-värdet, där ett mindre värde indikerar en bättre modell. Observera att AIC straffar modeller där onödiga kovariater lagts till.

Liksom för vanliga linjära regressionsmodeller kan förklaringsgraden, R^2, beräknas för överlevnadsmodeller. I analys av överlevnadsdata är detta dock inte standard och rapporteras inte av SAS. En anledning till detta kan vara att det finns många sätt att beräkna förklarad varians, men ingen anses vara mer tillförlitlig än någon annan i det generella fallet (Collett 2003). Magee (1990) beskriver dock en möjlighet som lämpar sig väl för modeller som inte är linjära. Utgångspunkten är chi-två värdet för likelihoodratio statistikan, *LR* (se senare avsnitt 3.5.2.), och antal observationer, *n*, då $R^2 = 1 - exp(-LR/n)$.

3.5 Parametriska modeller

För en parametrisk överlevnadsmodell antas överlevnadstiden följa en känd fördelning. Täthetsfunktionen för tiden kan skrivas med okända parametrar och därigenom kan överlevnads- och hasardfunktionen bestämmas (se avsnitt 3.3).

Flera av de parametriska regressionsmodellerna tillhör gruppen accelerated failure time (AFT) modeller. För dessa kan överlevnadstiden skrivas som en funktion av förklarande variabler. Ett antagande för dessa modeller är att kovariaterna påverkar överlevnadstiden multiplikativt och kan sägas accelerera överlevnadstiden (för proportionella hasardmodeller är effekten multiplikativ med hasarden). Antagandet kan illustreras med ett exempel där överlevnadstiden för två grupper av individer ska jämföras enligt: $S_2(t) = S_1(\gamma t)$, γ representerar den konstanta accelererande faktorn, vilket är en kvot av överlevnadstiderna och beskriver skillnaden mellan grupperna.

En generell formel för en accelerated failure time modell med tiden för händelsen *T* för individ *i* kan skrivas som: $\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi} + \sigma\varepsilon$; där μ är interceptet α_i avser okända koefficienter till *p* förklarande variabler. Den slumpmässiga feltermen, ϵ, multipliceras

med en skalparameter, σ, vilken fångar upp variansen i ϵ och möjliggör på så vis att ϵ kan ha en fix varians. Modellen påminner om en vanlig linjär regressionsmodell, men oftast görs inget antagande om att feltermen är normalfördelad med noll i väntevärde. Den beroende variabeln är även logaritmerad för att säkerställa att predikterade värden av *T* är positiva. Alternativt kan modellen skrivas som: $T_i = \exp(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi} + \sigma\varepsilon)$.

Från den generella formeln av AFT-modellen kan hasardfunktionen härledas. Hasardfunktionen för den *i*:te individen skrivs: $h_i(t) = \exp(-\eta_i) h_0(t / \exp(\eta_i))$, där $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_p x_{pi}$. Modellen skrivs utan intercept och felterm, men inkluderar en term, h_0, för baseline hasardfunktionen. h_0 är hasardfunktionen då kovariaterna är lika med noll (Collett 2003).

Till skillnad mot icke-parametriska eller semiparametriska modeller genererar parametriska modeller estimat som är mer konsistenta med den teoretiska överlevnadskurvan, detta givet att den faktiska överlevnadstiden följer den fördelning som antas. Parametrarna kan då estimeras så att överlevnads- och hasardfunktionen specificeras fullständigt. Ett problem med de parametriska modellerna kan dock vara att hitta rätt underliggande fördelning för hasardfunktionen.

3.5.1 Fördelningar för accelerated time modeller

Frekvent använda fördelningar med tillhörande överlevnadsmodeller är exponential, weibull, log-logistic, lognormal och generaliserad gamma. Nedan beskrivs hasardfunktionerna för nämnda modeller.

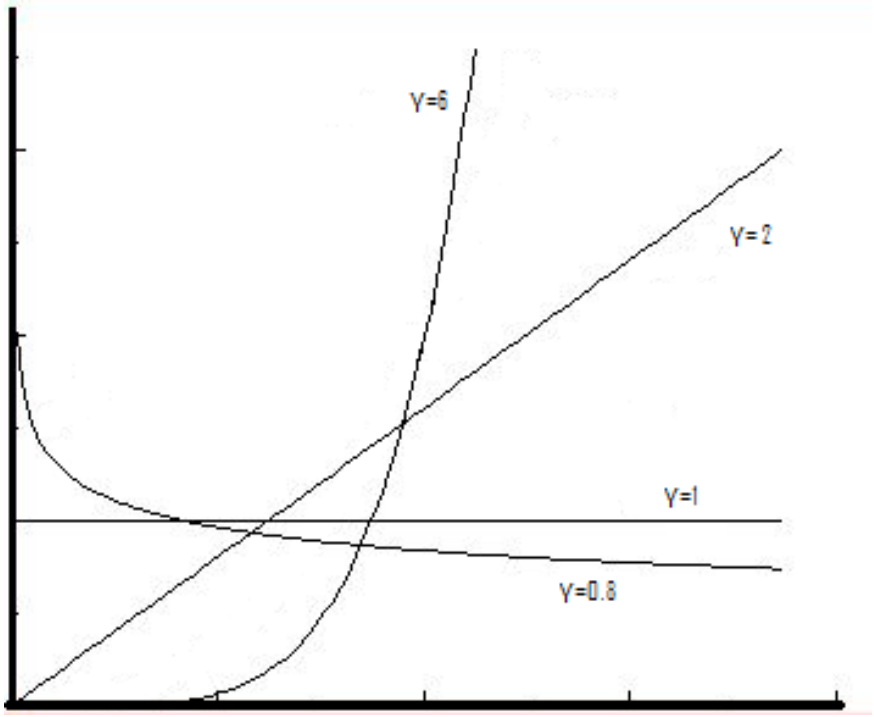
I den enklaste parametriska modellen följer överlevnadstiden en exponentialfördelning, där namnet kommer sig av att täthetsfunktionen *f(t)* har en exponentialfördelning. Modellen antar däremot att hasarden är konstant över tid enligt: $h(t) = \lambda$. Eftersom hasarden antas vara konstant följer att modellen är proportionell, det vill säga kvoten av hasarden mellan olika grupper/individer är konstant. Således tillhör modellen egentligen inte AFT.

En något mer generell modell än den exponentiella är weibullmodellen. Hasardfunktionen innehåller ett tillägg, jämfört med den exponentiella modellen, det vill säga en formparameter *p* och skrivs: $h_0(t) = \lambda\gamma t^{\gamma-1}$ och för individ *i*:

$h_i(t) = e^{-\eta_i}\lambda\gamma(e^{-\eta_i}t)^{\gamma-1}$. Värdet på $\gamma(= 1/\sigma)$ bestämmer formen av hasardfunktionen, då γ <1

eller γ >1 är hasarden avtagande respektive tilltagande med tiden. Weibullmodellen reduceras till en exponentiell modell då γ =1, det vill säga den exponentiella modellen är ett specialfall av weibullmodellen. Således kan modellen beskriva en avtagande, tilltagande och konstant hasard. I nedan figur visas exempel på hasardfunktioner givet vissa värden på γ.

Figur 1 Hasardfunktionen för weibull (källa http://www.engineeredsoftware.com)



För den log-logistiska modellen skrivs hasardfunktionen:
$h_0(t) = \frac{e^{\theta} K t^{K-1}}{1+e^{\theta} t^{K}}$. Fördelningen innehåller två parametrar: $K(= 1/\sigma)$och $\theta(= \mu/\sigma)$. Funktionen

implicerar att hasarden kan ta en maximipunkt, det vill säga hasarden tilltar till en viss punkt därefter är den avtagande, detta då K>1. Modellen har inte ett proportionalitetsantagande om hasarden, men ett liknande antagande som innebär att kvoten av överlevnadsoddsen ska vara konstant över tid[2].

Liksom den log-logistiska modellen tillåter log-normal modellen hasarden att öka till en maximipunkt för att sedan avta och gå mot noll när tiden går mot oändligheten. Modellen är inte en proportionell hasard modell, men ln$T$ antas vara normalfördelad med medelvärde μ och varians σ^2. Liksom för övriga modeller kan hasardfunktionen uttryckas som kvoten av

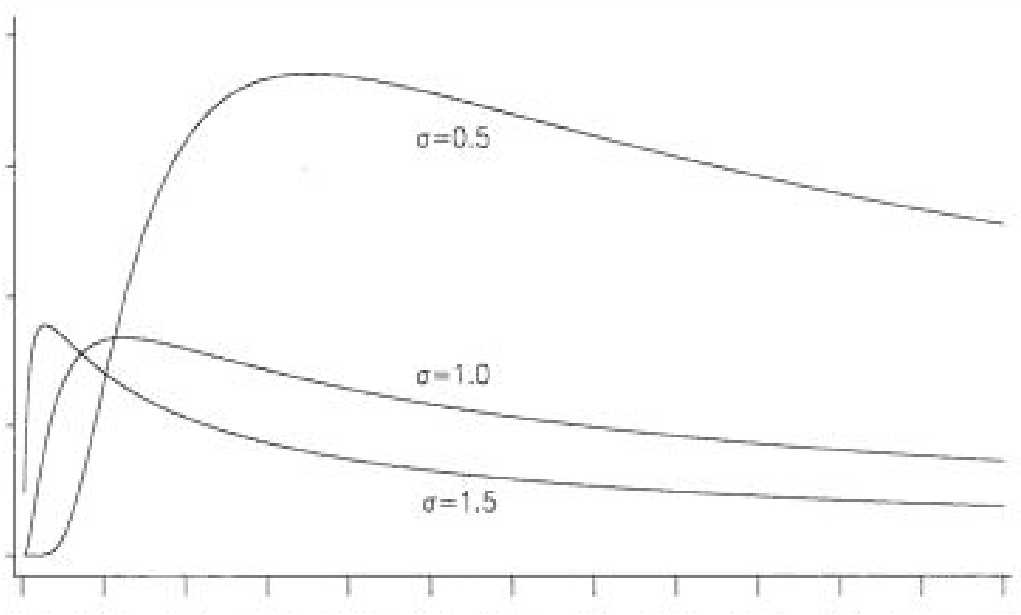
[2] Överlevnadsoddsen beräknas med: $\frac{S(t)}{(1-S(t))}$

täthetsfunktionen och överlevnadsfunktionen, där täthetsfunktionen är:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} t^{-1} \exp\left(-(\log t - \mu)^2 / 2\sigma^2\right)$$

och överlevnadsfunktionen: $S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$, där $\Phi(.)$ är fördelningsfunktionen för normalfördelning[3]. Hasardfunktionen är noll när *t* är noll och tilltar till ett maxivärde för att sedan gå mot noll när *t* går mot oändligheten.

Figur 2 Hasardfunktionen för log-normal givet olika värden på σ (källa Allison 2010)



Den generaliserade gammamodellen har ytterligare en parameter, ρ, att estimera, vilket innebär att hasardfunktionen kan ta fler former än de andra modellerna. Till exempel tillåter modellen en minimipunkt. Generaliserade gamma modellen kan dock inte hantera hasardfunktioner som byter riktning mer än en gång, vilket innebär att den till exempel inte kan beskriva en funktion med två maximipunkter.

[3] Vilket ges av:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-u^2/2\right) du$$

En nackdel med modellen är att den är ganska komplex vilket gör det svårt att bedöma formen på hasardfunktionen som estimeras med parametrarna. Täthetsfunktionen är: $f(t) = \frac{\theta \lambda^{\rho\theta-1} \exp\left[-(\lambda t)^{\theta}\right]}{\tau(\rho)}$ och överlevnadsfunktionen: $S(t) = 1 - \tau_{(\lambda t)^{\theta}}(\rho)$. (Collett 2003)

Då den generaliserande gamma funktionen är flexibel kan den anta flera av de andra modellerna, det vill säga de är specialfall av den generaliserande gamma modellen. När ρ=1 är fördelningen weibull, när ρ=1 och λ=1 exponential och när ρ=0 antar funktionen log-normal. Dessa egenskaper används för att utvärdera om de andra modellerna passar för ett datamaterial. (Allison 2010)

3.5.2 Test av fördelningen

Vilken modell med tillhörande fördelning som är lämplig för datamaterialet kan utvärderas med olika test. Visuellt kan detta testas med en graf över den estimerade hasard- eller överlevnadsfunktionen med hjälp av Kaplan-Meier eller life-table. Utvärderingen av grafen kan förenklas genom att ta logaritmen av överlevnadsfunktionen, multiplicera med minus ett och ytterligare en gång ta logaritmen, vilket motsvarar logaritmen av den kumulativa hasardfunktionen. För weibullmodellen skrivs denna funktion som: $\log\left(-logS(t)\right) = \log\lambda + p\log t$ (Collett 2003), där *S(t)* är estimaten från lifetable eller Kaplan-Meier. Om modellen passar väl ska den logaritmerade kumulativa hasradfunktionen uppträda linjärt mot den logaritmerade tiden.

Likelihoodfunktionen ger ett värde på hur mycket information datamaterialet innehåller om parametrarna i en modell. Likelihoodfunktionen är beroende av antal observationer och värdet i sig kan inte ge en indikation på hur bra modellen passar. Värdena för likelihoodfunktionen med olika modeller kan därmed enbart jämföras med samma datamaterial. Ett statistiskt test som inkluderar likelihoodfunktionen är *likelihood-ratio statistic*. Testet jämför om en modell är ett specialfall av en annan. Två modeller jämförs genom att beräkna skillnaden i log-likelihood mellan modellerna. Skillnaden multipliceras med två och ger ett chi-två värde för likelihood-ratio. Nollhypotesen i testet är att modellen är ett specialfall av den andra modellen. Alternativhypotesen är således att modellen inte är ett specialfall.

Då testet kräver att en modell kan vara ett specialfall av en annan modell så testas den exponentiella modellen mot weibullmodellen och weibull och log-normal modellerna testas mot den generaliserade gamma modellen. Testet bygger på att den mer generella modellen passar för datamaterialet. En nackdel med testet är således att den generaliserade gamma modellen inte kan testas mot någon annan modell, eftersom den inte är ett specialfall av någon annan modell. En befogad fråga att svara på är således varför inte den mest generella modellen, nämligen den generaliserade gamma modellen, alltid används? Allison (2010) nämner två anledningar: det är svårt att bedöma formen på hasardfunktionen och att funktionen rent matematiskt är komplicerad vilket innebär att det kan ta mycket datortid vid användning.

3.6 Piecewise exponential modellen

De parametriska modellerna som beskrivits ovan antar att hasarden är en relativt enkel funktion av tiden, då de endast tillåter en jämn form med som mest en maximi- eller minimipunkt. En möjlighet att rucka på detta antagande ges med piecewise exponential modellen. Modellen är inte strikt parametrisk men använder sig av exponentialfördelningen.

För att använda modellen delas tiden in i intervall. Inom varje intervall antas hasarden vara konstant, men den kan variera obegränsat mellan intervallen. Modellen kan således vara användbar i situationer där risken är konstant under en viss tid för att sedan abrupt tillta eller avta till en ny nivå för en annan tid. Intervallen bestäms något godtyckligt till exempel genom att studera den estimerade hasardfunktionen genom Kaplan-Meier. (Allison 2010)

3.7 Cox proportional hasard modell

I dagsläget är proportional hasard modell, även kallad Cox regressionsmodell, den övervägande vanligaste metoden inom överlevnadsanalys. Dess namn kan dock vara missledande på grund av att modellen lätt kan modifieras till att tillåta icke-proportionalitet.

Statistikern David Cox presenterade, förutom själva modellen, även en ny skattningsmetod som kom att heta maximum partial likelihood (partial likelihood beskrivs i avsnitt 3.8.2) och begreppet Cox regression syftar på kombinationen av modellen och skattningsmetoden. Begreppet proportional hasard modell är egentligen en generalisering av bland annat weibullmodellen medan maximum partial likelihood metoden är betydligt mer komplicerad. Det tog åratal för statistiker att begripa detta nya sätt att gripa sig an skattningar (Allison 2010).

Vi undersöker modellen och dess ändamål och behandlar den basala modellen utan tidsberoende kovariater. Modellen är vanligtvis skriven enligt följande ekvation:

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i x_i}$$

(1.1)

Modellen ger ett uttryck för hasarden i tidpunkt *t* för en individ med en bestämd uppsättning förklarande variabler betecknad som *X*. Detta innebär att *X* representerar en uppsättning kovariater som modelleras för att predicera en individs hasard. Cox proportional hasard modellens formel visar att hasarden vid tidpunkt *t* är produkten av två delar varav den första av dessa $h_0(t)$ kallas för *baseline hazard function* medan den andra är det exponentiella uttrycket *e* upphöjt till den linjära summan av $\beta_i X_i$ där summan är över *p* förklarande *X* variabler. En viktig egenskap hos denna formel som avser proportional hasard antagandet är att baseline hasarden är en funktion av *t* men innefattar inte *X* variablerna. Å andra sidan så är *X* variablerna innefattande i det exponentiella uttrycket:

$e^{\sum_{i=1}^{p} \beta_i x_i}$, men innefattar följaktligen inte tidpunkt *t*. *X* variablerna kallas i detta fall

tidsoberoende *X* variabler. En lämplig begreppsförklaring för tids-oberoende *X* variabler, till exempel kön och ursprung, är att dessa har värden som inte förändras över tid. Observera att fastän variabler, såsom ålder, förändras över tid så kan det vara lämpligt att behandla dessa variabler som tidsoberoende i analysen. Detta då dess värden inte förändras nämnvärt över tid eller ifall påverkan av sådana variabler på överlevnadsrisken huvudsakligen beror på värdet vid endast en mätning.

