CHALMERS

GÖTEBORGS UNIVERSITET

Förskjutning av döden kring födelsedagen

Kan viljestyrka skänka ett förlängt liv?

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Kandidatarbete inom civilingenjörsutbildningen vid Chalmers

Natalia Andreeva

Furqan Farooqi

Ingrid Ingemarsson

Lucas Lazaroo

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2019

Förskjutning av döden kring födelsedagen

Kan viljestyrka skänka ett förlängt liv?

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet
Natalia Andreeva

Kandidatarbete i matematik inom civilingenjörsprogrammet Kemiteknik vid Chalmers
Furqan Farooqi

Kandidatarbete i matematik inom civilingenjörsprogrammet Teknisk fysik vid Chalmers
Ingrid Ingemarsson

Kandidatarbete i matematik inom civilingenjörsprogrammet Maskinteknik vid Chalmers
Lucas Lazaroo

Handledare: Sergei Zuyev
Examinatorer: Maria Roginskaya och Ulla Dinger

Institutionen för Matematiska vetenskaper
CHALMERS TEKNISKA HÖGSKOLA
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2019

Kan döden vänta in födelsedagen?

Idéen om att människor har den inombordliga kraften att skjuta upp sin död (givet naturliga orsaker) för att erfara en sista meningsfull händelse har länge cirkulerat på ett anekdotiskt plan. Det har tidigare utförts ett fåtal studier kring fenomenet på händelser så som Pesach och Månfesten i Kina med tvetydiga resultat. Detta fenomen har på engelska kallats ”postponement hypothesis”, vilket hädanefter kommer översättas till förskjutningshypotesen.

Att vidare statistiskt verifiera huruvida det ligger någon sanning bakom denna frigörelse av livsenergi kan potentiellt ge upphov till medicinsk forskning kring hypotesen. Ifall det finns en eller flera mekanismer som i någon mån håller dessa människor vid liv, kan en kartläggning av dessa möjligen ge upphov till behandlingar för andra människor med livshotande skador. Detta är dock i nuläget ett högst hypotetiskt scenario.

I denna studie har det undersökts huruvida det finns något statistiskt belägg för dennna förskjutningshypotes, med människors födelsedagar. Om hypotesen stämmer borde mortaliteten (dödsrisken) vara lägre en period innan födelsedagen och högre en period efteråt för dödsfall av naturliga skäl. Med hjälp av överlevnadsanalys - ett samlingsbegrepp för statistiska verktyg när tidsberoende händelser studeras - har en metod utvecklats för att analysera om fenomenet förekommer runt födelsedagen. Det har innefattat framtagandet av modeller som beskriver när människor dör och i vilken takt.

Först inhämtades data från olika populationer, med information om när människor avlidit. Den slutgiltiga gruppen av människor som studerades utgjordes av dödsstatistiken i Sydafrika år 2015 för människor mellan 50 - 80 år som gick bort av naturliga dödsorsaker. Utifrån de inhämtade dataseten har modeller konstruerats som beskriver hur mortaliteten borde se ut om ingen förskjutningsverkan existerar. Därefter har de framtagna modellernas värden jämförts med de faktiska dödsfallen runt födelsedagarna för de studerade data, det studerade intervallet omfattade 14 dagar före bemärkelsedagen samt samma period efteråt. Om en skillnad mellan dessa återfinns är det en indikator på att dödsrisken är annorlunda runt födelsedagen.

Analys av befolkningen gav ett resultat som pekade på att förskjutningshypotesen inte stämmer. Istället pekar utfallet på att människor dör i större utsträckning runt födelsedagen överlag, både innan och efter bemärkelsedagen. Detta innebär alltså att människor vid en viss ålder verkar löpa större risk att avlida inom en 4 veckors period runt sin födelsedag. Möjliga förklaringar till detta kan vara att händelsen medför både fysiska och psykiska påfrestningar för åldrande individer som ökar risken för att död.

Om en förskjutningseffekten faktiskt existerar är det för diffust för att uttyda i fallet med födelsedagar för de undersökta datamängderna. Det utesluter inte möjligheten att fenomenet kan manifisteras i andra händelser som exempelvis högtider av stor kulturell eller religiös betydelse, vilket inte studerats i detta arbete.

Sammanfattning

I denna uppsats undersöks en hypotes som säger att människor kan förskjuta sin nalkande död (givet naturliga orsaker). Fenomenet har fått namnet ”Postponement hypothesis” och antar att en betydelsefull händelse kan agera motivation för att förlänga livet under en kort tid. Vi antar att en persons födelsedag är en sådan meningsfull händelse och analyserar hypotesen omkring detta datum med hjälp av verktyg från en statistisk gren kallad överlevnadsanalys. Om hypotesen är sann kan vi förvänta oss att mortaliteten borde vara lägre under en period innan en persons födelsedag, och kanske högre under en period efter. Vi väljer att begränsa denna period till 14 dagar. Detta skulle indikera att det finns en kraft som förskjuter döden för den givna populationen. Dataseten som använts i analysen är mortalitetsdata över personer som levt till att bli superhundraåringar och italienare som blivit äldre än 105 år, samt primärt ett dataset över alla som dog i Sydafrika under år 2015. Mortaliteten sammanställs i så kallade hazardfunktioner, som vid varje given ålder uttryckt i dagar beskriver dödshastigheten för människor som överlevt till denna givna ålder. Därefter appliceras diverse parametriska modeller på hazardfunktionerna, i syfte att upptäcka avvikelser kring födelsedagarna. För detta är ett t-test utfört på medelvärdet av residualernas mortalitet under födelsedagsperioden, för att se om medelvärdet är skillt från 0. Detta borde vara fallet om dagarna kring födelsedagarna inte skiljer sig i termer av mortalitet från andra dagar under året. Resultatet av vår analys visar att ingen förskjutningseffekt kan uttydas för de aktuella dataseten. Däremot kan den omvända effekten observeras, vilket indikerar att mortaliteten egentligen är högre både före och efter födelsedagen. Spekulationer kring varför detta är fallet skulle kunna vara att människor löper högre risk att dö kring sin födelsedag på grund av stress relaterat till förberedeler.

Abstract

In this thesis we adress a hypothesis that suggests that people can postpone their imminent death (given natural causes). The so-called "Postponement hypothesis" assumes that a meaningful occasion can act as a motivator to prolong life for a short amount of time. We consider a persons birthday as that meaningful occasion and analyze the hypothesis around this date by using tools from a statistical discipline known as Survival analysis. If the hypothesis is true it can be expected that the mortality rate should be lower a period before a person's birthday and, perhaps, higher shortly afterwards. We choose to set this period to a limit of 14 days. This would indicate that there is a force which postpones death for the population concerned. The datasets used in analysis are mortality data over people who lived to be Supercentenarian and Italian people who became older than 105 years, and also primarily a dataset for South African people who died in the year 2015. The mortality rate is summarised by hazard functions, which at each age expressed in days describes the dying rate of people who survived to this day. We thereafter apply various parametric models to the hazard, in order to discover any discrepancy around the birthdays. For this a t-test is conducted on the mean of the residuals mortality in the birthday period, to see if the mean is non-zero. This should be the case if the days around the birthdays are no different in terms of mortality rate compared to other days of the year. The results of our analysis show that no postponement of death can be seen for the examined dataset. Instead the data suggest that the mortality rate is actually higher both before and after the birthday. Speculations as to why this is the case might be a higher risk associated with the stress of preparing for the birthday.