Cox modellens formel har egenskapen att om alla *X* variabler är lika med noll så reduceras formeln till baseline hasard funktionen. Detta innebär att den exponentiella delen i formeln blir e^0 vilket blir ett. Från ett något annorlunda perspektiv så reduceras Cox modellen till baseline funktionen när ingen *X* variabel finns med i modellen. $h_0(t)$ kan följaktligen ses som en måttstock av hasard funktionen innan man inkluderar *X* variabler (Kleinbaum och Klein 2005).

Då man logaritmerar båda sidor kan ekvation 1.1 omskrivas som

$$log\ h_i(t) = \alpha\ (t) + (\beta_1\ x_{i1} + \ldots + \beta_k\ x_{ik})$$

där $\alpha\ (t) = log\ h_0\ (t)$. Om vi specificerar $\alpha\ (t) = \alpha$ så får man den exponentiella modellen. I fall man vidare specificerar α (t) = α log t så får man fram weibullmodellen. Fördelen med Cox regressionsmodell är att dessa val inte är nödvändiga eftersom funktionen $\alpha\ (t)$ kan ha vilken form som helst, det vill säga den är inte beroende av att fördelningen specificeras.

Hasarden för varje individ är en bestämd proportion av hasarden relaterat för varje annan individ. För att se detta så kan man ta kvoten av hasarden för individerna *i* och *j* och applicera detta i ekvation (1.1).

$$\frac{h_i(t)}{h_j(t)} = exp\{\beta_1(x_{i1} - x_{j1}) + \ldots + \beta_k(x_{ik} - x_{jk})\} \qquad (1.2)$$

Det väsentliga med denna ekvation är att $h_o(t)$ tar ut varandra och som ett resultat av detta så är kvoten av hasardfunktionerna konstant över tiden (Allison 2010).

Formel 1.2 visar på det underliggande antagandet om proportionalitet som Cox proportional modellen vilar på, det vill säga att hasardkvoten är konstant över tid när man exempelvis jämför två bestämda uppsättningar kovariater. Detta innebär att hasarden för en individ följaktligen är stående i proportion till hasarden för varje annan individ och inte förändras över tid. Man kan exempelvis påträffa icke-proportionalitet i hasarden mellan två eller fler grupper ifall dessa korsar varandra i grafer mätt över tid.

En betydelsefull egenskap hos Cox proportional modellen, och som bidrar till att den är allmänt omtyckt, är att baseline hasard funktionen $h_0(t)$ är en ospecificerad funktion som gör att Cox proportional hasard modell är semiparametrisk. Denna egenskap skiljer sig från parametriska modeller, såsom exempelvis weibull och gamma, vars funktioner är fullständigt specificerade med undantag för värden på okända parametrar (Kleinbaum och Klein 2005).

3.7.1 Proportionalitetsantagande för Cox

Vilket beskrivits tidigare så vilar Cox proportional hasard modellen på antagandet om en proportionell hasard, eller annorlunda uttryckt en konstant hasard ratio över tid. Allmänt anses Cox modellen vara robust oavsett om antagandet är uppfyllt eller inte. Det råder dock delade meningar om hur ”viktigt” det är att modellen har en proportionell hasard. Allison (2010) menar att proportionalitetsantagandet är överdrivet och att anledningen till det stora fokuset på antagandet är att det gett namn till modellen. Vidare anser han att proportionalitetsantagandet sällan är fullständigt uppfyllt, vilket är fallet för nästan alla statistiska antaganden. Schemper (1992) anser däremot att användning av Cox-modellen då antagandet inte är uppfyllt kan leda till att tester av modellen inte är tillförlitliga och att hasarden för kovariaterna riskerar att under- och överskattas. Därför menar han att Cox-modellen måste kontrolleras innan slutsatser kan baseras på modellen.

3.7.2 Test för proportionalitetsantagandet

Vi redogör härmed för populära tester som tillhandahåller p-värden samt teststatistika för att undersöka antagandet om proportionalitet för en bestämd kovariat. Sådana statistiska test kan ge bättre möjligheter att dra slutsatser än vad som är möjligt enbart med hjälp av grafer. Antagandet om proportionalitet utvärderas ofta med hjälp av residualer.

Man kom att intressera sig för användningen av icke-linjära regressionsmodeller som antog att oberoende variabler påverkade hasardfunktionen (Kay 1977). Residualer var tidigare vanligtvis avgränsade i samband med linjära modeller(Cox och Snell 1968). Residualer används för att undersöka en modells lämplighet med avseende på ett datamaterial. För Cox proportional hasard modellen så finns det inte ett jämförligt enkelt tillvägagångssätt såsom ”observerat minus predicerad skattning” som används i linjära regressioner.

Det finns olika varianter av residualer för att utvärdera antagandet om proportionalitet, bland annat Schoenfeld och Martingale.

3.7.3 Schoenfeld residualer

David Schoenfeld presenterade 1982 ett sätt att testa proportionalitetsantagandet med hjälp av en uppsättning residualer som i idag kallas Schoenfeld residualer och kom att användas för Cox proportional hasard modell.

Dessa residualer kan kartläggas över tid för att testa antagandet för proportionalitet i Cox proportional hasard modell. Schoenfeld residualer kan användas för att identifiera värden som

avviker från det allmänna mönstret (s.k. outliers), som kan störa analysen (Schoenfeld 1982). Istället för endast en residual för varje individ så är det en separat residual för varje individ för varenda kovariat. För var och en av kovariaterna i modellen så är Schoenfeld residualer avgränsad för varje objekt som har en händelse. För att ta ett exempel så kan man tänka sig en Cox proportional hasard modell med två kovariater såsom ålder och utbildning.

I detta exempel så får vi två Schoenfeld residualer specificerade för varje objekt som har en händelse och innebär följaktligen en residual för varje enskild kovariat. Det huvudsakliga syftet med dessa residualer är att detektera avvikelser från antagandet om propotionalitet. Resonemanget till stöd för detta statistiska test är att i fall antagandet om proportionalitet uppfylls för en särskild kovariat så kan man konstatera att Schoenfeld residualerna för kovariaten inte har ett samband med tiden. Att testa antagandet om proportionalitet är således ekvivalent med att testa om Schoenfeld residualer är korrelerade med tiden eller en funktion av tiden. Om antagandet om proportionalitet är uppfyllt så skall korrelationen var nära noll (Kleinbaum och Klein 2005). P-värden för tester av korrelationer presenteras i korrelationsmatris i exempelvis SAS och nollhypotesen är följaktligen att antagandet om proportionalitet uppfylls.

Schoenfeld residualerna bör inte visa på några trender utan skall kretsa kring noll för att proportionalitet skall antas vara tillfredsställande. Om detta inte är fallet och hasarden förändras långsamt över tiden så bör en Schoenfeld residual plot påvisa detta. En positiv(negativ) trend antyder att hasarden ökar(minskar) över tiden (Farrington 2000).

Schoenfeld residualer är inte definierade för censurerade observationer utan behandlar dessa som ofullständiga värden.

3.7.4 Assess med martingale

I SAS används en av de mest förekommande residualer när man utvärderar proportionalitet i Cox modellen. Dessa residualer benämns martingale residualer och inkorporeras i SAS med hjälp av en metod som kallas Assess. Martingale residualer erhålls genom att transformera Cox-Snell residualer (för mer detaljer kring Cox-snell se bilaga 1) och ytterligare en känd residual, som dock inte behandlas i texten men benämns *deviance residualer* och är en transformering av martingale (Allison 1995). Till den som önskar en fördjupning kring residualer rekommenderas att läsa Collett (2003).

Till skillnad från metoden med Schoenfeld residualer så fungerar dock inte Assess i händelse av att Cox proportional hasard modell innehåller tidsberoende kovariater.

Med hjälp av Assess så efterfrågas en utvärdering av antagandet om proportionalitet och för varje kovariat produceras grafer (se bilaga 13) och p-värden som är baserade på martingale residualer. Däremot uppvisar Assess inte någonting om karaktären för överträdelsen av antagandet för proportionalitet.

3.7.5 Grafisk utvärdering med kumulativa hasardfunktionen

Vilket beskrivits tidigare kan logaritmen av den kumulativa hasardfunktionen användas för att utvärdera passformen av en parametrisk modell för datamaterialet. Funktionen kan även användas för att grafiskt utvärdera om proportionalitetsantagandet kan anses vara uppfyllt.

Utvärdering sker genom att studera den logaritmerade hasardfunktionen mot logaritmen av tiden för en kovariats alla kategorier. Om antagandet är uppfyllt ska kategorierna uppträda proportionellt mot varandra i grafen, det vill säga inte korsa varandra, konvergera eller divergera gentemot varandra (Collett 2003). Steffensmeier och Zorn (1998) skriver att denna grafiska utvärdering kanske är den vanligaste, men ifrågasätter användbarheten och rekommenderar istället mått som bygger på residualer. De menar att metoden ofta misslyckas med att korrekt diagnostisera icke-proportionalitet.

3.7.6 Metoder för att hantera icke-proportionalitet

Schemper (1992) föreslår tre lämpliga metoder för att hantera en modell där antagandet inte är uppfyllt: 1) stratifiera modellen med avseende på kovariaten som inte har en proportionell hasard, 2) dela in datamaterialet i disjunkta tidsperioder, 3) använd en tidsberoende kovariat i modellen.

Stratifiering är en lämplig metod då kovariaten som inte uppfyller antagandet om proportionalitet är kategorisk. Genom att stratifiera datamaterialet tillåts varje strata ha olika baseline hasardfunktioner och den proportionella hasardfunktionen för strata *s* ges av: $h_s = h_{so}(t)e^{x'\beta}$.

Skattningarna inom varje strata vägs samman och skattningar av β görs på hela datamaterialet. En nackdel med stratifiering är att det inte går att få estimat för kovariaten som datamaterialet stratifieras på. En förutsättning är därför att denna inte är av direkt intresse för analysen. (Homser 1999)

Grundproblemet då en kovariat inte uppfyller antagandet är att den influerar hasarden olika för olika tidpunkter. Kovariaten kan till exempel vara förknippad med en hög hasard i början av den undersökta tidsperioden och i slutet en väldigt låg hasard. För att komma till rätta med problemet kan datamaterialet delas in i två eller fler tidsperioder så att varje tidsperiod uppfyller antagandet om proportionalitet. Varje tidsperiod tillåts därmed ha olika baseline hasardfunktioner. En Cox-modell anpassas sedan för varje tidsperiod och separata analyser görs. Indelningen i tidsperioder kan baseras på grafer av den kumulativa hasardfunktionen för kovariaten (Schemper 1992).

För att lösa proportionalitetsproblemet kan även en ny variabel som är en interaktion med tiden introduceras i modellen. Denna variabel är en funktion av originalvariabeln, *x*, som uppvisar icke-proportionalitet och överlevnadstiden, *t*. Om till exempel variabeln x ökar exponentiellt med tiden kan den nya variabeln konstrueras genom att multiplicera *x* med kvadraten av *t*. Om interaktionen med tiden är rätt specificerad inrymmer den nya variabeln icke-proportionaliteten och modellen kommer således att vara proportionell. En nackdel med metoden är att det kan vara svårt att tolka original och den nya variabeln.

3.8 Estimering med maximum likelihood och partial likelihood

3.8.1 Maximum likelihood

I SAS skattas de parametriska modellerna med maximum likelihood metoden (ML). Sannolikheten (*eng. likelihood*) för datamaterialet räknas ut med varje individs täthetsfunktion (som är associerad med överlevnadstiden) enligt: $L = \prod_{i=1}^{n} f(t_i)$. Denna likelihoodfunktion

behandlar inte de observationer som censurerats utan bara de som faktiskt upplevt händelsen vid tidpunkt t_i. För att utnyttja den information som de censurerade observationerna ger används ett tillägg till likelihoodfunktionen i form av deras överlevnadsfunktion, enligt:
$L = \prod_{j=1}^{r} f(t_j) \prod_{l=1}^{n-r} S(t_l^*)$, där *r* samt *n-r* hänförs till icke-censurerade respektive censurerade

observationer. De censurerade observationernas överlevnadsfunktion används eftersom deras exakta överlevnadstid inte är känd, utan enbart att den är större än tidpunkten för när de censurerades [S(t)= P>t*]. Likelihoodfunktionen maximeras givet de okända parametrarna i täthets- och överlevnadsfunktionen, vilka beror på vilken parametrisk modell/fördelning som används. Observera således att hasardfunktionen måste vara känd. Maximeringen kan ske genom att derivera likelihoodfunktionen med avseende på beta-koefficienterna. Uttrycket sätts lika med noll och löses för beta. (Collett 2003)

3.8.2 Partial likelihood

Till skillnad mot de parametriska modellerna estimeras beta-koefficienterna något annorlunda för Cox proportional hasard modellen. Cox visade, 1972, hur betakoefficienterna kan estimeras då baseline hasardfunktionen är godtycklig. Cox jämför situationen med då parametern för feltermen är okänd.

Likelihoodfunktionen skrivs utan hasardfunktionen[4] och informationen om beta ges av de individer för vilka händelsen av intresse inträffar. Av denna anledning görs ett antagande om att de censurerade observationerna inte får vara informativa. Detta antagande påverkas främst då flera händelser är av intresse och det finns konkurrerande risker mellan händelserna. Eftersom denna studie endast fokuserar på skilsmässor kommer inte detta problem att tas upp här. Vidare estimeras beta med maximum likelihood. Dessa estimat används för att göra en separat maximum likelihood estimering av fördelningen som är associerad med hasarden för varje tidpunkt då händelsen inträffar. Baseline hasardfunktionen antas vara noll då händelser inte har inträffat (Cox 1972).

Likelihoodfunktionen som erhålls kom att kallas *partial likelihood function* eftersom den inte direkt använder sig av de faktiska censurerade och icke-censurerade överlevnadstiderna. Till skillnad mot maximum likelihood estimeringen vid de parametriska modellerna används inte

[4] Likelihoodfunktionen skrivs som ett uttryck där baseline hazardfunktionen finns i både nämnaren och täljaren, vilket tar ut varandra.

information om baseline hasardfunktion vid partial likelihood, vilket innebär att de estimerade beta-estimaten inte är fullt så effektiva (Collett 2003).

4. Beskrivning av datamaterial och variabler

I detta avsnitt redogörs för datamaterialet som ligger till grund för studien. En allmän beskrivning följs av en mer variabelspecifik beskrivning. Även tanken bakom konstruktionen av variablerna förklaras.

Rådatamaterialet kommer från SCB:s register över totalbefolkningen (RTB), flergenerationsregister och utbildningsregister (UREG). I nämnda register finns bland andra variabler som avser personnummer, civilstånd, registrerad partner, födelsedatum, födelseland, föräldrarnas födelseland, nyblivna änkor och änklingar, nyblivna gifta, nyblivna frånskilda, utvandring och utbildningsnivå.

I datamaterialet i denna studie hänförs varje observation till ett par som gifte sig för första gången någon gång under 1998. Individerna i paret följs till den 31 december 2008. Studietiden för en observation omfattar således som mest elva år. Förändringar under studietiden till exempel civilstånd registreras med datum vilket innebär att tiden mäts i dagar. Antal personer som ingår i studien och gifte sig för första gången under 1998 uppgick till 40 846 stycken. Detta resulterar i 20 423 observationer, det vill säga par som ingick äktenskap. Den faktiska siffran som gifte sig för första gången under 1998 uppgick till cirka 100 ytterligare par än nämnda siffra. På grund av att uppgifter saknades för dessa togs de bort från datamaterialet.

Det ges inte någon möjlighet att följa individer som är folkbokförda utanför Sverige, därmed behandlas de par som emigrerat som censurerade observationer. En observation behandlas även som censurerad från den dagen som någon i paret avlider. Under studietiden högercensurerades 202 observationer på grund av dödsfall, 624 observationer på grund av att både mannen och kvinnan emigrerat samt 15731 då de fortfarande var gifta efter studietidens sista dag. Totalt skilde sig 3866 par, vilket motsvara cirka 19 procent av det totala antalet par vid studietidens startdatum.

Händelsen som analyseras är skilsmässa vilken mäts i antal dagar från giftermålet. Denna tidsvariabel är den beroende variabeln i analysen. De oberoende variablerna är inte tidsberoende, det vill säga värdena på kovariaterna mättes enbart när giftermålet ägde rum. Till exempel används utbildningsnivån när paret gifte sig och förändringar under studietiden tas ej hänsyn till.

Kovariaterna är mannens ålder vid giftermålet, ålderskillnad mellan kvinnan och mannen vid giftermålet, utbildningsnivå för de ingående parterna samt en för ursprung.

4.1 Ursprungsvariabel

Ursprungsvariabeln är indelad efter var individerna samt deras föräldrar är födda. För tio observationer saknades uppgifter om ursprung, därför ingår de inte i analysen. Indelningen av denna variabel presenteras i nedan tabell.

Kategori	Makens samt makans ursprung	Makens respektive makans föräldrars ursprung	Antal observationer (antal skilsmässor) samt skilsmässor i procent
1.	Född utomlands	-	815 (179)
	Född utomlands	-	22 %
2.	Född i Sverige	Födda i Sverige	3418 (734)
	Född i Sverige	Minst en född utomlands	21 %
3.	Född i Sverige	Födda i Sverige	1740 (411)
	Född utomlands	-	24 %
4.	Född i Sverige	Minst en född utomlands	335 (92)
	Född i Sverige	Minst en född utomlands	27 %
5.	Född i Sverige	Minst en född utomlands	505 (139)
	Född utomlands	-	28 %
6.	Född i Sverige	Födda i Sverige	
(Jmfr.-kategori)	Född i Sverige	Födda i Sverige	13610 (2314) 17 %

Kategorierna i kovariaten ursprung tar hänsyn till när individerna i paret har samma eller olika ursprung, med reservation för att de par som är födda utrikes kan vara födda i olika länder. Kategoriseringen möjliggör att tillgodose olika konstellationer för vilka parterna kan antas ha olika starka band till ett utrikes land. En alternativ indelning har också använts där hänsyn tagits till om det är mannen eller kvinnan som har svensk eller utländsk bakgrund; läs mer om denna indelning i bilaga 14.

Personerna som är utrikesfödda analyseras som en ”homogen” grupp vilket är ett relativt starkt antagande, då det innebär att personer från närbelägna länder till Sverige tillhör samma kategori som länder långt ifrån Sverige. Det primära syftet med studien är dock att se till hur personer

med utländsk bakgrund tenderar att separera i förhållande till de med svensk bakgrund. Om födelseregion skulle beaktats skulle det innebära många fler kategorier, där vissa av dessa skulle ha ytterst få observationer. Därför fokuserar denna studie på om individerna som ingår äktenskapet samt deras föräldrar är födda utrikes. För att få en uppfattning vilken födelseregion individer i Sverige har se bilaga 2.

4.2 Utbildningsvariabel

Utbildningsnivå är en relativt vanligt förekommande variabel i dylika analyser. Förutom att ta hänsyn till just utbildningsnivån kan variabeln fungera som en proxyvariabel för inkomst, då högre utbildning kan antas leda till högre inkomst. För varje individ delas variabeln in efter förgymnasial, gymnasial och eftergymnasial. Förgymnasial definieras här som studier under nio eller färre år på grundskolenivå. Kravet för gymnasial utbildning definieras som en examen från en två- eller treårig gymnasial utbildning. För eftergymnasial utbildning måste individen studerat minst ett år vid högskola. Värt att observera för denna indelning är att vi ej kunnat urskilja om individen är student, vid giftermålet, och således kan antas erhålla en fullständig högskoleutbildning (exempelvis kandidat- eller magisterexamen) under studietiden. Av denna anledning anses individen genomgått en ”högskoleutbildning” efter endast ett års studerande vid högskola.

Ett problem med utbildningsvariabeln är att utbildningsnivå inte fanns registrerad för 802 individer. De flesta av dessa individer emigrerade någon gång under 1998. Eftersom utbildningsnivån registreras den sista dagen varje år finns ingen utbildningsstatistik om dessa individer att tillgå. Då klassificeringen av enskilda utbildningar (SUN) genomgick en omfattande revidering under 1998-1999 är det inte möjligt att använda 1997 års utbildningsnivå för dessa individer. Även då det rör sig om relativt få individer utgör de en viktig del, eftersom en stor andel av dessa tillhör den redan förhållandevis ringa gruppen *födda utomlands*. För att fortfarande behålla dessa individer i studien infördes en extra kategori då utbildning inte fanns registrerad på någon av individerna i paret.

Liksom för ursprungsvariabeln har även utbildningen delats in i kategorier efter båda individerna i paret. Detta för att kunna se till effekter där individerna i paret har olika utbildningsnivå. Variabeln har totalt sju kategorier vilka som beskrivs i nedan tabell.

Kategori	1.	2.	3.	4.	5.	6.	7. (Jmfr.-kategori)
Makens samt makans utbildningsnivå	Båda gymnasial utb.	Båda eftergymnasial utb.	En med förgymnasial utb. resp. en med gymnasial utb.	En med eftergymnasial utb. resp. en med förgymnasial utb.	En med gymnasial utb. resp. en med eftergymnasial utb.	Uppgifter saknas på minst en av individerna	Båda förgymnasial utb.
Antal (procentuell andel av totala antalet)	6184 (30)	5051 (25)	2589 (13)	593 (3)	4772 (23)	647 (3)	587 (3)

4.3 Variabler för ålder samt åldersskillnad

I tidigare studier har det visats att ålder vid giftermålet haft effekt på risken för separation. Denna variabel är ofta kategorisk för att lättare kunna tolka resultatet mot en jämförelsegrupp. Alternativet hade varit att mäta ålder ”kontinuerligt” i exempelvis dagar eller år. Tolkningen blir då effekt för varje vald tidsenhet. Vi har inte funnit att tidigare studiers åldersindelning varit relevant för vår studie, dessa studier har varit amerikanska och åldern vid giftermål kan antas vara betydligt mycket lägre i USA än i Sverige[5]. Åldersvariabeln är något godtyckligt framtagen där utgångspunkten varit att varje kategori ska vara någorlunda homogen. Vi har även valt att dela in variabeln så att den kan sägas representera kohorter; i alla fall vad avser de två äldsta grupperna samt den yngsta gruppen, för vilka var födda innan 60-talet, huvuddelen under 60-talet samt under 70-talet.

Eftersom vi även ser till effekten av åldersskillnaden mellan parterna skulle hög korrelation råda mellan åldersvariablerna (dvs. åldersskillnad, mannens samt kvinnans ålder) om bådas ålder beaktades. Därför baseras variabeln ålder vid giftermål endast på mannens ålder. Kategorierna för hur variabeln delas in återfinns i tabellen nedan.

[5] Sweeney och Phillips samt Zhang och Hook har den yngsta gruppen som kvinnor <20 år respektive den äldsta som kvinnor >30 år.

Kategori	1.	2.	3.	4. (Jmfr.-kategori)
Mannens ålder vid giftermålet	>39 år	30-39 år	27-29 år	<27 år
Antal (procentuell andel av totala antalet)	1842 (9)	10541 (52)	4792 (23)	3248 (16)

Eftersom åldersskillnaden mellan parterna kan tänkas påverka risken för skilsmässa har en kategorisk variabel som tillgodoser detta funnits med i modellen. Liksom Sweeney och Phillips (2004) delar vi in denna variabel enligt följande: 1) mannen mer än 5 år äldre än kvinnan [M>5 år], 2) mannen inte mer än 2 år yngre och som mest 5 år äldre än kvinnan [M>-3 år till M<6 år], 3) mannen minst 3 år yngre än kvinnan [M<-2 år].

5. Resultat

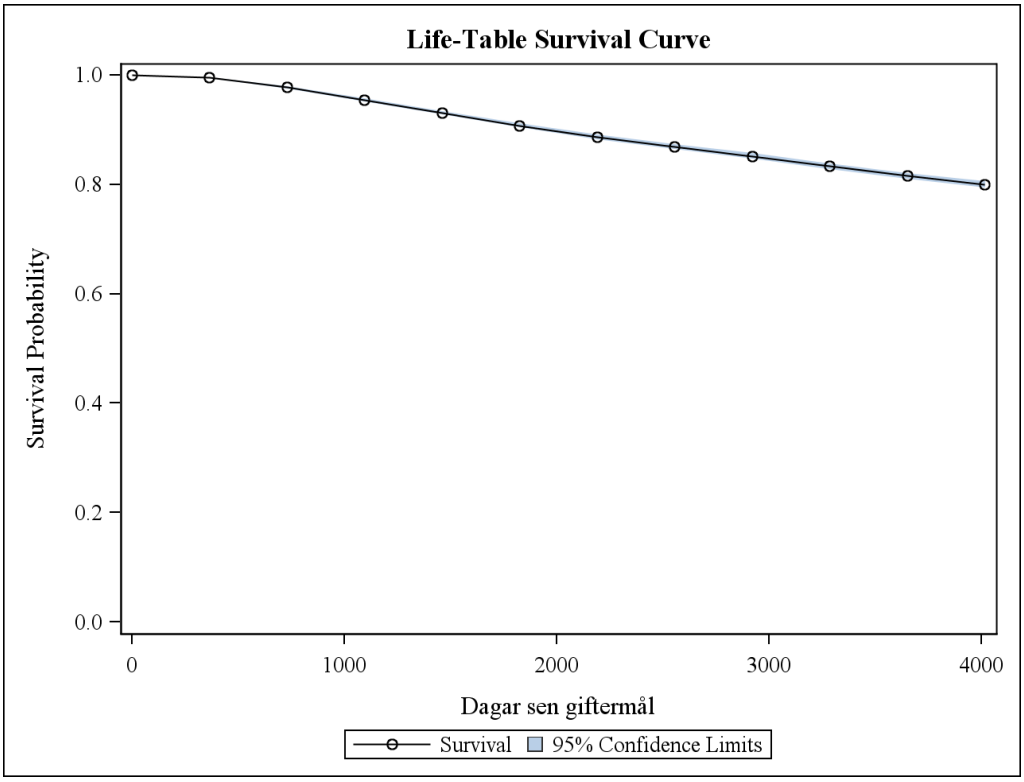
Kapitlet inleds med en skattning av överlevnadskurvan och hasardfunktionen. Därefter utvärderas kovariaternas effekt på överlevnadstiden. Vidare jämförs de parametriska modellerna med Cox regressionsmodell. Deras antaganden samt modifieringar av modellerna beskrivs.

5.1 Skattning av överlevnads- och hasardfunktion

För att få en preliminär uppfattning av datamaterialet skattas överlevnads- och hasardfunktionen. Skattningarna tar inte hänsyn till kovariaterna, då funktionerna endast beräknas på överlevnadstiderna. Därmed antas att hela datamaterialet har samma överlevnads- och hasardfunktion. Funktionerna skattas med Life-table och Kaplan-Meier metoden.

I figur 3 visas överlevnadskurvan skattad med Life-Table metoden. Tiden delas in i elva intervall där varje intervall representerar 365 dagar. Punkter på kurvan beskriver sannolikheten att ett (genomsnittligt) par fortfarande är gifta till tidpunkten *t*. Således är sannolikheten att paret fortfarande är gifta efter elva år (eller 4015 dagar) cirka 80 procent. Även då tiden delas in i betydligt fler intervall med Kaplan-Meier metoden ges ett snarlikt utseende (se bilaga 3). Av vad som framgår av konfidensbanden i figuren (knappt synbara) råder det ytterst liten osäkerhet i estimeringen.

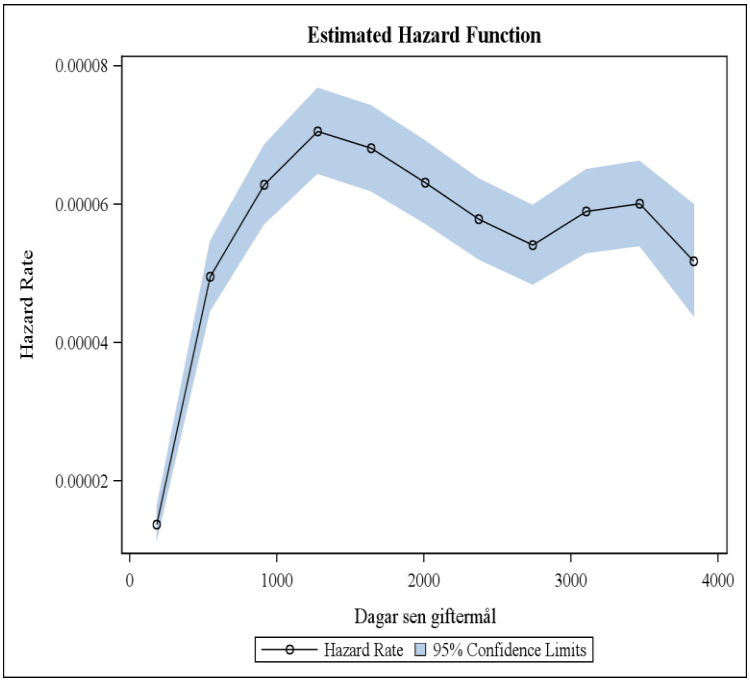
Figur 3 Överlevnadsfunktionen med life-table metoden (indelat per år)



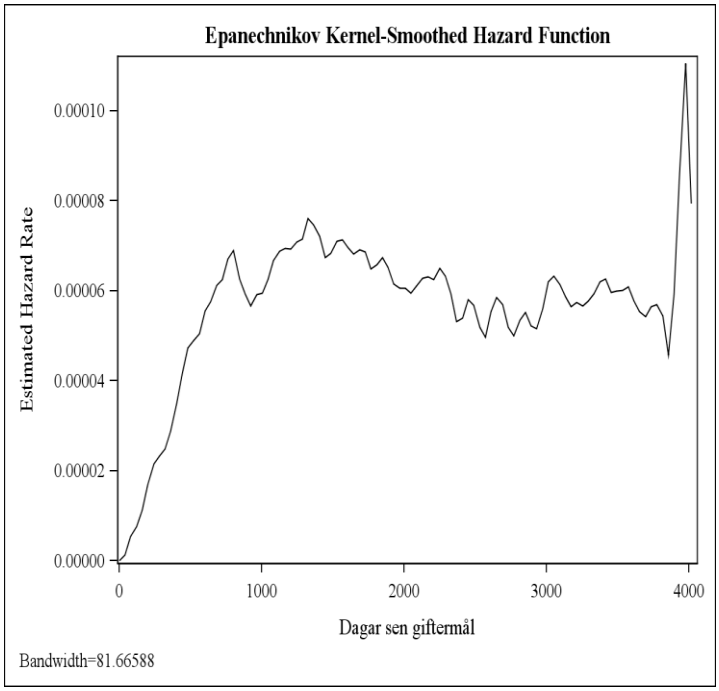
En estimering av hasardfunktionen med life-table metoden visas i figur 4. Enligt figuren ökar risken för skilsmässa under de tre första åren för att där nå sin högsta nivå. Intressant är även att risknivån tycks höjas mellan sju och nio år efter giftermålet.

Till skillnad mot överlevnadskurvan som inte skiljde sig nämnvärt mellan Kaplan-Meier och lifetable metoden så kan vi för hasardfunktionen se att kurvorna skiljer sig. Kaplan-Meier funktionen i figur 5 har utjämnats för att enklare kunna urskilja trender i figuren. Denna estimering visar ett mer oregelbundet mönster, vilket inte är förvånande då intervallen för vilka metoden använder vid beräkningen av risken (eller intensiteten) är betydligt kortare. I stort sett följer riskintensiteten samma mönster som vid life-table. Detta gäller fram till början av det elfte året då grafen uppvisar en kraftigt förhöjd skilsmässorisk. Det går dock inte att uttala sig om slutet av studieperioden då relativt få par finns kvar vid denna tid (endast de som gifte sig under 1998 års första månader och inte censurerats eller skilt sig). Slutet av studiperioden är således förknippad med en väldigt stor osäkerhet, det är även av denna anledning som funktionen visas utan konfidensband. Då konfidensbanden inkluderas i figuren komprimeras y-axeln på grund av osäkerheten för sista delen av tidsperioden, således blir det svårt att uttyda hasardfunktion (för figur med konfidensband se bilaga 4).

Figur 4 Hasardfunktionen med life-table metoden (indelat per år)



Figur 5 Hasardfunktionen med Kaplan-Meier



I bilaga 5 återfinns bland annat siffror för vilka life-table estimaten baseras på. Även hasarden, överlevnaden och täthetsfunktionen presenteras. Från denna tabell visas till exempel att hasarden var som högst under år fyra då totalt 490 personer genomgick skilsmässa.

5.2 Test av skillnader mellan grupper och kovariaternas effekt

Då syftet i studien i mångt och mycket bygger på att utvärdera skillnader mellan olika grupper är det relevant att testa om överlevnadsfunktionerna för dessa grupper skiljer sig åt. Det ger även en indikation på om indelningarna är relevanta att använda. För att tillgodose skillnader mellan grupperna används log-rank och Wilcoxon test. Nollhypotesen i dessa test är att grupperna är homogena sett till överlevnadsfunktionerna. I tabellen nedan visas chi-två värdena med tillhörande p-värden för de kovariater vi valt att använda.