Innehåll

1 Inledning **1**
- 1.1 Syfte . . . 1
- 1.2 Tidigare studier . . . 1
- 1.3 Metod . . . 2
- 1.4 Avgränsningar och antaganden . . . 2
- 1.5 Datakällor . . . 2
- 1.6 Populationsöversikt . . . 3

2 Överlevnadsanalys **5**
- 2.1 Grundläggande definitioner . . . 5
- 2.2 Icke-parametrisk estimering av hazard- och överlevnadsfunktion . . . 8
- 2.3 Parametrisk estimering av hazardfunktion . . . 9

3 Analys och hypotestest **10**
- 3.1 Utförande av hazardfunktion . . . 10
- 3.2 Bootstrap-metoden . . . 13
- 3.3 Konfidensintervall . . . 13
- 3.4 Regressionsmodeller . . . 14
- 3.5 Residualanalys . . . 16

4 Diskussion **18**
- 4.1 Arbetets utfall . . . 18
- 4.2 Resultat jämfört med tidigare studier och framtida rekommendationer . . . 19

A Ytterligare tabeller för analys **23**

B Kod **24**
- B.1 Histogram . . . 24
- B.2 Periodiskt histogram . . . 24
- B.3 Grafer . . . 25
- B.4 Hazardfunktion . . . 26
- B.5 Integrerad hazarftunktion . . . 27
- B.6 Bootstrap och konfidensintervall . . . 27
  - B.6.1 Bootstrap . . . 27
  - B.6.2 Vektorbaserad inverse hazard . . . 28
  - B.6.3 Plot för bootstrappad konfidensintervall . . . 29
- B.7 Regressionsmodeller . . . 29
  - B.7.1 Regression med linjär modell . . . 29
  - B.7.2 Regression med generaliserad linjär modell . . . 31
  - B.7.3 Inverse hazard med upp av två-årsintervall . . . 33

Förord

Vi vill börja med att tacka vår handledare Sergei Zuev för hans vägledning och enorma engagemang i arbetet.

På grund av arbetet och gruppens storlek har projektets olika moment utförts gemensamt av samtliga medlemmar i gruppen. Olika personer har axlat ytterligare ansvar för diverse områden när det förefallit naturligt, detta ämnar att återspeglas i tabell 1 men det understryks att hela gruppen har samverkat för att uppnå projektets syfte med en delad insats av problemlösning, genomförande och revidering. En individuell loggbok och gemensam dagbok har kontinuerligt förts under arbetets gång.

Avsnitt	Delavsnitt	Natalia	Furqan	Ingrid	Lucas
Kan döden vänta in födelsedagen?			×		×
Sammanfattning/Abstract					×
Inledning					
	Tidigare studier	×			×
	Datakällor				×
	Populationsöversikt			×	×
Överlevnadsanalys					
	Grundläggande definitioner			×	×
	Icke-parametrisk ...	×		×	
	Parametrisk ...				×
Analys och hypotestest					
	Utförande av hazardfunktion			×	×
	Bootstrap-metoden		×		
	Konfidensintervall	×			
	Regressionsmodeller	×	×		
	Residualanalys	×	×		
Diskussion					
	Arbetets utfall	×			×
	Resultat jämfört med tidigare studier och framtida rekommendationer		×	×	
Appendix					
	Ytterligare tabeller för analys		×	×	
	Kod (infogning)			×	
Grafer och tabellutformning				×	

Tabell 1: Huvudansvariga för respektive kapitel, max två gruppmedlemmar har tilldelats varje delavsnitt. Dock har även övriga bidragit till dessa.

1 Inledning

Det finns en hypotes som säger att döende människor kan förskjuta sin död tills dess att en viktig händelse har skett [1]. Hypotesen menar att personer som ligger på sin dödsbädd kan lyckas hålla sig vid liv tills det att de fått ta del av en sista meningsfull händelse i sina liv, varpå de snabbt avlider. Detta har kommit att kallas "Postponement Hypothesis" eller "Death Postponement Hypothesis", hädanerfter benämnt "förskjutningshypotesen". Tidigare studier av effekten har utförts med olika resultat [2]. Vidare utröning av fenomenet är meriterande för att nå konsensus bakom hypotesens validitet och är därför av intresse.

1.1 Syfte

Arbetet ämnar att undersöka om det föreligger något statistiskt underlag för förskjutningshypotesen. Detta uppnås genom att analysera beteendet kring personers födelsedagar med hjälp av överlevnadsanalys och andra statistiska verktyg.

1.2 Tidigare studier

Det förekommer tidigare studier som har undersökt förskjutningseffekten och huruvida det finns något underlag för dess existens som gjorts med andra metoder än den som tillämpas i detta arbete. I detta delavsnitt redovisas några av dessa för att kontextualisera detta projekt med tidigare arbeten. En framträdande förespråkare för teorin, David P. Philips har utfört flera studier (med olika medförfattare vid varje enskilt arbete) för att undersöka detta fenomen. Många av de artiklar som diskuteras innefattar honom som författare vilket innebär att slutsatserna från de olika artiklarna konsekvent kommer från en enskild källa.

Phillips utförde en studie där den judiska påsken, Pesach, användes som betydelsefull händelse för att undersöka om denna förskjutningseffekt kan observeras [1]. Ett dataset av 1919 vuxna människor, som dog av naturliga orsaker under åren 1966-1984 användes som underlag för analys och jämfördes med en kontrollgrupp bestående av icke-judiska personer. Med regressionsanalys och binomialtest på dödsfrekvensen fick de p-värden som var signifikanta för gruppen av judiska människor. En nedgång ("dip") av antalet dödsfall observerades strax innan Pesach och efter högtiden observerades istället en uppgång ("peak"), medan kontrollgruppen ej visade samma mönster. Kritik har dock riktats mot arbetet angående säkerheten i den insamlade datamängden, Gary Smith påpekar att i urvalsprocessen valdes personer från Kalifornien som endast antogs vara judiska baserat på namnen [3]. En liknande studie av Philips undersökte dödligheten hos 1288 kineser som dog mellan åren 1960-1984 kring Månfesten, en traditionell högtid inom kinesisk kultur [4]. En linjär och kurvlinjär analys utfördes som uppvisade ett dylikt "dip-peak" mönster som var signifikant för kineserna och förekom inte hos en icke-kinesisk kontrollgrupp. Philips noterade att stickprovstorleken i både studierna av Pesach och Månfesten var små till storleken.

I ett senare arbete, benämnt 'The Birthday: Lifeline or Deadline?', anmärker Philips att det är problematiskt att generalisera resultaten från föregående studier för en hel befolkning eftersom dessa händelser var specifika för små undergrupper av populationen [5]. I denna studie användes ett större stickprov av 1309334 (år 1969-1977) och 1435815 (år 1978-1990) av personer (ålder 18+) som dog av naturliga orsaker med födelsedagar som betydelsefull händelse. Kontingens-tabeller och t-tester tillämpades och gav resultatet av en nedgång i dödsfall före födelsedagen och en uppgång efter (0-6 dagar efter) för det sammanslagna stickprovet. När stickproven analyserades separat observerades olika resultat för kvinnor och män. För kvinnor var dödsantalet signifikant fler efter födelsedagen, det omvända observerades för männen som dog i större omfattning innan deras födelsedag.