Variabler	**Frihetsgrader** [= antal kategorier-1]	**Log-Rank: Chi-två (p-värde)**	**Wilcoxon Chi-två (p-värde)**
Ursprung	5	140.2 (<.0001)	139.2 (<.0001)
Åldersskillnad	2	91.4 (<.0001)	92.5 (<.0001)
Mannens ålder	3	193.8 (<.0001)	205.1 (<.0001)
Utbildning	6	817.5 (<.0001)	826.0 (<.0001)

Oberoende av vilket test som används blir slutsatsen att nollhypotesen kan förkastas för samtliga kovariater, vilket innebär att vi finner bevis för att minst en överlevnadsfunktion i varje kategori

är skiljd från de övriga. För den intressantaste kovariaten *ursprung* presenteras var och en av kategoriernas överlevnadsfunktioner grafiskt i bilaga 6.

Log rank och Wilcoxon testen kan även användas för att studera om kovariaterna är förknippade med tiden till skilsmässa. För att genomföra detta test krävs dock att kovariaterna är kvantitativa, vilket innebär att vår kategorisering inte kan användas. För detta test har således åldersskillnaden beräknats som antal år som skiljer mannen och kvinnan åt och mannens ålder behandlas i år. För utbildning och ursprung är en kvantifiering dock inte möjlig vilket innebär att dessa uteslutits från detta test.

Univariat test med log-rank	Test statistika	Standard fel	Chi-två	P-värde
Mannens ålder	4045.7	351.2	132.7	(<.0001)
Åldersskillnad	-1618.5	236.4	46.9	(<.0001)

Det univariata testet, där kovariaterna behandlas oberoende av varandra, av log-rank uppvisar höga chi-två värden för båda variablerna. Slutsatsen blir således att koefficienterna för kovariaterna är skiljda från noll. Det negativa värdet på test statistikan för åldersskillnad tolkas som ju större åldersskillnad desto kortare äktenskap.

Det är även önskvärt att se om kovariaterna är högt korrelerade till varandra det vill säga förklarar samma sak. Av denna anledning genomförs ett *forward stepwise* test med log-rank. Testet visar visserligen att chi-två värdet för åldersskillnaden förändras när vi kontrollerar för mannens ålder, men att åldersskillnaden fortfarande är signifikant. Båda kovariater är således betydelsefulla för att förklara tiden till skilsmässa.

Även då vi enbart visat att kovariaterna är relevanta för modellen i kvantitativ form torde vi kunna dra slutsatsen att de även tillför modellen uppdelade i kategorier, speciellt som kategorierna tycks vara skiljda från varandra.

Forward stepwise med log-rank	Frihetsgrader	Chi-två	P-värde
Mannens ålder	1	132.7	(<.0001)
Åldersskillnad	2	168.6	(<.0001)

5.3 Parametriska metoder

5.3.1 Utvärdering av fördelning/modell

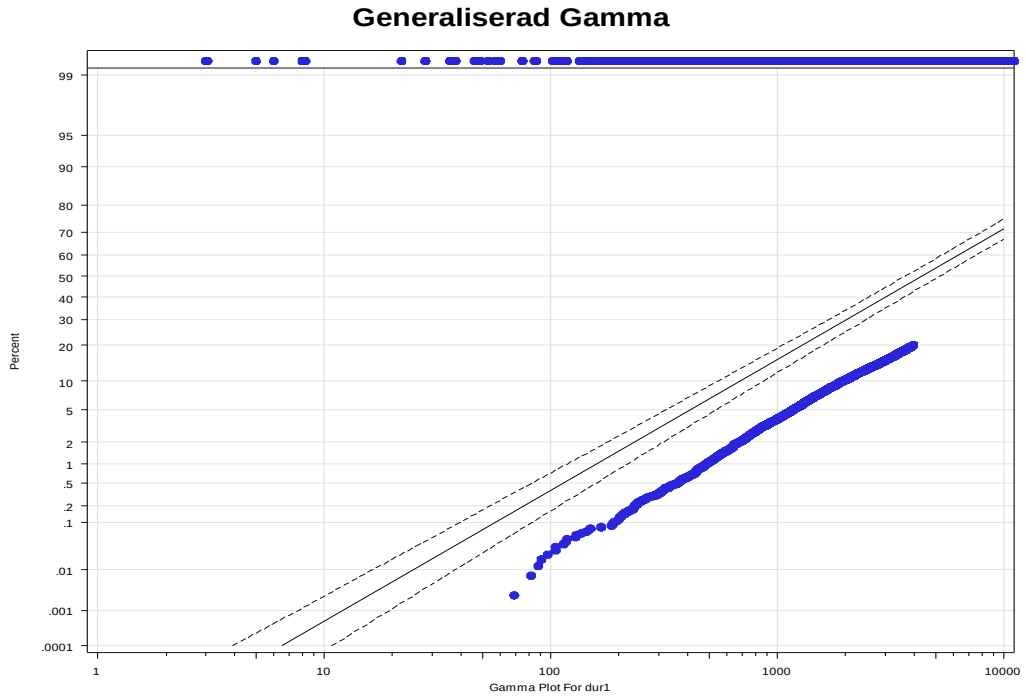
Hasardfunktionerna i figur 4 och 5 kan användas för att ge en preliminär indikation på vilken fördelning och således vilket modell som är lämplig för datamaterialet. Av vad som beskrivits i

metodavsnittet finns även statistiska test för utvärdering. För att genomföra testen krävs att materialet körs med respektive parametrisk modell. De modeller som stöds av SAS i *lifereg*-proceduren och som vi ämnar testa är exponentiell, weibull, log-logistic, log-normal och generaliserade gamma. I nedan tabell visas log-likelihood och likelihood ratio testet. Likelihood ratio används för att testa om log-normal, weibull eller den exponentiella modellen är ett specialfall av generaliserade gamma.

Modell	Log-likelihood	Frihetsgrader	Likelihood ratio statistika	P-värde
Exponentiell	-12601.8	2	436	<0.01
Weibull	-12475.9	1	184.2	<0.01
Log-logistic	-12439.8	-	-	-
Log-normal	-12384.4	1	1.2	>0.1
Gen. gamma	-12383.8	-	-	-

Testet ger vid handen att varken den exponentiella eller weibull är ett specialfall av generaliserade gamma modellen. Således kan inte datamaterialet sägas följa någon av dessa fördelningar. Då den log-logistiska modellen inte kan vara ett specialfall av generaliserade gamma kan inte likelihood ration utvärderas. Men log likelihood-värdet för den log-logistiska modellen är lägre än motsvarande värde för generaliserade gamma, vilket tyder på att även denna modell inte är lämplig. För log-normal kan vi dock inte förkasta nollhypotesen vilket ger stöd för att den är ett specialfall av generaliserade gamma. Vilket har nämnts i metodavsnittet kan den mest generaliserbara metoden, generaliserade gamma, inte utvärderas genom detta test.

Det är dock möjligt att utvärdera passformen av generaliserade gamma visuellt genom en sannolikhetsplot där en transformation av överlevnadsestimaten används. Den estimerade överlevnadsfunktionen visas som en linje med ett 95 procentigt konfidensband runtom.



Vilket visas i sannolikhetsploten av generaliserade gamma följer inte händelserna den estimerade överlevnadskurvan. Modellen passar dock betydligt mycket bättre än exempelvis exponentiell, för vilken finns en graf i bilaga 7. Med den visuella utvärderingsmetoden får vi ytterligare stöd för att log-normal är ett specialfall av generaliserade gamma då modellerna har en approximativt liknande passform (se bilaga 8). Värt att notera här är att då inte utbildning inkluderas i modellen får log-normal en tillsynes bra passform (se bilaga 9).

5.3.2 Skattning med log-normal och generaliserade gamma

Slutsatsen från testen av de parametriska modellerna är att ingen fördelning med tillhörande modell kan sägas passa datamaterialet. Log-normal och generaliserade gamma verkar dock vara de mest lämpliga, för dessa modeller presenteras estimaten i tabellen nedan. Modellen som estimeras kan skrivas på den allmänna formeln för AFT-modeller enligt:

$$logT = \beta_0 + \beta_{11}x_1 + \beta_{12}x_2 + \ldots + \beta_{14}x_4 + \beta_{21}x_5 + \beta_{22}x_6 + \ldots + \beta_{26}x_{10} + \beta_{31}x_{11} + \beta_{32}x_{12} + \beta_{33}x_{13} + \beta_{41}x_{14} + \beta_{42}x_{15} + \ldots + \beta_{47}x_{20} + \sigma\varepsilon$$

	Log-normal			Gen. gamma			
	Beta-estimat	Std.fel	Chi-två	Beta-estimat	Std.fel	Chi-två	$e^{\hat{\beta}}$
Intercept (k_0)	8.36	0.09	9322**	8.34	0.09	8603 **	
Mannens ålder >39 år (k_{11})	0.80	0.06	155**	0.81	0.06	155**	2.25
30-39 år (k_{12})	0.41	0.04	109**	0.42	0.04	108**	1.52
27-29 år (k_{13})	0.31	0.04	49**	0.31	0.04	49**	1.36
<27 år (k_{14} Jmfr.-kat.)	0.00	.	.	0.00	.	.	1.00
Ursprung[6] 1 (k_{21})	-0.06	0.07	0.6	-0.07	0.07	0.8	0.93
2 (k_{22})	-0.20	0.04	28**	-0.20	0.04	27**	0.82
3 (k_{23})	-0.32	0.05	40**	-0.32	0.05	40**	0.73
4 (k_{24})	-0.32	0.10	10**	-0.32	0.10	10**	0.73
5 (k_{25})	-0.33	0.08	16**	-0.34	0.08	16**	0.71
6 (k_{26} Jmfr.-kat.)	0.00	.	.	0.00	.	.	1.00
Åldersskillnad M>5år (k_{31})	-0.35	0.06	31**	-0.35	0.06	31**	0.70
M>-3år till M<6år (k_{32})	0.04	0.05	0.6	0.04	0.05	0.6	1.04
M<-2år ((k_{33} Jmfr.-kat.)	0.00	.	.	0.00	.	.	1.00
Utbildning 1 (k_{41})	0.77	0.07	113**	0.78	0.07	110**	2.18
2 (k_{42})	1.30	0.08	284**	1.30	0.08	276**	3.67
3 (k_{43})	0.43	0.08	32**	0.44	0.08	32**	1.55
4 (k_{44})	0.60	0.10	35**	0.61	0.10	35**	1.84
5 (k_{45})	1.11	0.08	215**	1.12	0.08	209**	3.06
6 (k_{46})	0.77	0.11	51**	0.78	0.11	50**	2.18
7 (k_{47} Jmfr.-kat.)	0.00	.	.	0.00	.	.	1.00
Scale	1.41	0.02		1.47	0.05		
Shape	-0.09	0.08					

**** signifikansnivå 0.01 * signifikansnivå 0.05**

[6] Indelning för *Ursprung* och *Utbildning* beskrivs på följande sida.

Indelning för kovariaterna Ursprung och Utbildning

Utbildning: 1, båda gymnasial 2, båda eftergymnasial 3, förgymnasial resp. gymnasial 4, eftergymnasial resp. förgymnasial 5, gymnasial resp. eftergymnasial 6, uppgifter saknas på minst en 7, båda förgymnasial.

Ursprung: 1, födda utrikes 2, båda födda i Sve. För en individ är minst en förälder född utrikes 3, född i Sve. med föräldrar födda i Sve. resp. född utrikes 4, båda födda i Sve. båda har minst en förälder född utrikes 5, född i Sve. med minst en utrikes förälder resp. född utomlands 6, födda i Sve. med föräldrar födda i Sve.

Av vad som framgår i ovan tabell ger båda modeller mycket liknande resultat, då standardfelen och estimaten oftast endast skiljer på hundradelen. I tabellen anges även *shape-parametern*, ρ, för generaliserade gamma, detta värde är nära noll vilket ger ytterligare stöd för att log-normal är ett specialfall.

Eftersom vi har kategoriska variabler tolkas koefficienterna efter vilket tecken de antar. Ett negativt värde indikerar att gruppen har en kortare tid till skilsmässa än jämförelsegruppen. Wald test genererar ett chi-två värde som används för att testa om koefficienten är signifikant skiljd från noll det vill säga från jämförelsegruppen. För att få en mer intuitiv förståelse för beta-estimaten, kan exponenten av dessa beräknas. Då denna term subtraheras med ett och multipliceras med 100 kan man visa att den kan tolkas som den förväntade tiden till skilsmässa i procent jämfört med jämförelsegruppen. I tabellen har e^{β} beräknats på estimaten med generaliserade gamma modellen.

Alla kategoriseringar för mannens ålder är signifikant skiljda från de som är yngre än 27 år. Samtliga övriga grupperna det vill säga när mannen är 27 år eller äldre visar sig ha en längre tid till skilsmässa. Koefficienterna indikerar även att de äldsta männen tenderar att vara gifta längst tid och ju yngre mannen är vid giftermålet desto kortare tid varar giftermålet. De äldsta männens äktenskap förväntas hålla 125 procent längre än de yngsta männens äktenskap.

Tolkningen av koefficienterna för ursprung indikerar att samtliga grupper har kortare äktenskap än jämförelsegruppen, det vill säga de par där båda samt deras föräldrar är födda i Sverige. Koefficienterna är signifikanta för alla grupper utom den som refererar till par där båda är födda utrikes. Sett till varaktighet av äktenskapet kan således inte denna grupp sägas vara annorlunda än jämförelsegruppen. Resterande grupper, för vilka en är född utrikes eller minst en är andra generationens invandrare, uppvisar liknande koefficient-värden. Grupp 2 tenderar visserligen att ha något längre äktenskap än grupp 3,4 och 5. I grupp 2 är båda födda i Sverige och en individ är andra generationens invandrare (med minst en förälder född utrikes). De konstellationer av par där ena parten är född utrikes och den andra är andra generationens invandrare förväntas ha ett äktenskap som varar 29 procent kortare än jämförelsegruppen.

Beta-estimaten för åldersskillnad visar att då mannen är mer än fem år äldre än kvinnan tenderar äktenskapet att hålla kortare tid än då de är någorlunda jämngamla eller om kvinnan är tre eller fler år äldre.

Utbildningsnivån av individerna i paret påverkar äktenskapets varaktighet, enligt estimaten av koefficienterna. Alla kategorier under denna kovariat har signifikant mycket längre äktenskap än de par där båda innehar en förgymnasial utbildning, det vill säga studerat nio år eller kortare. De par som tycks vara gifta längst tid är de där båda studerat på högskola. Dessa förväntas vara gifta mer än tre gånger längre än par där båda har förgymnasial utbildning. Även de par där parterna har olika långa studieerfarenheter är markant skiljda från jämförelsegruppen. För de par där skillnaderna är som störst det vill säga när en med förgymnasial utbildning är gift med en med eftergymnasial utbildning tenderar att ha längre äktenskap än par med förgymnasial respektive gymnasial. Generellt tyder resultaten på att de par där minst en har förgymnasial utbildning har kortast "äktenskapstid".

Förrutom de parametriska modellerna som presenteras ovan har en piecewise exponential modell använts. Skattningarna med denna modell är inte nämnvärt skiljda jämfört med skattningarna för generalsierade gamma eller log-normal. Resultatet för piecewise exponential presenteras i bilaga 10.

5.4 Semiparametrisk modell med Cox proportional hasard

Ingen av de parametriska modellerna tycks ha en optimal passform för datamaterialet. Generaliserade gamma och log-normal är de modeller som hitintills passat bäst. Det är därför intressant att se hur en modell som inte bygger på ett antagande om materialets fördelning lämpar sig. Cox proportionell hasard kräver inte att fördelningen specificeras, men bygger istället på att hasardfunktionerna är proportionella.

För att utvärdera om detta antagande uppfylls testas modellen med Schoenfelds residualer, SAS Assess test och en plot av den logaritmerade kumulativa hasarden.

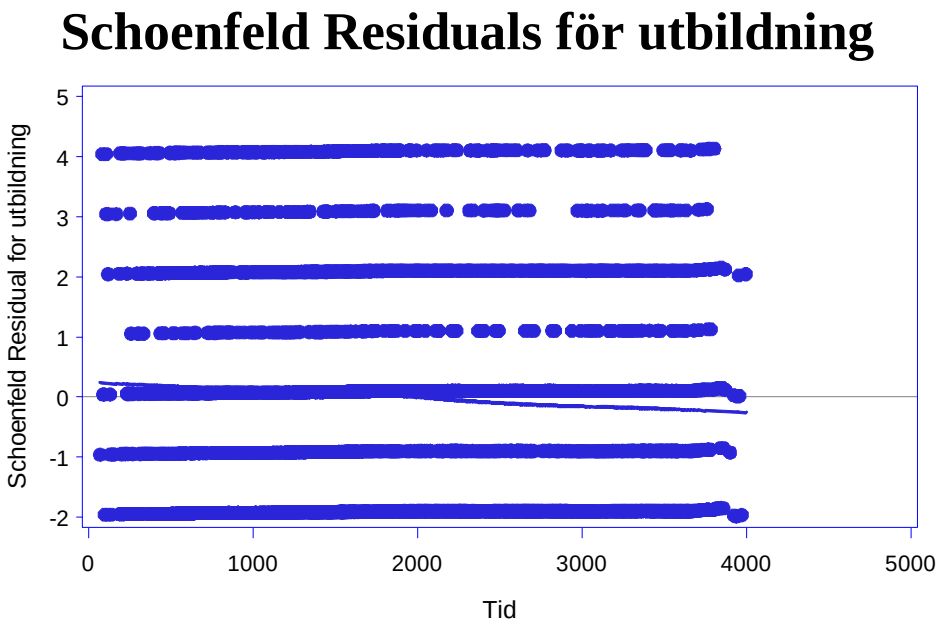
5.4.1 Schoenfeld residualer

Om proportionalitetsantagandet är uppfyllt ska Schoenfeld residualerna inte vara korrelerade med tiden. Sambandet mellan kovariaternas residualer och olika funktioner av tiden visas i nedan tabell. Residualerna testas mot vanliga funktioner av tiden, vilka förrutom tiden innefattar den kvadrerade tiden och logaritmen av tiden.

Korrelationsmatris mellan Schoenfeld residualerna och tiden (p-värde)

Residualer för:	Tid	Log. av tid	Tid^2
Mannens alder	-0.09 (<.01)	-0.09 (<.01)	-0.09 (<.01)
Åldersskillnad	-0.01 (0.37)	-0.01 (0.53)	-0.01 (0.37)
Ursprung	0.01 (0.37)	0.02 (0.17)	0.01 (0.54)
Utbildning	-0.07 (<.01)	-0.07 (<.01)	-0.07 (<.01)

Vi ser tydliga problem kring antagandet för proportionalitet med hänseende till mannens ålder och utbildningsnivån. I ovanstående tabell visar dessa på mycket låga p-värden för alla tre funktioner av tiden och kan därmed antas vara korrelerade med tiden. Däremot förefaller antagandet för proportionalitet, för alla tre tidsfunktioner, uppfyllas för åldersskillnad samt ursprung. Nedan illustreras korrelationen mellan residualerna och tiden för utbildning.



Observationerna redovisas med grövre streckade linjer för de sju grupperingarna relaterat till utbildning och den tunna linjen som kretsar kring nollstrecket representerar Schoenfeld residualerna. För att antagandet om proportionalitet skall vara uppfyllt så skall korrelationen mellan residualerna och tiden var nära noll. Som vi ser i figuren avviker Schoenfeld residualerna från noll, och visar i detta fall en negativ trend, som indikerar att utbildning är en beroende kovariat av tid. Detta innebär att proportionalitetsantagandet inte är uppfyllt och överensstämmer således med det statistiska testet. Schoenfeld Residualerna för de övriga kovariaterna presenteras i bilaga 12.

5.4.2 Assess

I Assess-tabellen exponeras p-värden för varje kovariat i vår modell och med hjälp av dessa utvärderas huruvida de följer proportionalitetsantagandet.

Supremum test för proportionalitetsantagandet

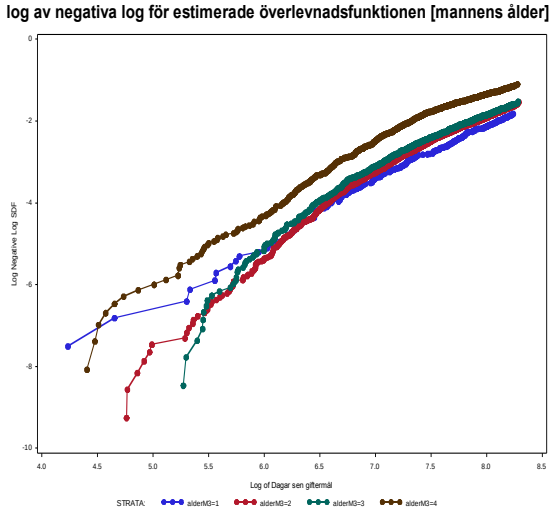
Residualer för:	Maximum absolut värde	p-värde
Mannens alder	3.32	<0.001
Åldersskillnad	1.28	0.172
Ursprung	0.83	0.442
Utbildning	2.73	<0.001

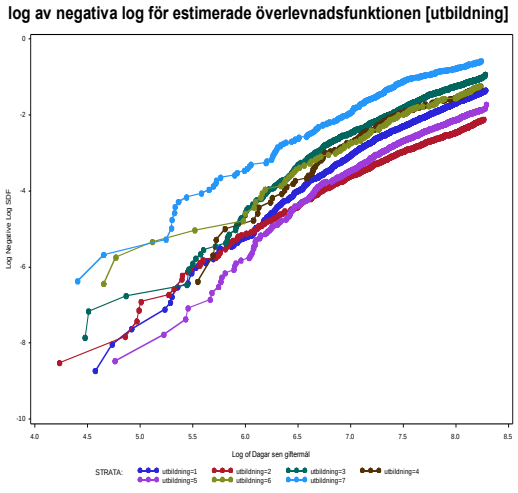
I tabellen ovan summeras resultaten av Assess för varje kovariat i vår modell. Liksom testet med Schoenfeld residualer kan vi tyda att åldern för män samt individernas utbildningsnivå inte kan antas vara proportionella. Detta ser vi genom att mycket låga p-värden exponeras för dessa kovariater.

Emellertid uppvisar tabellen tämligen höga p-värdena för ålderskillnad och ursprung, 0.172 respektive 0.442. Även i detta test förefaller åldersskillnaden mellan individerna som ingår äktenskap samt ursprung uppfylla antagandet om proportionalitet vilket innebär att betydelsen för dessa kovariater är konstanta över tid.

5.4.3 Grafisk utvärdering med den kumulativa hasardfunktionen

Proportionaliteten kan även utvärderas visuellt genom en plot av den logaritmerade kumulativa hasarden mot den logaritmerade tiden. Vilket beskrivits tidigare ska kategorierna för kovariaten uppträda proportionellt med varandra. Nedan visas illustrationer för utbildning och mannens ålder.





Från ovan grafer framgår att kategorierna under utbildningen och mannens ålder inte uppträder proportionellt. Detta visas då linjerna korsar varandra. Samtliga test som presenteras visar således på att antagandet om proportionalitet inte är uppfyllt för mannens ålder och utbildningskovariaterna.

5.5 Modifierad Cox-modell

Eftersom testen visar tecken på att mannens ålder och utbildning inte är proportionella försöker vi åtgärda modellen för att uppfylla antagandet.

Vilket beskrivits tidigare kan en lämplig metod vara att stratifiera modellen med avseende på en kovariat som är icke-proportionell. Nackdelen med metoden är att det inte går att få estimat för kovariaterna som datamaterialets startifieras med avseende på. Då det i vårt fall rör sig om två kovariater som inte uppfyller antagandet försöker vi lösa problemet genom att dela upp studietiden i intervall. Förhoppningen är att kovariaterna inom varje intervall uppträder proportionellt.

Genom att studera plottarna av de logaritmerade kumulativa hasardfunktionerna för mannens ålder och utbildning är det svårt att se någon naturlig brytpunkt för indelning i tidsintervall där proportionalitet kan tänkas uppfyllas. Schemper rekommenderar att tidsperioden delas in efter mittpunkten för tiden. På detta vis får vi två tidsperioder med relativt många observationer i varje. Första tidsperioden inkluderar således de 2008 första dagarna och den sista dag 2009 till 4017, vilken är den sista observerade dagen i studien. I den första tidsperioden behandlas alla observationer som inte genomgått skilsmässa efter 2008 dagar som högercensurerade. För den andra tidsperioden antar modellen att studien startar dag 2009, det vill säga de observationer som skilt sig eller censurerats tidigare finns ej med.

Assess och schoenfeld används återigen för att utvärdera de båda tidsperioderna. I nedan tabell visas resultaten.

	Schoenfelds residualer		**Assess supremum test**	
	Linjär korrelation med tiden (p-värde)		**Max. absolut värde (p-värde)**	
	Första tidsperioden	**Andra tidsperioden**	**Första tidsperioden**	**Andra tidsperioden**
Mannens ålder	0.031 (0.16)	-0.012 (0.60)	0.945 (0.44)	0.814 (0.56)
Ålderskillnad	0.019 (0.40)	0.032 (0.18)	0.954 (0.51)	1.508 (0.06)
Ursprung	0.019 (0.38)	-0.006 (0.81)	0.742 (0.60)	0.639 (0.77)
Utbildning	-0.002 (0.93)	-0.004 (0.88)	0.723 (0.71)	0.776 (0.61)

Varken Schoenfelds residualer eller Assess uppvisar tecken på att mannens ålder och utbildning för någon av tidsperioderna inte är proportionella. Åldersskillnad för den andra tidsperioden är den enda kovariaten där nollhypotesen kan förkastas (det vill säga är ej proportionell) med en signifikansnivå på 10-procent. Assess testet i SAS illustreras även med grafer där de empiriska

residualerna visas med 20 simuleringar. Exempel på grafer för den andra tidsperioden visas i bilaga 13.

På grund av det potentiella problemet med åldersskillnad har en separat körning gjorts på andra tidsperioden där modellen stratifierats med avseende på åldersskillnad. Resultatet av denna körning presenteras i bilaga 11. När kovariaternas estimat jämförs med modellen utan stratifiering framkommer att båda modellerna har mycket liknande skattningar. Av denna anledningen beslutar vi oss för att modellen inte behöver anpassas efter olika strata för åldersskillnaden.

5.5.1. Resultat med modifierad Cox-modell

För att råda bot på proportionalitetsproblemet delades tidsperioden in i två intervall. Där den första tidsperioden avser de första 2008 dagarna och den andra perioden dag 2009 till 4017.

En tolkning bör baseras på båda tidsperioderna var för sig då vi vet att proportionalitetsantagandena är uppfyllt för dessa. Tolkningen för första tidsperioden blir enkel då enda skillnaden mot hela tidsperioden är att vi endast undersöker en studieperiod som motsvarar fem och ett halvt år istället för elva. Tolkningen för den andra tidsperioden är dock något mindre intuitiv, då denna modellen undersöker risken för skilsmässa för individer som redan varit gifta i högst fem och ett halvt år.

I nedan tabell presenteras parametervärdena och värdena för hasard ration (riskkvoten mellan en kategori och jämförelsekategorin) för modellen som avser hela tidsperioden, första tidsperioden och den andra tidsperioden. Till skillnad mot de parametriska modellerna har parameterestimaten här motsatt tecken. Förklaringen härleds till att Cox-modellen estimerar logaritmen av hasarden istället för logaritmen av tiden till en händelse som de parametriska metoderna.

Vid jämförelse av kategorierna under kovariaterna visar det sig att de allra flesta parameterestimaten är liknande oavsett hur tiden delas in. De kategorierna under varje variabel som uppvisar signifikanta estimat ändrar inte tecken, det vill säga hasarad ration är antingen över ett eller under ett oavsett vilken tidsperiod som avses. Detta innebär att resultaten är relativt konsistenta för båda tidsperioder, vilka i sin tur är konsistenta med hela tidsperioden. Vissa estimat blir dock inte signifikanta (på 5 % nivå) för båda tidsperioderna, till exempel uppvisar den näst yngsta ålderskategorin för mannens ålder ett värde som inte är signifikant för den andra tidsperioden. I stort uppvisar dock resultaten liknande värden, vilket kan komma sig av att Cox modellen allmänt är robust eller att icke-proportionaliteten som visades för hela tidsperioden inte var något betydande problem för denna studie.

	Hela tidsperioden		Första tidsperioden		Andra tidsperioden	
	Parameter estimat	**Hasard ratio**	**Parameter estimat**	**Hasard ratio**	**Parameter estimat**	**Hasard ratio**
Mannens ålder >39 år	-0.92**	0.40	-1.20**	0.30	-0.56**	0.57
30-39 år	-0.45**	0.64	-0.62**	0.54	-0.22**	0.80
27-29 år	-0.34**	0.72	-0.46**	0.63	-0.14	0.87
<27 år **(Jmfr.-kat.)**	-	1	-	1	-	1
Åldersskillnad M>5år	0.38**	1.46	0.32**	1.38	0.46**	1.59
M>-3år till M<6år	-0.06	0.94	-0.12	0.89	0.02	1.02
M<-2år **(Jmfr.-kat.)**	-	1	-	1	-	1
Ursprung[7] 1	-0.04	0.96	0.04	1.04	-0.17	0.84
2	0.23**	1.26	0.20**	1.23	0.26**	1.29
3	0.34**	1.41	0.30**	1.36	0.39**	1.48
4	0.29**	1.34	0.24	1.28	0.35*	1.43
5	0.32**	1.37	0.26*	1.29	0.40**	1.49
6 **(Jmfr.-kat.)**	-	1	-	1	-	1
Utbildning[8] 1	-0.79**	0.45	-1.06**	0.35	-0.33**	0.72
2	-1.47**	0.23	-1.68**	0.19	-1.08**	0.34
3	-0.43**	0.65	-0.59**	0.55	-0.09	0.92
4	-0.61**	0.54	-0.69**	0.50	-0.40*	0.67
5	-1.22**	0.30	-1.41**	0.25	-0.84**	0.43
6	-0.76**	0.47	-0.87**	0.42	-0.51**	0.60
7 **(Jmfr.-kat.)**	-	1	-	1	-	1

**** signifikansnivå 0.01 * signifikansnivå 0.05**

Genom likelihood ratio statistikan kan förklaringasagraden, R^2, för hela modellen beräknas. Förklaingsagraden uppvisar ett värde på knappa fem procent, vilket innebär att det sannolikt finns många fler variabler som skulle behöva inkluderas i modellen.

För att på ett pedagogiskt vis framställa resultaten kommer tolkningen att främst avse modellen för hela tidsperioden. Av säkerhetsskäl tolkar vi de kategorier som inte har signifikanta värden för båda tidsperioderna med mindre tillit.

[7] Ursprung: 1, födda utrikes 2, båda födda i Sve. För en individ är minst en förälder född utrikes 3, född i Sve. med föräldrar födda i Sve. resp. född utrikes 4, båda födda i Sve. båda har minst en förälder född utrikes 5, född i Sve. med minst en utrikes förälder resp. född utomlands 6, födda i Sve. med föräldrar födda i Sve.

[8] Utbildning: 1, båda gymnasial 2, båda eftergymnasial 3, förgymnasial resp. gymnasial 4, eftergymnasial resp. förgymnasial 5, gymnasial resp. eftergymnasial 6, uppgifter saknas på minst en 7, båda förgymnasial.

Liksom för de parametriska metoderna som använts uppvisar mannens ålder vid giftermålet med Cox regressionen ett samband med risken för skilsmässa. Desto äldre mannen är desto lägre risk har paret för skilsmässa. Risken är endast cirka 40 procent för de par där mannen är äldre än 39 år jämfört med de par där mannen är yngre än 27 år. Vi kan inte med säkerhet säga att par där mannen är mellan 27 till 29 år skiljer sig från jämförelsegruppen sett till hela tidsperioden, då dessa estimat inte är signifikanta för den andra tidsperioden.

För de par där kvinnan är mer än fem år yngre än mannen vid giftermålet löper en betydligt mycket större risk att skilja sig än då kvinnan är minst tre år äldre än mannen.

Parameterestimaten för ursprung tyder på att de par där båda parter är födda utrikes har minst risk för skilsmässa under de första elva åren av äktenskapet. Denna kategori är dock inte signifikant skiljd från jämförelsekategorin, vilket indikerar att dessa grupper har en liknande risk över hela tidsperioden. En intressant iaktagelse är att estimaten uppvisar ett negativt värde för första perioden men ett positivt värde för den andra. Obeaktat signifikansnivån indikerar detta att de utrikesfödda har en lägre risk än jämförelsegruppen under de första fem och ett halvt åren för att sedan i högre utsträckning riskera skilsmässa den andra perioden. Risken för skilsmässa för de par där ena parten är andra generationens invandrare och den andra är född utrikes är cirka 30 procent högre än då båda är födda i Sverige med föräldrar födda i Sverige.

Parameterestimaten för utbildningsnivån med Cox uppvisar i hög grad samma mönster som med de parametriska metoderna. Jämförelsekategorin där båda har förgymnasial utbildningen löper högst risk relativt övriga kategorier för skilsmässa. De par där båda har gymnasial utbildning löper endast ungefär hälften stor risk som jämförelsegruppen.

En alternativ indelning för ursprung har även använts, denna indelning tar hänsyn till om det är mannen eller kvinnan som är andra generationens invandrare eller är född utrikes. Resultaten för denna indelning uppvisar inte några stora skillnader. En något förhöjd risk kan dock urskiljas då kvinnan är född i Sverige och mannen född utomlands jämfört med när motsatt förhållande råder. Estimeringen med Cox visas i bilaga 14.

6. Analys och diskussion

Hasardfunktionen med lifetable respektive Kaplan-Meier upvisar i båda fall att risknivån för skilsmässa når en högsta punkt tre år efter ingånget äktenskap i Sverige, detta om man bortser från den substantiella ökningen av risknivån i Kaplan-Meier efter elfte året. Det uppvisade mönstret förefaller inte enbart gälla i Sverige utan även Clarke-Stewart och Brentano (2007) har observerat att studier beträffande risknivån för skilsmässor i USA når en högsta punkt runt två och ett halvt år.