Den alternativa hypotesen påstår att det finns en "birthday effect", hädanefter kallad "födelsedagseffekt", vilket innebär att dödligheten ökar innan födelsedagen. Ajdacic-Gross et al. analyserar kopplingen mellan dödsdagen och födelsedagen med hjälp av schweiziska mortalitetsdatabaser från åren 1969-2008 [6]. Skillnaderna mellan dödsdagar och födelsedagar bildas som tidsserier. Au-

toregressivt integrerat flytande medelvärdesmodeller (ARIMA) används som tillämpligt verktyg. Stickprovstorleken var mer än 2 miljoner observationer av dödsfall. I övergripande tidsserier fanns det en 13,8-procentig ökning av dödsfall vid födelsedagar med variationer på mellan 11 och 18 procent hos män och kvinnor äldre än 60. I gruppen med dödsfall av naturliga orsaker var hjärtsjukdomar och cancer överrepresenterade som dödsorsak. Slutsatsen är att födelsedagar har fler dödsfall än förväntat främst för hjärtsjukdomar (infarkt och stroke) på grund av extra stress.

Ytterligare en studie utfördes av en annan forskare, H. Leerhoff, som genomförde en undersökning på mer än 4 miljoner dödsfall i Tyskland under perioden 1992-2011 [7]. Som meningsfull händelse användes både jul och födelsedag. Medianåldern var 72 år med den första kvartilen $Q1 = 63$, och den tredje kvartilen $Q3 = 81$. Varje ”dip-peak”-företeelse tilldelades ett index och antogs följa en binomialfördelning för att testa om förskjutningseffekten kunde observeras. Resultatet visade att ingen effekt hittades i födelsedagsgruppen, däremot återfanns en liten effekt för julgruppen med ett negativt ”dip-peak” index (fler dödsfall före julen).

1.3 Metod

För att avgöra om det finns en förskjutningseffekt kommer data analyseras för att kvantifiera mortalitet runt födelsedagen som sedan jämförs med resten av året. Ifall en försjuktningseffekt finns, borde ett ”dip-peak”-mönster uppträda runt födelsedagen. Alla kvantitativa beräkningar kommer genomföras i programspråket R.

1.4 Avgränsningar och antaganden

Som nämnt kommer arbetet endast undersöka om förskjutningeffekten kan verifieras med avseende på en persons födelsedag, därför görs antagandet att detta är en viktig händelse för de allra flesta människorna. Födelsedagen är en händelse som är gemensam för alla människor oavsett kulturell och religiös bakrund. Detta förenklar införskaffandet av data samt underlättar bearbetningen. Endast dödsorskar av naturliga skäl beaktas eftersom olycksrelaterade fall tillför brus utan att bidra med information om förskjutningshypotesen.

1.5 Datakällor

Arbetet har nyttjat tre olika datakällor som beskrivs nedan. Gemensamt för samtliga dataset är att de består av så kallad mikrodata, vilket avser information på individnivå. I arbetets fall innebär det att datamängderna innefattar information om ålder (mätt i år och dagar) för varje datapunkt. Detta är nödvändigt för att kunna undersöka förskjutningsfenomenet. Följande är de källor som har behandlats:

Data över superhundraåringar

Gentrology Research Group är en grupp bestående av läkare och vetenskapsmän med syfte att motverka åldersrelaterade nedgångar i hälsa [8]. Organisationen har data över alla bekräftade superhundraåringar som har använts i arbetet. Begreppet superhundraåringar definierar en person som har levt längre än 110 år, totalt innefattar datamängden 1015 datapunkter.

Italienskt dataset

Ett dataset innehållande ålder av italienska invånare som mellan 2009 och 2015 blev 105 år eller äldre. Källan används i studien *The plateau of human mortality: Demography of longevity pioneers* och tillhandagavs av proffessor Holger Rootzén [9]. Datamängden innehåller både levande och döda människor, således filtrerades data bort för att endast behandla de personer som avlidit, ur detta erhålls 2883 datapunkter.

Sydafrikanskt dataset

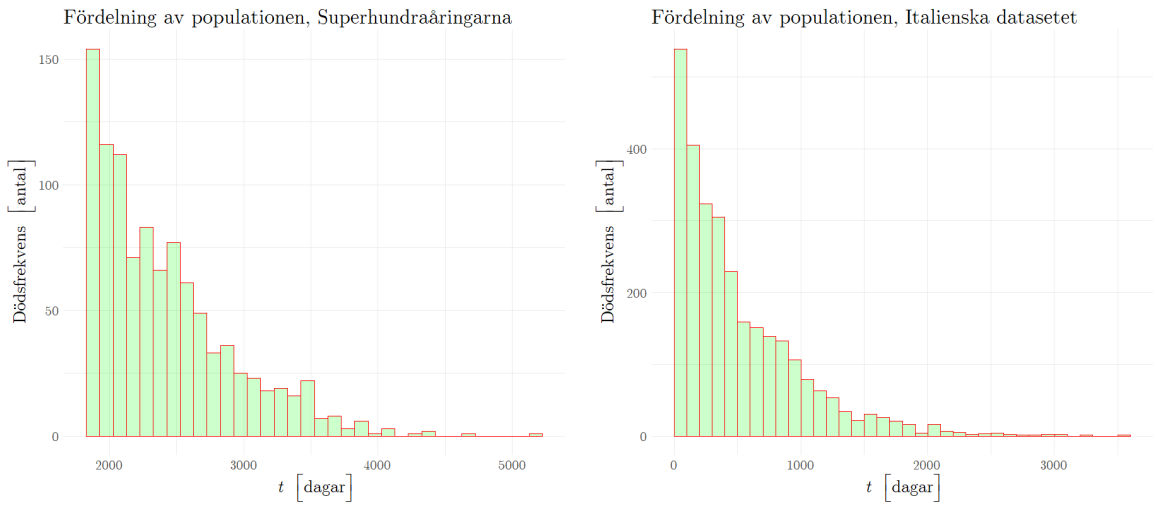
Det tredje datasetet behandlar dödsstatistiken i Sydafrika år 2015. Källan är DataFirst [10], en forskningsenhet på University of Cape Town med en uppdragsbeskrivning att erbjuda öppen tillgång av mikrodata i forskningssyfte [11]. Institutionen aggregerar data som samlas in av diverse organisationer verksamma i Sydafrika och andra afrikanska länder. Det specifika datasetet

som används i detta arbete har producerats av Sydafrikas inrikesministerium och statistiska byrå. Datamängden innehåller 460 236 datapunkter med 48 olika variabler som inte enbart avser parametrar relaterat till livslängd [12]. För att plocka fram relevant information filtreras data att endast inkludera dödsfall av naturliga orsaker för människor äldre än 50 år. Detta genererar ett dataset med 248 881 datapunkter. Begreppet naturliga dödsorsaker används enligt dataproducenternas definition som exkluderar dödsfall på grund av våld eller olyckor. Den filtrerade informationen är i jämförelse med övriga källor avsevärt större.

1.6 Populationsöversikt

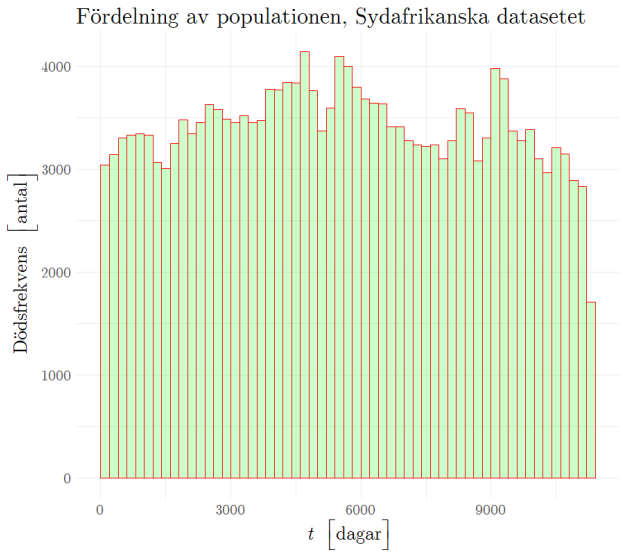
För att initiellt överblicka de befintliga dataseten presenteras deras frekvensfördelning i histogram, vilka delar in data i intervall i form av staplar med samma bredd. Om samtliga intervall är av samma storlek kan frekvensen för varje område utgöras av stapelns höjd, är detta inte fallet representeras frekvensen av stapelns area. Två olika histogram utförs för vardera dataset.

Det första histogramet (figur 1 och 2) visar i kronologisk ordning när människor över 105 års ålder har avlidit. Respektive histogram visas i figur 1(a) för superhundraåringarna och i figur 1(b) för det italienska datasetet. Den horisontella axeln återger antal dagar efter 105 år är fyllda för varje person medan den vertikala axeln visar dödsfrekvensen för respektive partition. Figur 2 visar histogram av samma slag för det Sydafrikanska datasetet, men den horisontella axeln återger antal dagar efter 50 års ålder.



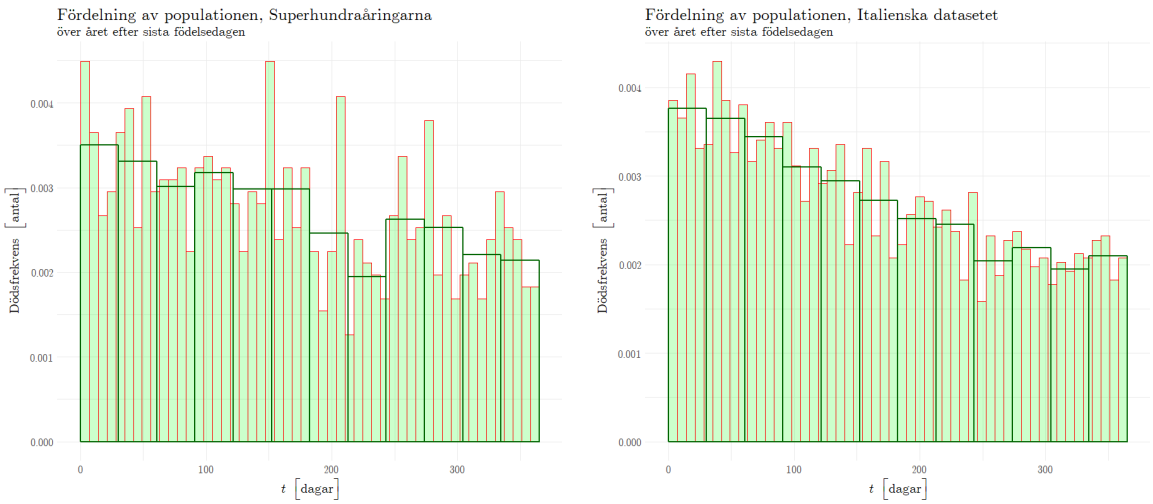
(a) (b)

Figur 1: Histogram för två dataset, den horisontella axeln visar antal dagar efter 105 år ålder och den vertikala speglar dödsantal för respektive partition. Delfigurer enligt följande: (a) Histogram för datasetet över superhundraåringarna; (b) Histogram för det italienska datasetet.



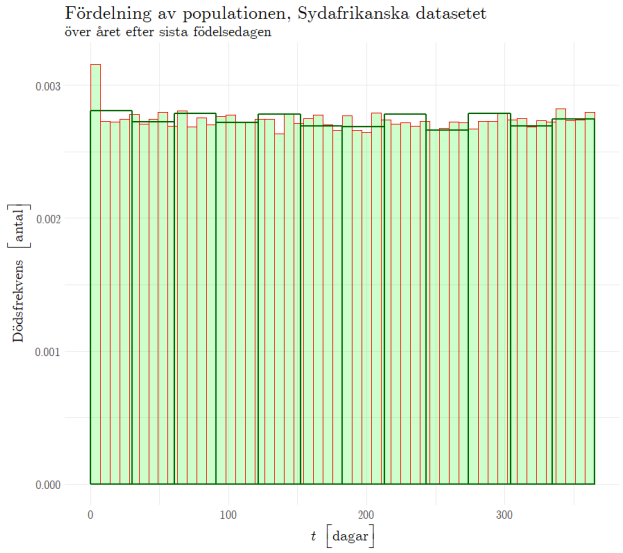
Figur 2: Histogram för det sydafrikanska datasetet, den horisontella axeln visar antal dagar efter 50 år ålder och den vertikala speglar dödsantal för respektive partition.

Det andra histogrammet (figur 3 och 4) utförs med avsikten att vi visuellt skall kunna undersöka om dödsfrekvensen avviker märkbart kring födelsedagen. Grafen konstrueras för att visa under vilken period på året människor avlider, mätt utifrån en persons födelsedag. Startdatum 0 är födelsedagen och dag 365 är dagen innan nästa födelsedag. Denna period representeras av den horisontella axeln, den verktikala återspeglar dödsfrekvensen och den kumulativa frekvesen är 1 eftersom de omfattar samtliga dödsfall. De mindre staplarna motsvarar ett tidsspann på en vecka och är 52 stycken till antalet. De bredare staplarna motsvarar månader. Samtliga staplar beräknas utifrån antagandet att ett år utgörs av 365 dagar och delas in i lika stora delar. Datasetet över superhundraåringar visas i figur 3(a), det italienska datasetet visas i figur 3(b) och det sydafrikanska i figur 4.



(a) (b)

Figur 3: Periodiska histogram för de två mindre dataseten. Den horisontella axeln visar antal dagar under året och den vertikala speglar dödsantal för respektive partition. Delfigur (a) korresponderar till datasetet över superhundraåringarna, medan (b) hör till det italienska datasetet.



Figur 4: Periodiskt histogram för det sydafrikanska datasetet, den horisontella axeln visar antal dagar under året och den vertikala speglar dödsantal för respektive partition.

De streckade delområderna i figur 3(a) och 3(b) framhäver en trend som visar att dödsfrekvenserna tycks vara lägst månaden innan födelsedagen och därefter som högst de efterkommande 30 dagarna. Vid första anblick kan detta misstas stödja förskjutningshypotesen för de tillhörande datamängderna. Detta är inte fallet, observera figur 1(a) och 1(b), i dessa framgår det att antalet överlevande människor minskar med tiden. När de periodiska histogrammen konstrueras bevaras denna trend från tidigare histogram som ger att mindre antal människor är vid liv vid slutet av året.