Sju-årskris är ett återkommande begrepp när man studerar litteratur och undersökningar som rör äktenskapsskillnader. I vår undersökning observerar vi en påtaglig ökning av risken för skilsmässa omkring sju till nio år efter giftermål vilket exempelvis har likheter med en studie av

United States Census Bureau[9] som uppvisar en hög risknivå sju år efter ingånget äktenskap i USA.

Risken för skilsmässa i början av äktenskapet visar på en låg risknivå. Man bör dock vara medveten om att en skilsmässa inte sker över en dag utan att det kan handla om en utdragen process som varar i månader eller år innan äktenskapsskillnad är fullbordad (Clarke-Stewart och Brentano 2006).

En av målsättningarna för denna studie var att finna en lämplig modell för datamaterialet. De modeller som presenetrats är diverse parametriska, Cox proportional regression och piecewise exponential. För att framgångsrikt använda en parametrisk modell behöver fördelningen av datamaterialet specificeras. De parametriska modellerna kontrollerades mot generaliserade gamma med likelihood ratio test, med log-likelihood samt visuellt. Ingen av dessa modeller kan sägas passa materialet perfekt, men log-normal vilken är ett specialfall av generaliserade gamma är den mest lämpliga. Det kan tänkas att log-normal (och generaliserade gamma) modellen inte fullt ut lyckas skatta den fullständiga hasardfunktionen, då denna tycks ha två maximipunkter enligt estimering med life-table och Kaplan-Meier. För att ta beakta båda maximipunkterna behövs således en mer anpassningsbar modell.

Det var därför intressant att jämföra de parametriska modellerna med piecewise exponential modellen. Modellen bygger på att datamaterialet delas upp i disjunkta intervall för vilka det kan antas råda konstant hasard (eller en exponentialfördelning). Utifrån utseendet på hasardfunktionen tycks antagandet vara någorlunda väl uppfyllt efter cirka tre år. Modellen lyckas dock inte fånga upp den linjärt ökande trenden som uppvisas för de första åren. På det hela taget gav de parametriska modellerna och piecewise exponential ett mycket liknande resultat.

Den tredje modellen som användes var den semiparametriska modellen: Cox proportional hasard. Modellen ger något mindre effektiva skattningar än de parametriska men har fördelen att hasardfunktionens fördelning inte behöver specificeras. Modellen vilar dock på antagandet om en proportionell hasard, vilket testades med Schoenfeld residualer, Assess och den kumulativa hasardfunktionen. Samtliga av dessa tester visade tydliga tecken på att både utbildningsnivå och mannens ålder inte uppfyllde förutsättningarna för modellen. Två modeller för två separata tidsperioder visade sig dock kunna råda bot på problematiken. Skattningarna för de separata tidsperioderna visade sig vara lika skattningen för hela tidsperioden. Visserligen blev somliga kategorier inte signifikanta för båda tidsperioderna till exempel mannens ålder 27-29 år. Dessa kategorier bidrar sannolikt till proportionalitetsproblemet, då det innebär att dessa grupper och jämförelsegrupperna inte har en statistiskt skiljd hasard under en tidsperiod. Utifrån skattningarna tycks inte proportionaliteten vara ett särdeles stort bekymmer, då parameterestimaten inte byter tecken mellan tidsperioderna för kovariaterna som är signifikanta för hela tidsperioden. Av denna anledning bör parameteresimaten för hela tidsperioden kunna tolkas trots att antagandet inte är uppfyllt.

[9] myndighet i USA som ansvarar för folkräkningen

Som helhet visar R^2-värdet att modellen endast förklarar en liten del av risken för skilsmässa. Även då detta mått har ifrågasatts vid överlevnadsanalys är det ändå troligt att detta sociala beteende är betydligt mycket mer komplext än vad våra kovariater kan förklara.

Det är svårt att avgöra vilken modell som är mest lämplig för materialet. Alla undersökta modeller visar mycket liknande skattningar och samma kategorier är insignifikanta oavsett modell. De mest lämpliga parametriska modellerna vilka utgörs av log-normal och generaliserade gamma har uppenbart en passform som inte är optimal. Av denna anledning faller vår rekommendation på Cox.

Alla kovariater i modellen visades påverka risken för skilsmässa. Åldern på mannen vid giftermålet följer ett tydligt mönster där jämförelsegruppen det vill säga de yngsta löper högst risk och ju äldre mannen är desto lägre risk. Zhang och Hook (2009) finner samma mönster för kvinnans ålder i deras studie av amerikanska par. South och Lloyd (1995) menar att åldersvariabeln vid giftermålet är en av de viktigaste för att förklara risken för skilsmässa. Detta då yngre personer mer sannolikt än äldre fortsätter att leta partner även då de är gifta.

Kovariaten för ålderskillnad visar att då mannen är mer än fem år äldre än kvinnan är risken som högst för paret. Risken då paret är någorlunda jämngamla uppvisar approximativt samma värden som då kvinnan är mer än två år äldre. Vårt resultat går stick i stäv med vad Zhang och Hook som visar att den högsta risken i USA innehas för gruppen då kvinnan är mer än två år äldre än mannen och den lägsta risken då paret är jämngamla. Vi kan bara spekulera i varför det finns en skillnad mellan länderna. Kanske är det vanligare i Sverige att kvinnan är några år äldre än mannen och denna kategori inte kan sägas beskriva en udda konstellation, det vill säga de skiljer sig inte särdeles mycket från normen. Både vår och Zhang och Hook undersökning visar dock att då mannen är väsentligt mycket äldre än kvinnan har paret en förhöjd risk för separation.

Clarke-Stewart och Brentano (2007) pekar även på att studier visar att ju mindre skillnader mellan individerna i paret med avseende på ålder och utbildningsnivå desto mer troligt att äktenskapet håller. Det är rimligt att anta att större likheter i individernas roller i äktenskapet förbättrar relationerna mellan makar och ger en mer gemensam grund samt förståelse av framgångar och utmaningar.

Individer i ett par som har olika utbildningsnivå uppges ha en förhöjd risk för skilsmässa. Anledningen anges vara att de har en potentiell källa till konflikter, olika värderingar samt obalans i maktfördelning; vilket kan förväntas förhöja instabiliteten i äktenskapet (Sweeney och Phillips 2004). Våra resultat tyder inte på att de par med störst skillnad i utbildningsnivå har högst risk, utan att de par där båda har en låg utbildningsgrad är mest utsatta. Liksom för mannens ålder vid giftermålet tyder resultaten för utbildningsnivå på att ju högre utbildning desto mindre risk för skilsmässa. Det är dock svårt att veta om det är just utbildningsnivån eller om det kan vara andra korrelerade variabler till utbildning som skapar denna markanta skillnad. Det är till exempel troligt att högre utbildning leder till högre inkomst vilket sannolikt bidrar till mindre konflikter i paret angående ekonomin. För att sortera ut effekten av utbildning skulle således även inkomst inkluderats i modellen.

Kovariaten som representerar ursprung visar att de par där båda är födda i Sverige och har föräldrar födda i Sverige samt de par där båda är födda utrikes har minst risk för skilsmässa. Dessa par kan sägas utgöra ytterligheterna av de konstellationer som undersökts. De representerar även de par som har mest gemensamt sett till härkomst. En möjlig förklaring är att det i dessa paren finns mindre kulturella skillnader än i övriga par, vilket leder till mindre missförstånd och mer stöd från släkt och vänner (Zhang och Hook 2009). Enligt denna teori skulle således par med stora kulturella skillnader leda till högre skilsmässorisk. I vår kategorisering utgörs dessa av par där ena individen är född utomlands och den andra i Sverige med föräldrar födda i Sverige. I enlighet med teorin uppvisar dessa par högre risk än andra kategoriseringar, men skillnaden är dock inte speciellt stor. Förutom kulturella skillnader nämner Zhang och Hook att den förhöjda risken även kan bero på rasism då paret utsätts för ytterligare en påfrestning. De menar att den grupp som har högst risk att separera, nämligen en svart gift med en vit, också utsätts för mest påtryckningar i form av social isolering.

I vår undersökning tas även hänsyn till om individerna är andra generationens invandrare, något som vi inte funnit i tidigare forskning. Dessa par har högre risk för skilsmässa än de par där de ingående individernas föräldrar är födda i Sverige. Något förvånande är att par där ena parten är andra generationens invandrare och den andra har föräldrar födda i Sverige uppvisar en lägre risk än då båda är andra generationens invandrare. Skillnaden är visserligen inte speciellt stor och det ska även tilläggas att vi inte vet var de som är andra generationens invandrare härstammar från, det är inte osannolikt att många av dessa har ursprung i olika länder.

På det hela taget skiljer sig resultaten mot våra förväntningar innan undersökningen. Utan några djupare kunskaper i ämnet trodde vi att personer med föräldrar födda i Sverige skulle vara mer benägna att ta ut skilsmässa. Förväntningarna grundade sig på att Sverige har relativt höga skilsmässotal samt att Sverige är relativt många andra länder mer sekulariserat och familjen kan väntas ha mindre stabila familjeförhållanden. Mehrdad Darvishpour (2004) anger att den främsta förklaringen till den ökade skilsmässorisken hos invandrare beror på att kvinnor anpassas snabbare än män till ett jämställt samhälle och att männen är ovana vid maktförskjutningen där kvinnor är mindre beroende av männen. Separationen blir en naturlig följd av kvinnors frigörelse i det nya hemlandet. De svenska männen har hunnit anpassa sig till maktförskjutningen under många decennier. Han påpekar även att det är stor skillnad mellan var paren härstammar från, utomeuropeiska förväntas ha högst risk för skilsmässa.

6.1 Källkritik samt förslag på framtida studier

Den relativt höga skilsmässorisken för andra generationens invandrare och de par där ena individen är utrikesfödd kan mycket väl bero på fler variabler än de som inkluderats i våra modeller. Vår förklaringsgrad är låg och det är sannolikt att en betydligt mycket större modell skulle behövas för att bättre förklara risken för skilsmässa. Till exempel kan det tänkas att socioekonomiska faktorer påverkar, såsom arbetslöshet och inkomst.

Vilket framgår av våra resultat finner vi ingen parametrisk modell som lämpar sig perfekt för datamaterialet. En modell som använts är piecewise exponential modellen, denna antar dock att de intervall som tiden delats in i har en konstant hasard. Vad som beskrivits tidigare är detta ett

antagande som inte uppfylls fullt ut. Det skulle därför vara intressant att se en modell som tillåter linjär hasard i intervallen. Detta kan uppnås med en piecewise log-linear modell.

I de modeller som vi använder oss av är kovariaterna oberoende av tiden. En mer avancerad modell tar hänsyn till att kovariaterna kan förändras med tiden. Speciellt för utbildningsnivå skulle det vara lämpligt att ta hänsyn till förändringar under studietiden. Detta skulle leda till en mer trovärdig indelning för högskolestudier, då det skulle vara möjligt att beakta en färdig utbildning mot endast påbörjade studier.

Det skulle vara intressant att använda fler år för när giftermål ägde rum och på så vis få ett större datamaterial i syfte att kunna studera skillnader mellan människor från till exempel olika världsdelar, det vill säga inte enbart utrikes födda.

Litteraturförteckning

Böcker:

Aalen, Odd O. Borgan, Ørnulf. Gjessing,Håkon K. 2008. *Survival and event history analysis : a process point of view*. New York: Springer.

Allison, Paul D. 1995. *Survival Analysis Using SAS®: A Practical Guide*. Cary, NC: SAS Institute Inc.

Allison, Paul D. 2010. *Survival Analysis Using SAS®: A Practical Guide*,Second Edition. Cary, NC: SAS Institute Inc.

Clarke-Stewart, Alison and Cornelia Brentano. 2006. *Divorce: Causes and Consequences.* New Haven, CT: Yale University Press.

Collett, David. 2003. *Modelling Survival Data in Medical Research*, Second Edition, London: Chapman & Hall.

Darvishpour, M. 2004. *Invandrarkvinnor som bryter mönstret*, Lund: Liber.

Hosmer, David W. Lameshow, Stanley.1999. *Applied survival analysis: regression modeling of time to event data*. New York: Wiley.

Kleinbaum, David G. Klein, Mitchel. 2005. *Survival analysis: a self-learning text*, Second Edition. New York: Springer.

Artiklar:

Cox, D.R and Snell, E.J. 1968. A general definition of residuals.
Journal of the Royal Statistical Society. 30, 248-275.

Cox, D.R. (1972). Regression Models and Life-tables. *Journal of the Royal Statistical Society,* 34, 187-220.

Farrington C.P. (2000). Residuals for Proportional Hazards Models with
Interval-Censored Survival Data. *Biometrics,* 56, 473-482.

Kay, R. (1977). Proportional Hazard Regression Models and the Analysis of Censored Survival Data. *Journal of the Royal Statistical Society*, 26, 227-237.

Magee, L. (1990). R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests. *American Statistical Association,* 44, 250-253.

Schemper, M. (1992). Cox Analysis of Survival Data with Non-Proportional Hazard Functions, *Journal of the Royal Statistical Society*, 41, 455-465.

Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model, *Biometrika*, 69, 239-241.

South, S.J. and Lloyd, K.M. (1995). Spousal Alternatives and Marital Dissolution, *American Sociological Review*, 60, 21-35.

Sweeney, M.M. and Phillips, J.A. (2004). Understanding Racial Differences in Marital Disruption: Recent Trends and Explanations, *Journal of Marriage and Family*, 66, 639-650.

Zhang, Y. and Van Hook, J. (2009). Marital Dissolution Among Interracial Couples, *Journal of Marriage and Family*, 71, 95-107.

Rapporter och presentationer:

Box-Steffensmeier, J.P and Zorn, C.J.W. (1998). *Duration Models and Proportional Hazards in Political Science, paper for presentation at the annual meeting of the Midwest political Science Association*. Version 1.9. Chicago Ohio State University.

Statistiska centralbyrån prognosinstitutet. (2010). Födda i Sverige – ändå olika? Betydelsen av föräldrarnas födelseland, *Demografiska rapporter 2010:2*. Stockholm, SCB.

Statistiska centralbyrån prognosinstitutet. (1995). Skilsmässor och separationer: bakgrund och utveckling, *Demografiska rapporter 1995:1*. Stockholm, SCB.

Internet-källor:

United nationens statistics division. (2006). *Demographic Yearbook 2006.* Hämtat 3 december 2010 från <http://unstats.un.org>.

Engineered Software. *Weibull distribution.* Hämtat 23 november 2010 från <http://www.engineeredsoftware.com/nasa/weibull.htm>.

United States Census Bureau. (2005). *Number, Timing, and Duration of Marriages and Divorces: 2001*. Household Economic Studies. Hämtat 20 december 2010 från <http://www.census.gov/prod/2005pubs/p70-97.pdf>.

Bilagor

Bilaga 1

Cox-snell residualer

En residual vi härigenom behandlar ytligt är Cox-snell residualer som introducerades av David Cox och E. Joyce Snell för att utvärdera validiteten av en överlevnadsfunktion som avser en mängd överlevnadsdata. Cox-snell residualer används alltjämt ofta i analyser av överlevnadsdata. Cox-snell residualen för *i*:te individen, *i*=1,2,3….,*n*, är given av följande funktion

$r_{ci} = \exp(\hat{\beta}` x_i)(\hat{H}_0(t_i)$

där $\hat{H}_0(t_i)$ är ett estimat av baseline kumulativa hasard funktionen i tiden t_i som är

överlevnadstiden för *i*:te individen(Collett, D. 112). Cox-Snell residualen r_{ci} är även värdet av $\hat{H}_i(t_i) = -\log\hat{S}_i(t_i)$ där $\hat{H}_i(t_i)$ och $\hat{S}_i(t_i)$ är skattade värden på kumulativa hasarden och

överlevnadsfunktionen för *i*:te individen vid tidpunkt t_i (Collett, D. 112).

I fall en modells lämplighet med hänseende till datamaterialet är tillfredställande så kommer estimat som grundar sig på modellen kunna uppvisa en överlevnadsfunktion för *i*:te individen vid tidpunkt t_i som ligger skapligt nära det motsvarande sanna värdet $S_i(t_i)$.

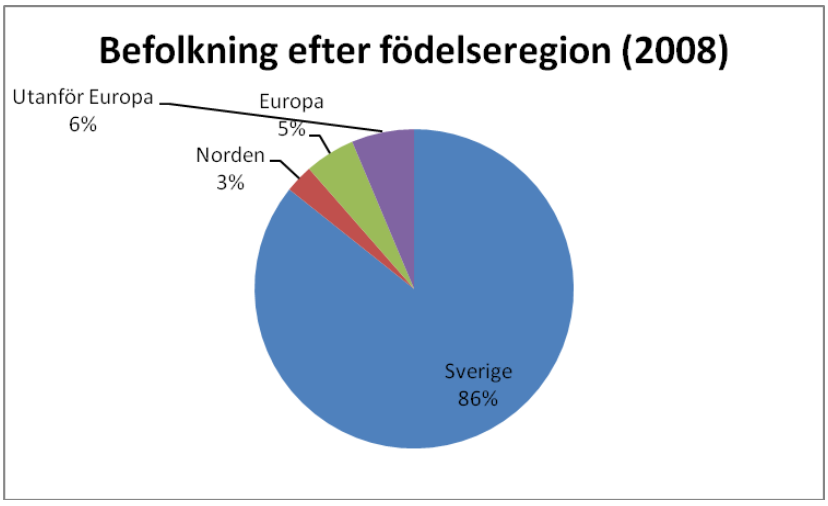
Detta antyder att om en lämplig modell passar datamaterialet så innebär detta att skattade värden $\hat{S}_i(t_i)$ har liknande egenskaper som det sanna värdet $S_i(t_i)$. Detta resulterar i att Cox-Snell

residualerna bör följa en exponentialfördelning.