2 Överlevnadsanalys

Eftersom informationen som behandlas i arbetet består av tidsangivelser för döds- och födelsedag är överlevnadsanalys (*eng.* Survival analysis) en lämplig statistisk gren för ändamålet att undersöka förskjutningshypotesen. Detta kapitel innefattar en intoduktion till detta område, och hur det appliceras i arbetet.

Överlevnadsanalys är en samling av metoder för att analysera sannolikheten att en händelse av intresse ska ske vid en viss tid (*eng.* time-to-event) [13]. Detta koncept är starkt kopplat till de två nyckel-funktionerna inom överlevnadsanalys: överlevnadsfunktionen (*eng.* survival function) och hazardfunktionen (*eng.* hazad function). Avsnitt 2.1 behandlar definitioner och egenskaper hos dessa.

2.1 Grundläggande definitioner

Låt $\tau \geq 0$ beteckna en slumpvariabel för tiden från början av en undersökning fram till den intressanta händelsen. I vårt fall är τ alltså tiden från födelsedag till dödsdag för en given individ. Denna tid τ kallar vi *överlevnadstiden*. Låt också tiden t beteckna ett specifikt värde för τ. Då kan vi definiera *överlevnadsfunktionen* enligt definition 2.1.

Definition 2.1 (Överlevnadsfunktion)**.** Överlevnadsfunktionen $S(t)$, *eng.* Survival function, definieras enligt

$$S(t) = P(\tau > t), \tag{1}$$

det vill säga att den anger sannolikheten för att den studerade händelsen (här dödsfallet) ännu inte inträffat vid ett givet t [14].

Funktionen är kritisk för överlevnadsanalys därför att den ger information om hur överlevnadssannolikheten ter sig för ett givet dataset, vilket är själva kärnan av ämnet.

En nära besläktad funktion är den *kumulativa fördelningsfunktionen* vilken vi definierar enligt defnition 2.2.

Definition 2.2 (Kumulativa fördelningsfunktionen)**.** Kumulativa fördelningsfunktionen $F(t)$, definieras enligt [14]

$$F(t) = P(\tau \leq t). \tag{2}$$

Alltså gäller

$$S(t) = 1 - F(t). \tag{3}$$

I denna rapport kommer två fall; det diskreta och det absolutkontinuerliga, att behandlas. I det diskreta fallet delas tiden in i lämpligt tidsintervall, till exempel dagar, veckor eller månader. I det kontinuerliga fallet betraktas tiden som steglös och kan anta alla positiva värden. Den data som behandlas är diskret, vilket meriterar diskret analys, men eftersom vi ska behandla parametrisk estimering är även det kontinuerliga fallet intressant. Vidare vill vi definiera ytterligare en viktig funktion, och för den behöver vi definition 2.3.

Definition 2.3 (Täthetsfunktionen/Sannolikhetsfunktionen)**.** Täthetsfunktionen $f(t)$ definieras enligt

$$f(t) = \lim_{\Delta t \downarrow 0} \frac{P(t \leq \tau < t + \Delta t)}{\Delta t} \tag{4}$$

där t är en kontinuerlig slumpvariabel. I det kontinuerliga fallet gäller även [15]

$$F(t) = \int_{-\infty}^{t} f(u)du. \tag{5}$$

Detta innebär att $f(t)$ är lika med tidsderivatan av $F(t)$. Den diskreta motsvarigheten, sannolikhetsfunktionen $f(t)$ blir

$$f(t) = P(\tau = t) \tag{6}$$

där t är en diskret slumpvariabel.

Därmed har vi underlag nog för att definiera hazardfunktionen enligt 2.4. David G. Kleinbaum beskriver funktionen enligt följande ”Likt idén om hastighet ger hazardfunktionen den momentana potentialen vid en tidsenhet t att en händelse skall inträffa som exempelvis döden, förutsatt att individen överlevt fram tills tiden t” [13].

Definition 2.4 (Hazardfunktion)**.** Definiera hazardfunktionen $h(t)$ (*eng.* hazard function) i diskreta fallet enligt [13]

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t | \tau \geq t)}{\Delta t}. \tag{7}$$

Detta är i det kontinuerliga fallet ekvivalent med

$$h(t) = \frac{f(t)}{S(t)} \tag{8}$$

ty

Härledning.

$$h(t) := \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t | \tau \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(\tau \geq t | t \leq \tau < t + \Delta t) P(t \leq \tau < t + \Delta t)}{P(\tau \geq t)\Delta t} =$$
$$= \frac{1}{P(\tau \geq t)} \lim_{\Delta t \to 0} \frac{P(\tau \geq t | t \leq \tau < t + \Delta t) P(t \leq \tau < t + \Delta t)}{\Delta t},$$

men

$$P(\tau \geq t | t \leq \tau < t + \Delta t) = \frac{P(\tau \geq t \cap t \leq \tau < t + \Delta t)}{P(t \leq \tau < t + \Delta t)} = \frac{P(t \leq \tau < t + \Delta t)}{P(t \leq \tau < t + \Delta t)} = 1$$

så

$$h(t) = \frac{1}{P(\tau \geq t)} \lim_{\Delta t \to 0} \frac{P(t \leq \tau < t + \Delta t)}{\Delta t} = \frac{1}{S(t)} f(t).$$

□

I det diskreta fallet gäller också ekvation (8) med innebörden

$$h(t) = \frac{P(\tau = t)}{P(\tau > t)}. \tag{9}$$

Hazardfunktionen är i detta fall den precisa sannolikheten att dö vid en viss tid delat med sannolikheten att fortfarande vara vid liv vid den tiden.

I det absolutkontinuerliga fallet kan även följande användbara samband härledas:

$$h(t) = -\frac{\frac{d}{dt} S(t)}{S(t)} = -\frac{d}{dt} \log S(t). \tag{10}$$

Härledning. Betrakta

$$\frac{d}{dt} S(t) = \frac{d}{dt}(1 - F(t)) = -\frac{d}{dt} F(t) = -f(t).$$

Därmed fås

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\frac{d}{dt} S(t)}{S(t)}.$$

Dessutom

$$-\frac{d}{dt} \log S(t) = -\frac{d}{dS} \log S(t) \frac{d}{dt} S(t) = -\frac{1}{S(t)} \frac{d}{dt} S(t).$$

Alltså gäller ekvation (10).

□

Integreralen av hazardfunktionen kan tänkas som en ackumulerad sannolikhet för en studerad händelse över ett längre tidsintervall. I vissa tidsperioder blir sannolikheten mindre än i andra tidsperioder. Den *integrerade hazardfunktionen* definieras enligt 2.5.

Definition 2.5 (Integrerad hazardfunktion)**.** Den integrerade hazardfunktionen (*alt.* kumulativa hazardfunktionen) $H(a, b)$ definieras enligt

$$H(a, b) = \int_a^b h(t) dt. \tag{11}$$

Dessutom gäller

$$H(a, b) = \frac{\log S(a)}{\log S(b)} \tag{12}$$

ty

Härledning.

$$H(a, b) = \int_a^b h(t) dt = \int_a^b \frac{f(t)}{S(t)} dt = \int_a^b \frac{\frac{d}{dt} F(t)}{S(t)} dt = \int_a^b \frac{-\frac{d}{dt} S(t)}{S(t)} dt = -\log S(t) \Big|_a^b = \frac{\log S(a)}{\log S(b)}$$

□

Vidare har vi $S(a) = P(\tau > a) = P(a < \tau < b) + P(\tau > b)$, där $P(\tau \geq b) = S(b)$, så vi får

$$H(a,b) = \log\left(1 + \frac{P(a < \tau < b)}{P(\tau > b)}\right). \tag{13}$$

2.2 Icke-parametrisk estimering av hazard- och överlevnadsfunktion

En naiv icke-parametrisk estimator av överlevnadsfunktionen $S(t)$ skulle kunna baseras på den *empiriska kumulativa fördelningsfunktionen* $F_n(t)$ som i

$$\hat{S}_{\text{naiv}}(t) = 1 - F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\tau_i > t}. \tag{14}$$

där n är totala antalet individer i en datamängd, och $I_{\tau_i > t}$ antar värdena 1 om $\tau_i > t$ (individen är vid liv), respektive 0 om $\tau_i \leq t$ (individen är ej vid liv).

Ett viktigt begrepp inom överlevnadsanalys är så kallad censurerad data, detta är data som innehåller ofullständig information om objektets överlevnadstid. Inom överlevnadsanalys finns metoder som möjliggör att både ocensurerade och censurerade data analyseras simultant vilket är en av verktygets styrkor [13]. Vi ska nu introducera en estimator som tar hänsyn till censurering.

Edward L. Kaplan och Paul Meier introducerade en icke-parametrisk "product-limit"-estimator för överlevnadsfunktionen $S(t)$ år 1958, vilken ges av

$$\hat{S}_{KM}(t) = \prod_{i=1}^{k}\left(1 - \frac{d_k}{n_k}\right), \tag{15}$$

där varje tid då dödsfall sker är arrangerade enligt $t_1 \leq t_2 \leq ... \leq t_k \leq t$, d_i är antalet dödsfall vid tiden t_i och n_i är antalet subjekt som lever vid tid t_i [16].

Eftersom detta arbete behandlar datapunkter över döda människor med fullständig information om dess ålder används ingen censurerad data. Därför vill vi visa att att den naiva estimatorn för $S(t)$ är ekvivalent med Kaplan-Meiers estimator när ingen censurering äger rum.

Härledning. Om vi arrangerar varje tid då dödsfall sker enligt $t_1 \leq t_2 \leq ... \leq t_k \leq t$ som ovan kan vi skriva om ekvation (14) som

$$\hat{S}_{\text{naiv}}(t) = \frac{1}{n}\left(n - \sum_{j=1}^{k} d_j\right)$$

där d_j är antalet individer som dog vid tid t_j. Vidare skriver vi om överlevnadsfunktionen som en produkt av betingade sannolikheter enligt följande

$$S(t) = P(\tau > t | \tau > t-1)P(\tau > t-1) = q(t)S(t-1),$$

där $q(t) \equiv 1 - P(\tau = t | \tau \geq t)$ och $P(\tau > t-1) = S(t-1)$ enligt definition 2.1, eftersom vi behandlar diskreta variabler. Utveckling av ovanstående formula vidare ger

$$S(t) = q(t)q(t-1)...q(0) = \prod_{t_i : i=0}^{k} q(t_i)$$

med samma indexering av tiden som förut. Vidare kan $q(t)$ omskrivas som $q(t) = 1 - \frac{P(\tau=t)}{P(\tau \geq t)}$. Dett kan estimeras som $\hat{q}_i(t) = 1 - \frac{d_i}{n_i}$. Följdaktligen kan Kaplan-Meierestimatorn skrivas som en produkt av $\hat{q}_i$ enligt

$$\hat{S}_{KM}(t) = \frac{n - d_1}{n} \cdot \frac{n - d_1 - d_2}{n - d_1} \cdot ... \cdot \frac{n - \sum_{j=1}^{k} d_j}{n - \sum_{i=1}^{k-1} d_i}.$$

Här förkortas majoriteten av täljarna bort mot efterföljande nämnare, vilket resulterar i en produkt som är lika ekvationen (14), alltså är Kaplan-Meierestimatorn lika med vår naiva estimator $\hat{S}_{\text{naiv}}$ i fallet utan censurering. □

Med likheten $\hat{S}_{\text{naiv}} = \hat{S}_{\text{KM}}$ i kommer vi estimera den diskreta hazardfunktionen (ekvation (9)) på det naiva tillvägagångsättet enligt ekvation (16). Observera att nämnaren består av $n\hat{S}_{\text{naiv}}$.

$$\hat{h}(t) = \frac{\sum_{i=1}^{n} I_{\tau_i = t}}{\sum_{i=1}^{n} I_{\tau_i > t}} = \frac{d_k}{n - \sum_{j=1}^{k} d_j}. \tag{16}$$

Dessutom kommer den integrerade hazardfunktionen $H(a, b)$ (ekvation (13)) estimeras enligt

$$\hat{H}(a, b) = \log\left(1 + \frac{\sum_{i=1}^{n} I_{a < \tau_i < b}}{\sum_{i=1}^{n} I_{\tau_i > b}}\right). \tag{17}$$

2.3 Parametrisk estimering av hazardfunktion

Det parametriska förfarandet kräver, till skillnad från det icke-parametriska, att vi anpassar en modell till vår data. Olika fördelningar har föreslagits för detta arbete, bland annat *exponentialfördelning* och *geometrisk fördelning*, samt *generaliserad Paretofördelning*. Den sistnämnda har exempelvis använts för att beskriva en hazardfunktion för mortalitet i en artikel av Rootzén och Zholud som undersöker den mänskliga livslängden [17].

Definitionen av denna fördelning ges av 2.6.

Definition 2.6 (Generaliserad Paretofördelning)**.** Definiera Generaliserade Paretofördelningen (*eng.* Generalized Pareto distribution) enligt

$$F_{\text{GP}}(t) = 1 - \left(1 + \frac{\xi t}{\sigma}\right)^{-1/\xi} \tag{18}$$

på $\{t : t > 0 \text{ och } (1 + \xi t/\sigma) > 0\}$ [18]. Här är ξ en formparameter och σ en parameter för skala.

Från ekvation (3) får vi att överlevnadsfunktionen $S_{\text{GP}}(t)$ ges av 1 minus ekvation (18). Från ekvation (10) får vi därmed

$$h_{\text{GP}}(t) = -\frac{\frac{d}{dt}\left(1 + \frac{\xi t}{\sigma}\right)^{-1/\xi}}{\left(1 + \frac{\xi t}{\sigma}\right)^{-1/\xi}} = \frac{1}{\sigma}\left(1 + \frac{\xi t}{\sigma}\right)^{-1} \tag{19}$$

för $\xi \neq 0$. Notera att detta innebär att hazardfunktionens invers blir den linjära funktionen enligt

$$\frac{1}{h_{\text{GP}}(t)} = \sigma\left(1 + \frac{\xi t}{\sigma}\right). \tag{20}$$

I fallet då $\xi \to 0$ gäller

$$S_{\text{GP}}(t) = e^{-\frac{t}{\sigma}}, \tag{21}$$

ty

Härledning.