I fall individers överlevnadstid är högercensurerade så är även motsvarande värden på dess residualer högercensurerade vilket innebär att residualerna icke kan läggas samman med residualer som härrör från ocensurerade observationer. På grund därav så är det möjligt att använda sig av modifierad Cox-Snell residual som fäster avseende vid censurering.

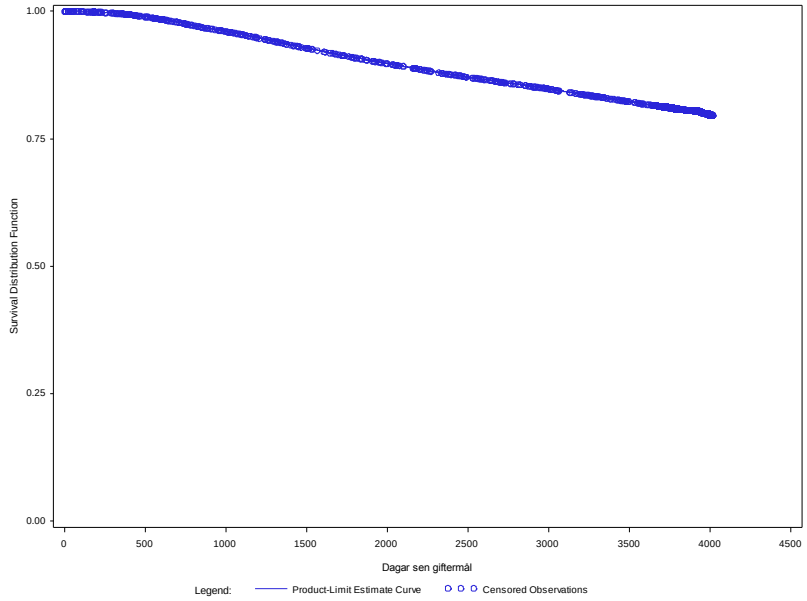
Bilaga 2

Cirkeldiagrammet nedan beskriver i vilken region Sveriges befolkning var födda år 2008.



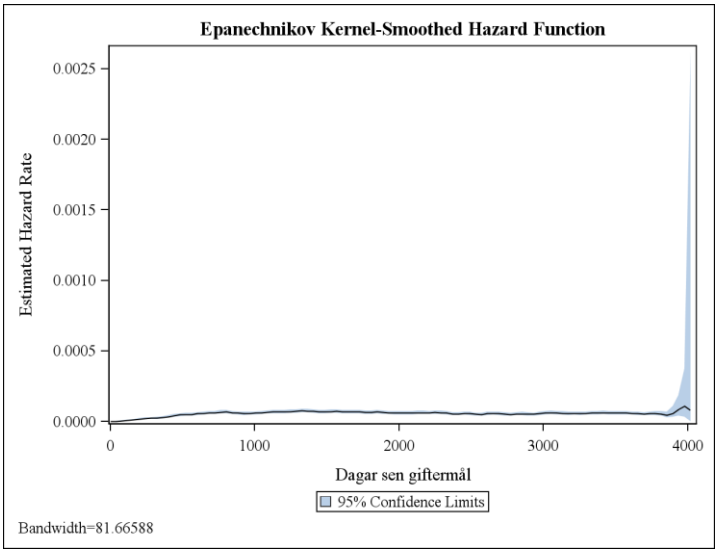
Bilaga 3

Figuren beskriver Kaplan-Meier estimeringen av överlevnadskurvan. Grafen skiljer sig inte nämnvärt från estimering med life-table metoden.



Bilaga 4

Nedan beskrivs den utjämnade hasardfunktion med Kapla-Meier. Figuren inkluderar ett 95-procentigt konfidensband, av vilket framgår osäkerheten i slutet av den undersökta perioden.



Bilaga 5

I tabellen life-table estimaten vilka ger en översiktlig bild över datamaterialet. Observera att tabellen fortsätter på nästa sida.

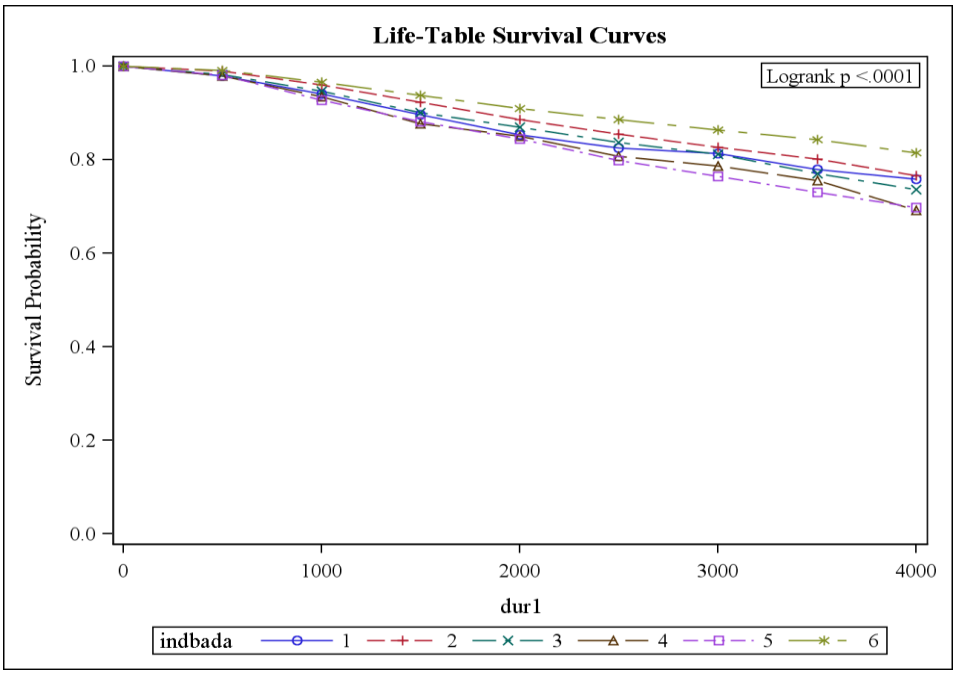
Life Table Survival Estimates									
Interval		Number Failed (skilsmässor)	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival (procent som inte skilt sig)	Failure (procent som skiljer sig)	Survival Standard Error
[Lower,	Upper)								
0	**365**	102	109	20368.5	0.00501	0.000495	1.0000	0	0
365	**730**	362	69	20177.5	0.0179	0.000934	0.9950	0.00501	0.000495
730	**1095**	448	48	19757.0	0.0227	0.00106	0.9771	0.0229	0.00105
1095	**1460**	490	44	19263.0	0.0254	0.00113	0.9550	0.0450	0.00146
1460	**1825**	460	28	18737.0	0.0246	0.00113	0.9307	0.0693	0.00179
1825	**2190**	416	35	18245.5	0.0228	0.00111	0.9078	0.0922	0.00204
2190	**2555**	372	38	17793.0	0.0209	0.00107	0.8871	0.1129	0.00223
2555	**2920**	340	45	17379.5	0.0196	0.00105	0.8686	0.1314	0.00238
2920	**3285**	362	45	16994.5	0.0213	0.00111	0.8516	0.1484	0.00250

Life Table Survival Estimates									
Interval		Number Failed (skilsmässor)	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival (procent som inte skilt sig)	Failure (procent som skiljer sig)	Survival Standard Error
[Lower,	Upper)								
3285	3650	360	52	16584.0	0.0217	0.00113	0.8335	0.1665	0.00263
3650	4015	154	15957	8219.5	0.0187	0.00150	0.8154	0.1846	0.00274
4015	.	0	87	43.5	0	0	0.8001	0.1999	0.00295

Interval		Evaluated at the Midpoint of the Interval			
[Lower,	Upper)	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	365	0.000014	1.355E-6	0.000014	1.362E-6
365	730	0.000049	2.547E-6	0.00005	2.607E-6
730	1095	0.000061	2.836E-6	0.000063	2.969E-6
1095	1460	0.000067	2.97E-6	0.000071	3.189E-6
1460	1825	0.000063	2.885E-6	0.000068	3.175E-6
1825	2190	0.000057	2.751E-6	0.000063	3.098E-6
2190	2555	0.000051	2.61E-6	0.000058	3.001E-6
2555	2920	0.000047	2.503E-6	0.000054	2.935E-6
2920	3285	0.000050	2.588E-6	0.000059	3.1E-6
3285	3650	0.000050	2.589E-6	0.00006	3.169E-6
3650	4015	0.000042	3.344E-6	0.000052	4.175E-6
4015	.	.	.	.	.

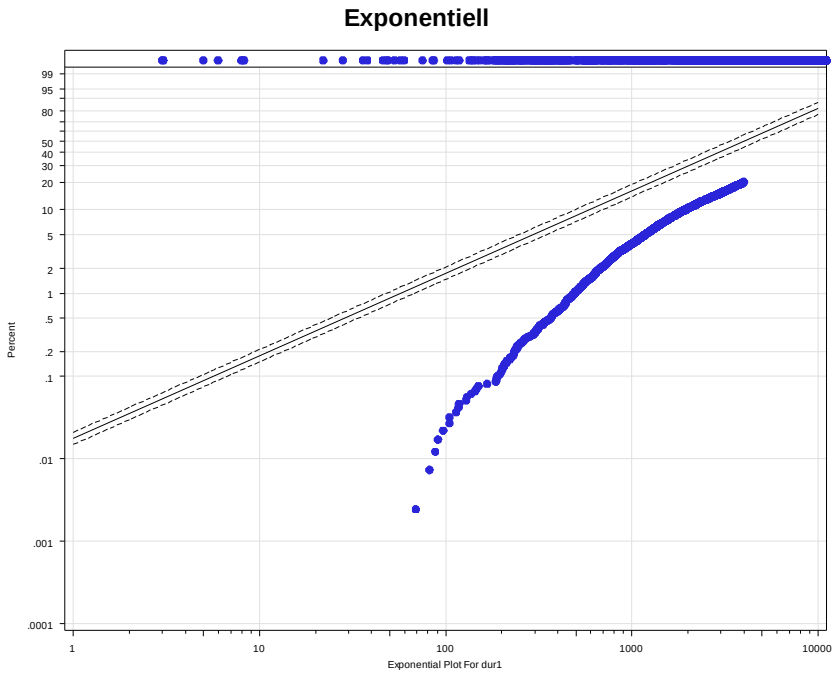
Bilaga 6

Grafisk illustration av överlevnadskurvorna för var och en av kategorierna i ursprungsvariabeln.



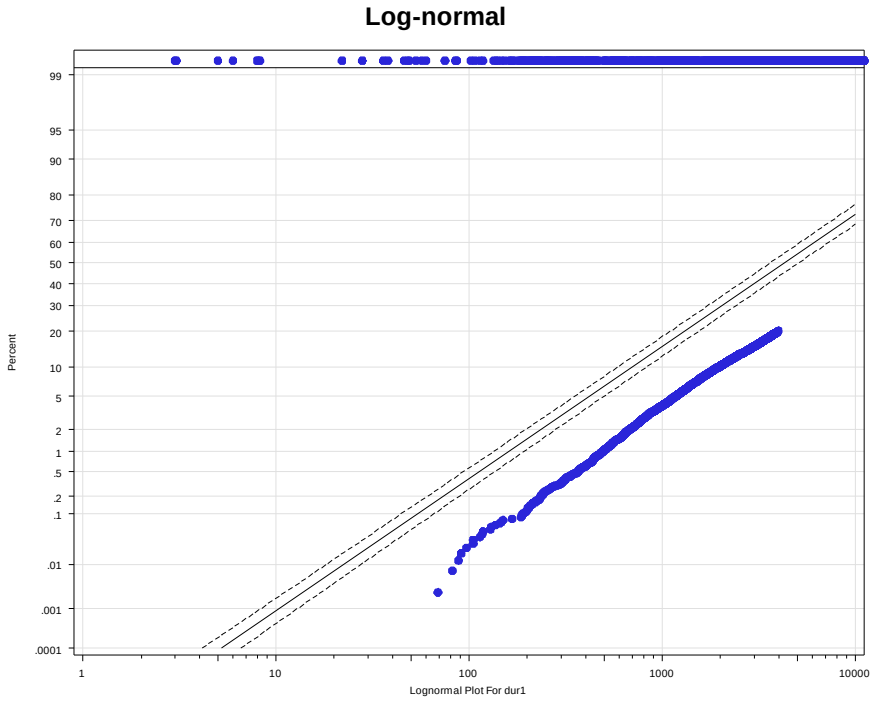
Bilaga 7

Grafisk utvärdering av exponentialfördelningen.



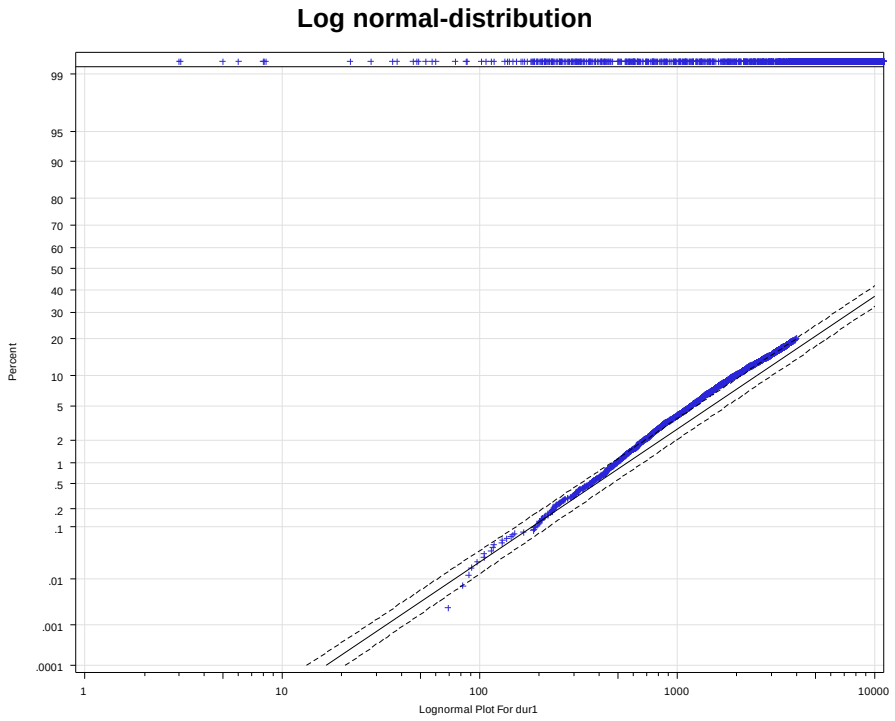
Bilaga 8

Grafisk utvärdering av log-normal fördelningen för den fullständiga modellen.



Bilaga 9

Grafisk utvärdering av log-normal fördelningen för modell utan kovariaten för utbildning.



Bilaga 10

Piecewise exponential modell

Hasardfunktionen (figur 4 och 5) är ganska komplex, då det finns få perioder som har tydliga trender den uppvisar även en antydan till två maximipunkter. De parametriska modellerna kan inte hantera denna fördelning fullt ut. Utifrån nämnda figurer tycks inte heller piecewise exponential modellen vara optimal. Detta då materialet skall delas upp i tidsintervall för vilka det kan antas råda en konstant hasard. För de tre första åren är dock hasarden oavbrutet ökande, men för resten av tidsperioden kan antagandet approximativt uppfyllas. Vid användning av modellen delades tidsperioden upp i 22 intervall där varje intervall representerar cirka sex månader. För varje par och intervall då paret inte skilt sig eller censurerats skapas en observation. För ett par som finns med i studien under hela studieperioden skapas således 22 observationer. Och för ett par som skiljer sig under första sex månaderna skapas en observation. Totalt använde modellen sig av 392 312 observationer. Tiden börjar mätas från starten av varje intervall.

Wald-statistikan för ”intervallvariabeln” uppvisade ett signifikant värde vilket antyder att hasarden inte är konstant över tid, således är intervallindelningen befogad.

Estimeringen av modellens koefficienter uppvisar inga överraskningar jämfört med generaliserade gamma och log-normal (se nedan tabell). Tabellen visar även skattningar för varje intervall vilka indikeras med j och jämförs med de sista sex månaderna för studietiden. Denna intervallvariabel visar att samtliga överlevnadstider var signifikant skiljda från sista

tidsintervallet. Endast de första sex månaderna visar på längre överlevnadstider än jämförelsekategorin.