$$\lim_{\xi \to 0}\left(1 + \frac{\xi t}{\sigma}\right)^{-1/\xi} = \lim_{\xi \to 0} e^{\log\left(1 + \frac{\xi t}{\sigma}\right)^{-1/\xi}} = \lim_{\xi \to 0} e^{-\frac{1}{\xi}\log\left(1 + \frac{\xi t}{\sigma}\right)} = \lim_{\xi \to 0} e^{-\frac{\log\left(1 + \frac{\xi t}{\sigma}\right)}{\frac{\xi t}{\sigma}}\frac{t}{\sigma}} = e^{-\frac{t}{\sigma}}$$

eftersom

$$\lim_{x \to 0} \frac{\log(1+x)}{x} = 1.$$

□

Från detta får vi, också från ekvation (10), att h_{GP} ges av

$$h_{\mathrm{GP}}(t) = -\frac{d}{dt} \log\left(e^{-\frac{t}{\sigma}}\right) = \frac{1}{\sigma}. \tag{22}$$

Vi kan alltså dela upp den inversa hazardfunktionens beteende i tre olika fall:

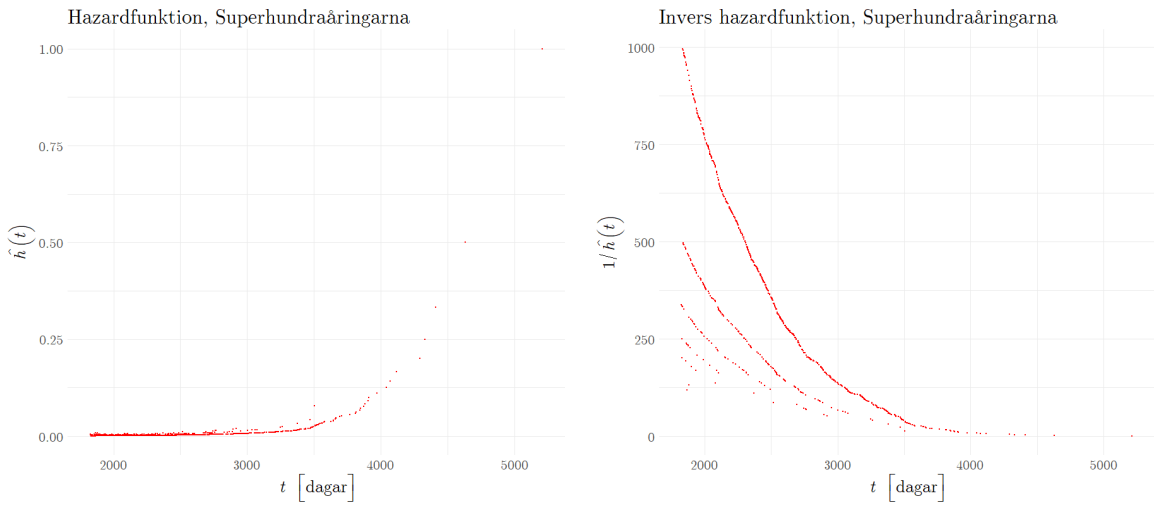
1. Om $\xi > 0$ är $1/h_{\mathrm{GP}}(t)$ en linjär funktion med en negativ lutning.
2. Om $\xi < 0$ är $1/h_{\mathrm{GP}}(t)$ en linjär funktion med en positiv lutning.
3. Om $\xi \to 0$ ger detta oss att hazardfunktionen $h_{\mathrm{GP}}(t) = \lambda$, en godtycklig konstant.

3 Analys och hypotestest

I detta kapitel beskrivs de metoder som används för att analysera de studerade datamängderna och vilka resultat som de genererar.

3.1 Utförande av hazardfunktion

För att inleda analysen undersöker vi hazardfunktionen för våra respektive dataset. Figur 5 och 6 visar icke-parametrisk estimering av hazardfunktion och dess invers i enlighet med ekvation (16) för dataseten över superhundraåringar och italienare. Den inversa hazardfunktionen visualiseras för att ett eventuellt linjärt samband, som i ekvation (20) för den Generaliserade Paretofördelningen, enklare ska kunna observeras. För båda dataseten (figur 5(b) och 6(b)) uppvisar inversen ett splittrat beteende och flera parallella linjer kan tydas. Detta kan förefalla märkligt, men kan förklaras av hazardfunktionens utseende.

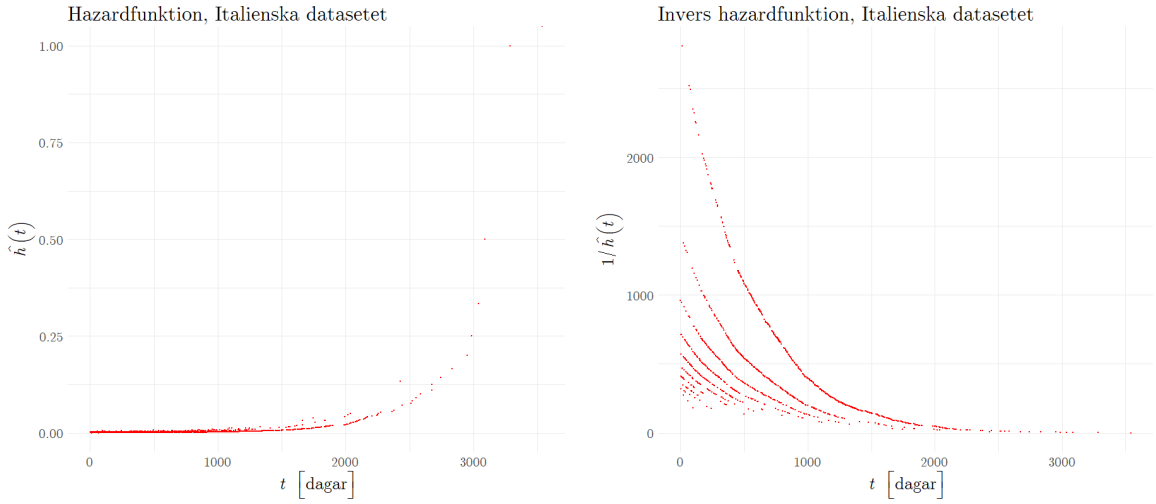


(a) (b)

Figur 5: Delfigurer enligt följande: (a) Hazardfunktion för datasetet superhundraåringar; (b) Invers hazardfunktion för datasetet superhundraåringar.