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept		1	12.9149	0.2454	12.4339	13.3958	2770.12	<.0001
>39 år	1	1	0.9519	0.0751	0.8046	1.0991	160.49	<.0001
30-39 år	2	1	0.4746	0.0438	0.3888	0.5604	117.55	<.0001
27-29 år	3	1	0.3503	0.0485	0.2552	0.4455	52.13	<.0001
<27 år (Jmfr.-kat.)	4	0	0.0000	.	.	.	.	.
Åldersskillnad M>5år	1	1	-0.3912	0.0694	-0.5274	-0.2551	31.74	<.0001
M>-3år till M<6år	2	1	0.0617	0.0594	-0.0548	0.1782	1.08	0.2991
M<-2år (Jmfr.-kat.)	3	0	0.0000	.	.	.	.	.
Ursprung 1	1	1	0.0469	0.0825	-0.1149	0.2086	0.32	0.5701
2	2	1	-0.2365	0.0424	-0.3197	-0.1533	31.06	<.0001
3	3	1	-0.3569	0.0552	-0.4651	-0.2487	41.79	<.0001
4	4	1	-0.2952	0.1070	-0.5048	-0.0855	7.62	0.0058
5	5	1	-0.3267	0.0888	-0.5009	-0.1526	13.52	0.0002
6	6	0	0.0000	.	.	.	.	.
Utbildning 1	1	1	0.8478	0.0706	0.7095	0.9861	144.33	<.0001
2	2	1	1.5410	0.0784	1.3874	1.6946	386.52	<.0001
3	3	1	0.4612	0.0735	0.3171	0.6053	39.35	<.0001
4	4	1	0.6584	0.1049	0.4529	0.8639	39.43	<.0001
5	5	1	1.2826	0.0757	1.1343	1.4310	287.19	<.0001
6	6	1	0.7960	0.1115	0.5774	1.0145	50.95	<.0001
7	7	0	0.0000	.	.	.	.	.
0-183 Intervall	1	1	1.0665	0.3339	0.4122	1.7209	10.21	0.0014
183-365	2	1	-0.5548	0.2535	-1.0517	-0.0580	4.79	0.0286
365-548	3	1	-1.1574	0.2429	-1.6335	-0.6813	22.70	<.0001
548-730	4	1	-1.4210	0.2399	-1.8911	-0.9509	35.10	<.0001

Analysis of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
730-913	5	1	-1.5440	0.2387	-2.0119	-1.0761	41.84	<.0001
913-1096	6	1	-1.5036	0.2391	-1.9723	-1.0349	39.54	<.0001
1096-1278	7	1	-1.6077	0.2382	-2.0746	-1.1408	45.55	<.0001
1278-1461	8	1	-1.6345	0.2380	-2.1010	-1.1679	47.14	<.0001
1461-1643	9	1	-1.6048	0.2384	-2.0720	-1.1376	45.33	<.0001
1643-1826	10	1	-1.5344	0.2391	-2.0030	-1.0658	41.19	<.0001
1826-2008	11	1	-1.5521	0.2390	-2.0205	-1.0837	42.18	<.0001
2008-2191	12	1	-1.4115	0.2405	-1.8829	-0.9401	34.45	<.0001
2191-2374	13	1	-1.4704	0.2400	-1.9408	-1.0001	37.55	<.0001
2374-2556	14	1	-1.2961	0.2420	-1.7705	-0.8217	28.67	<.0001
2556-2739	15	1	-1.3453	0.2416	-1.8188	-0.8719	31.02	<.0001
2739-2921	16	1	-1.3083	0.2421	-1.7828	-0.8337	29.20	<.0001
2921-3104	17	1	-1.3935	0.2412	-1.8663	-0.9207	33.37	<.0001
3104-3287	18	1	-1.4218	0.2410	-1.8942	-0.9494	34.79	<.0001
3287-3469	19	1	-1.4175	0.2412	-1.8903	-0.9447	34.53	<.0001
3469-3652	20	1	-1.4258	0.2413	-1.8987	-0.9529	34.92	<.0001
3652-3834	21	1	-1.1430	0.2453	-1.6237	-0.6623	21.72	<.0001
3834-4017	22	0	0.0000	.	.	.	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		
Weibull Shape		0	1.0000	0.0000	1.0000	1.0000		

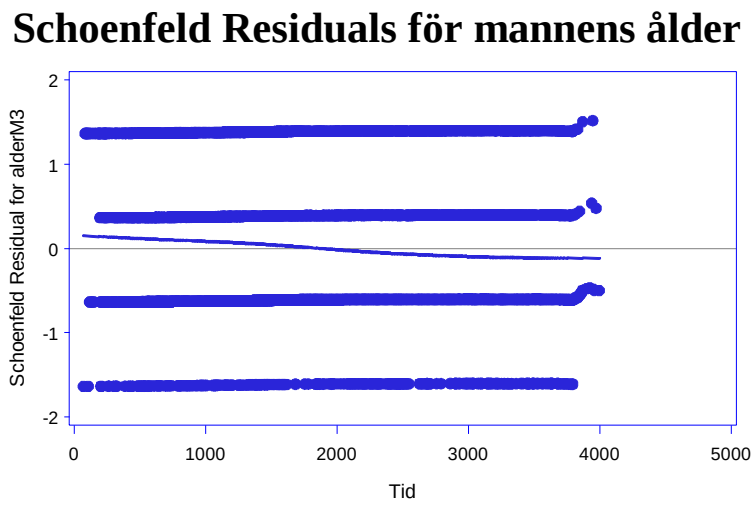
Bilaga 11

Nedan presenteras en tabell för resultaten då modellen för den andra tidsperioden stratifieras med avseende på kovariaten för åldersskillnad. Estimaten är i stort mycket lika de då stratifiering ej tillämpades.

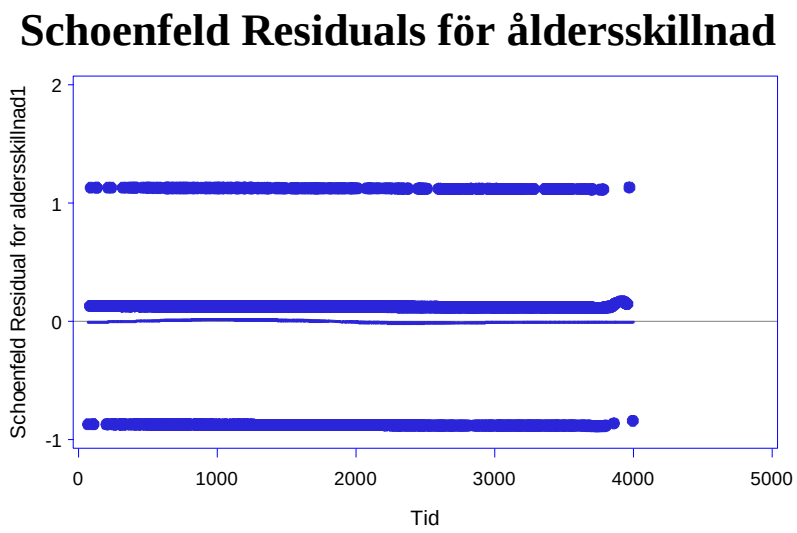
Analysis of Maximum Likelihood Estimates						
Parameter	**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
Ursprung 1	1	-0.17216	0.13624	1.5970	0.2063	0.842
2	1	0.25604	0.06201	17.0459	<.0001	1.292
3	1	0.39045	0.08145	22.9814	<.0001	1.478
4	1	0.35558	0.15989	4.9455	0.0262	1.427
5	1	0.40212	0.13267	9.1863	0.0024	1.495
>39 år	1	-0.55634	0.10806	26.5074	<.0001	0.573
30-39 år	1	-0.21935	0.06848	10.2594	0.0014	0.803
27-29 år	1	-0.14248	0.07543	3.5682	0.0589	0.867
Åldersskillnad	0	0	.	.	.	.
Utbildning 1	1	-0.33360	0.12713	6.8858	0.0087	0.716
2	1	-1.08416	0.13604	63.5068	<.0001	0.338
3	1	-0.08973	0.13201	0.4620	0.4967	0.914
4	1	-0.40252	0.17802	5.1124	0.0238	0.669
5	1	-0.84449	0.13333	40.1188	<.0001	0.430
6	1	-0.51587	0.18928	7.4278	0.0064	0.597

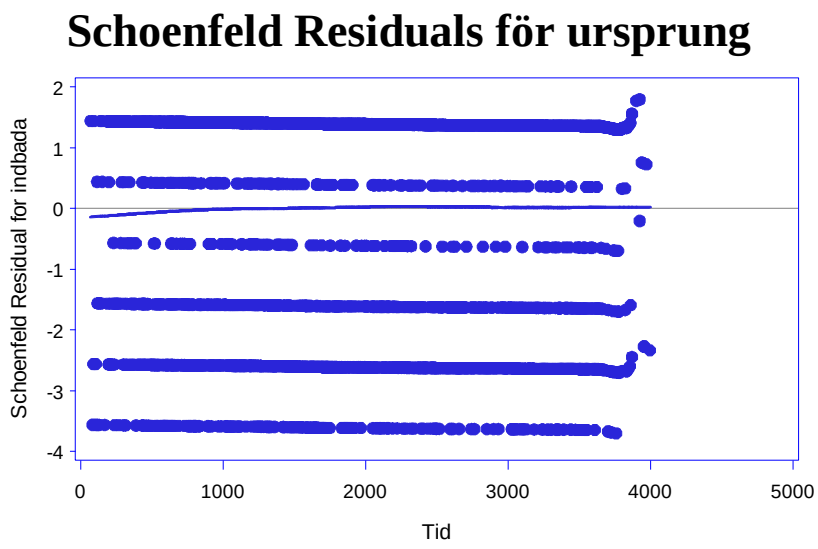
Bilaga 12

I nedan figur ser man tydligt hur linjen för Schoenfeld residualerna inte kretsar kring noll utan följer en negativ trend som innebär att kovariaten för männens ålder inte uppfyller proportionalitetsantagandet.



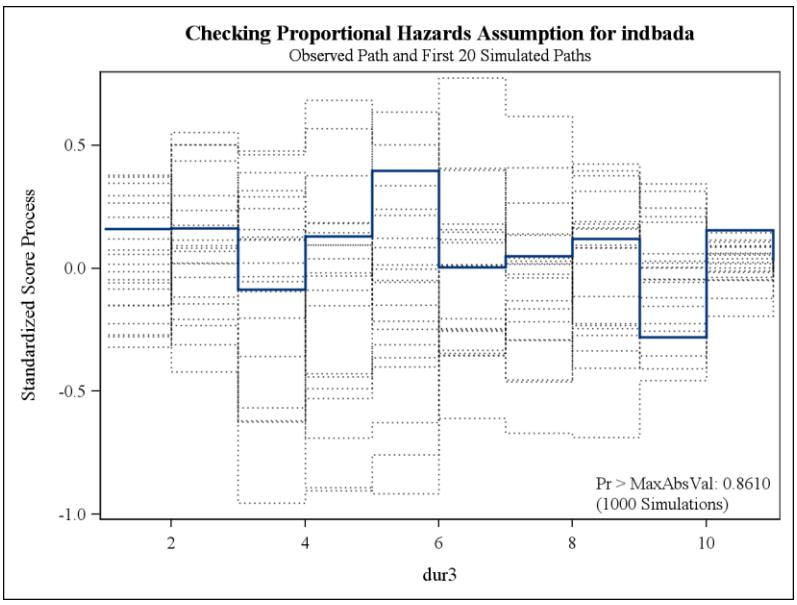
I enlighet med korrelationsmatrisen med Schoenfeld residualerna för åldersskillnad och ursprung ligger residualerna nära noll och antyder att proportionalitet för kovariaten är fullvärdig.





Bilaga 13

Observera att grafen som presenteras nedan är för andra tidsperioden. Vi har även delat upp tiden i 6-månadersintervall. Detta eftersom SAS inte förmår producera motsvarande graf när dagar används, möjligen på grund av storleken på datamaterialet som behandlas.



I Assess output produceras grafen ovan för tidsperiod två för kovariaten ursprung och våra observationer representeras av den heldragna blå linjen. De streckade linjerna är baserade på 20 slumpmässiga simulerade processer som omfattar antagandet om proportionalitet.

Om den heldragna blå linjen för våra observationer avviker substantiellt från de simulerade processerna så ger detta ett stöd för att antagandet om proportionalitet inte är uppfyllt.

I grafen förefaller inte den observerade processen för ursprung extrem i förhållande till de simulerade processerna.

Om p-värdet, som du finner längst ned till höger i grafen, skulle vara mycket lågt tyder detta på att antagandet om proportionalitet inte är tillfredsställt. P-värdet produceras genom ett tusen simulationer som även producerar en tabell som summerar resultat för varje kovariat. P-värdet 0.861 uppvisar att av 1000 simulationer så har dessa 86.1 procent extrema värden som överstiger de mest extrema värdena bland våra observationer. För övriga kovariater uppvisades en liknande grafisk framställning som för ursprung.

Bilaga 14

Resultat med indelning efter mannens respektive kvinnans härkomst

I nedan tabell presenteras estimering då hänsyn tagits till vilket ursprung mannnen och kvinnan har. Tidigare har vi inte beaktat vem av mannen eller kvinnan som härstammar varifrån. Enligt Zhang och Hook har, sett till etnicitet, par där mannen är svart och kvinnan vit högst risk för skilsmässa. Särskilt intressant att jämföra i tabellen är parkategorierna: 1 och 2, 3 och 4 samt 7 och 8, då varje par av dessa kategorier beror på om det är mannen respektive kvinnan som är född i Sverige. Det lägre talet i varje par av kategori representerar att mannen är född i Sverige och det högre talet att kvinnan är född i Sverige. Den fullständiga indelningen av kategorier återfinns efter tabellen med skattningar.

Resultatet visar små skillnader beroende på om det är kvinnan eller mannen som har utrikes härkomst. Skattningarna visar dock konsekvent att de par där kvinnan är född i Sverige och mannen är född utrikes har en högre risk för skilsmässa än då manen är född i Sverige och kvinnan utrikes. Störst skillnad uppvisas för de kategorier där antingen mannen eller kvinnan är andra generationens invandrare och deras partner är född utrikes. Jämfört med jämförelsegruppen det vill säga där både kvinnan och mannen samt deras föräldrar är födda i Sverige har dessa en förhöjd risk på 35 respektive 40 procent.

Analysis of Maximum Likelihood Estimates							
Parameter		**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**
Mannens ålder 1	**>39 år**	1	-0.91309	0.07513	147.7077	<.0001	0.401
2	**30-39 år**	1	-0.45204	0.04386	106.2358	<.0001	0.636
3	**27-29 år**	1	-0.33312	0.04859	47.0104	<.0001	0.717
Åldersskillnad 1	**M>5år**	1	0.37912	0.06947	29.7812	<.0001	1.461
2	**M>-3år till M<6år**	1	-0.05754	0.05944	0.9369	0.3331	0.944
Ursprung	**1**	1	0.21831	0.05620	15.0907	0.0001	1.244
	2	1	0.23953	0.05631	18.0941	<.0001	1.271
	3	1	0.34023	0.07113	22.8782	<.0001	1.405
	4	1	0.34747	0.07749	20.1069	<.0001	1.415
	5	1	0.06476	0.21932	0.0872	0.7678	1.067
	6	1	0.37249	0.12138	9.4180	0.0021	1.451
	7	1	0.30163	0.11479	6.9044	0.0086	1.352
	8	1	0.34067	0.13335	6.5270	0.0106	1.406
	9	1	-0.04115	0.08251	0.2488	0.6179	0.960
utbildning	**1**	1	-0.79271	0.07058	126.1372	<.0001	0.453
	2	1	-1.47074	0.07844	351.5844	<.0001	0.230
	3	1	-0.42516	0.07353	33.4322	<.0001	0.654
	4	1	-0.61505	0.10492	34.3623	<.0001	0.541
	5	1	-1.21636	0.07574	257.9279	<.0001	0.296
	6	1	-0.76254	0.11181	46.5131	<.0001	0.466

1. Mannen född i Sverige med föräldrar födda i Sverige. Kvinnan född i Sverige minst en förälder född utrikes.

2. Kvinnan född i Sverige med föräldrar födda i Sverige. Mannen född i Sverige minst en förälder född utrikes.

3. Mannen född i Sverige med föräldrar födda i Sverige. Kvinnan född utrikes.

4. Kvinnan född i Sverige med föräldrar födda i Sverige. Mannen född utrikes.

5. Mannen född i Sverige med minst en förälder född utrikes. Kvinnan född i Sverige minst en förälder född utrikes, men en mindre eller mer än mannen (dvs. när mannen har båda föräldrar utrikesfödda har kvinnan en förälder utrikesfödd och vice versa).

6. Mannen och kvinnan är födda i Sverige och deras föräldrar är antingen båda födda utrikes eller enbart en av dem (dvs. mannen och kvinnan i paret har lika många föräldrar som är utrikesfödda.

7. Mannen är född i Sverige och har en eller två föräldrar födda utrikes. Kvinnan är född utrikes.

8. Kvinnan är född i Sverige och har en eller två föräldrar födda utrikes. Mannen är född utrikes.

9. Mannen och kvinnan är födda utrikes.

10. (jämförelsekategori) Mannen och kvinnan är födda i Sverige med föräldrar födda i Sverige.