Nämnaren i ekvation (16) består av antalet individer som fortfarande är vid liv vid tiden t. Detta är ett förhållandevis stort värde (i det italienska datasetet 2883 individer i början av tidsintervallet)

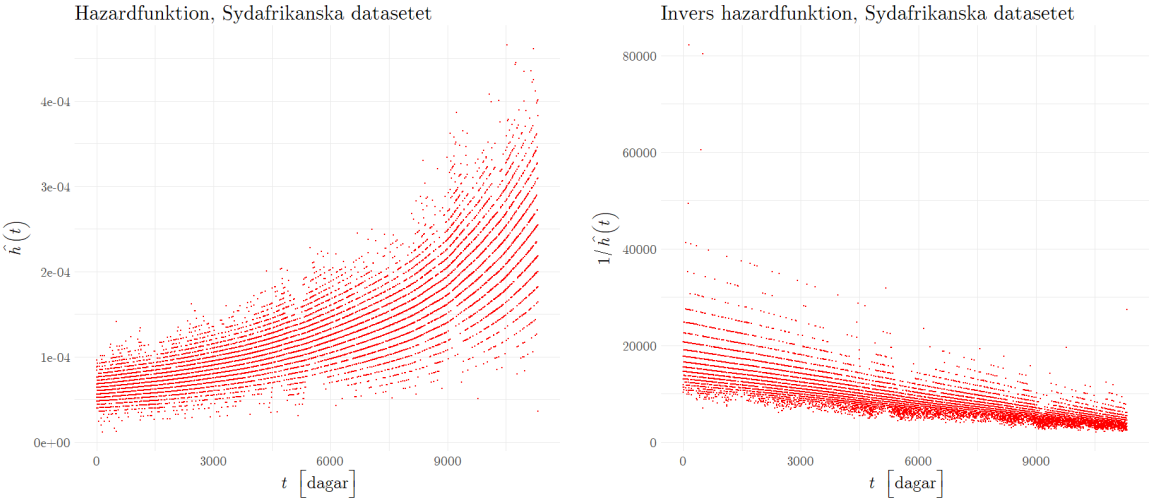
vid jämförelse med täljaren som består av antalet individer som dog vid just tid t (maximalt 13 personer i det italienska datasetet). Således kommer skillnaden mellan nämnarna $\hat{S}(t) - \hat{S}(t+1)$ vara försumbar i sammanhanget. Inverterar vi hazardfunktionen får vi ett stort värde som knappt ändras, dividerat med något litet. Till exempel: anta att värdet på $1/\hat{h}(t) = 2003/2 = 1001.5$ och nästföljande värde $2001/5 = 400.5$. Skillnaden på grund av antalet dödsfall per dag blir stor, och alla dagar då lika stort antal personer dör kommer uppfattas som en egen linje i grafen.



(a) (b)

Figur 6: Delfigurer enligt följande: (a) Hazardfunktion för datasetet Italien; och, (b) Invers hazardfunktion för datasetet Italien.

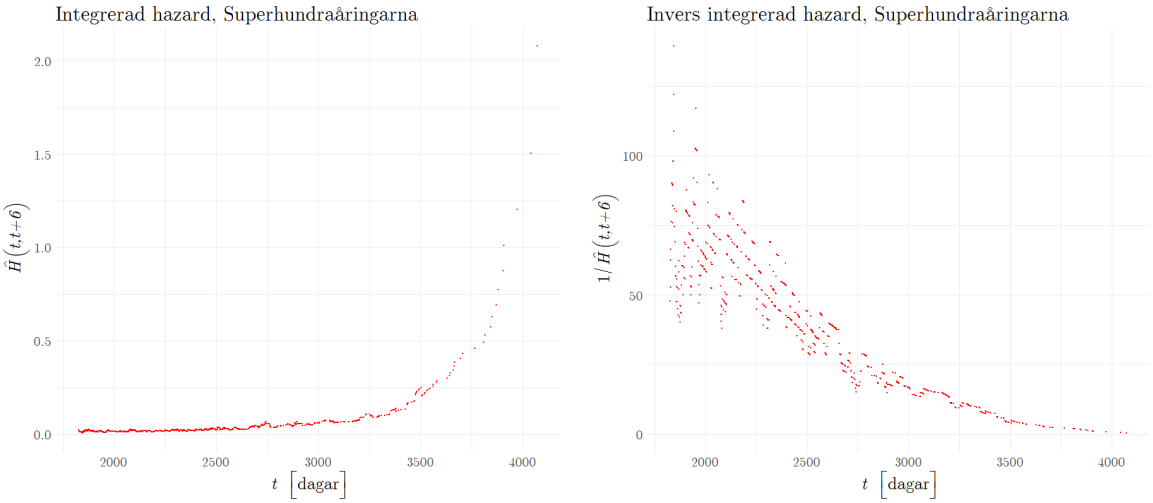
Hazardfunktion och invers hazardfunktion för det sydafrikanska datasetet presenteras i figur 7(a) respektive 7(b). Funktionerna undersöker intervallet mellan 50 - 80 år men beräknas utifrån samtliga dödsfall för individer över 50 års ålder. Därför blir intervallen på de vertikala axlarna annorlunda med ett betydligt mindre omfång när hazardfunktionen beaktas och det omvända gäller för den inversa hazardfunktionen. Den övre gränsen på intervallet sätts med anledning att dödsfallen över 80 ålder sker mindre frekvent, detta leder till en sämre estimering av datamängdens hazardfunktioner.



(a) (b)

Figur 7: Delfigurer enligt följande: (a) Hazardfunktion för det sydafrikanska datasetet; (b) Invers hazardfunktion för det sydafrikanska datasetet.

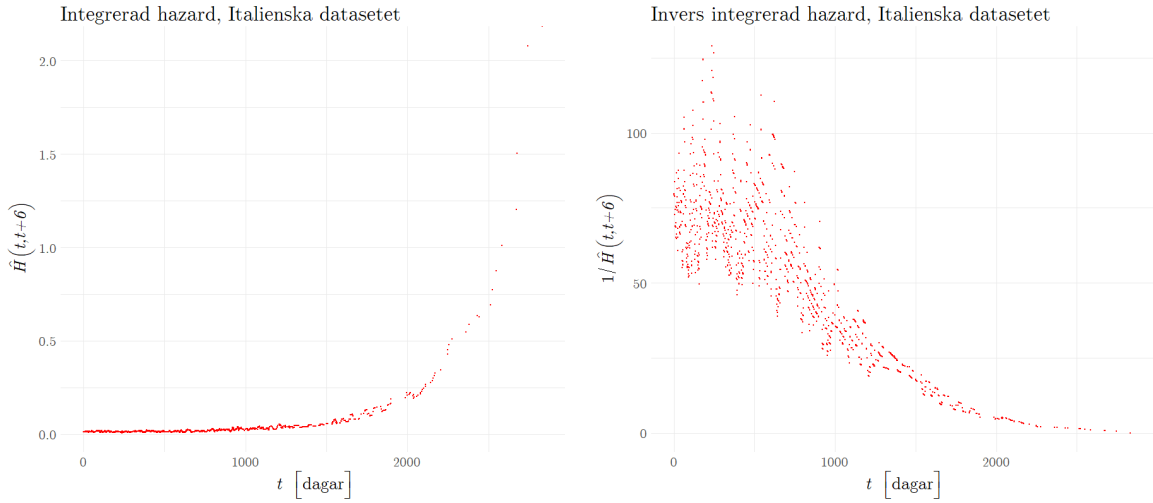
Vi ser att den estimerade inversa hazardfunktionen för det sydafrikanska datasetet uppvisar en antydan till linjärt beteende, något som är svårare att se för de mindre dataseten. Med anledning av det splittrade beteendet vill vi på något sätt behandla den så att vi lättare kan urskilja den dynamik vi letar efter. För att släta ut funktionen beräknar vi därför den integrerade hazardfunktionens estimator enligt ekvation (17) i kapitel 2.2 över ett veckointervall för varje tidssteg. För dag t blir alltså värdet på den vertikala axeln den integrerade hazardfunktionen från t till $t+6$ dagar, och för den därpå följande dagen blir värdet den integrerade hazardfunktionen från $t+1$ till $t+7$ dagar.



(a) (b)

Figur 8: Delfigurer enligt följande: (a) Integrerad hazardfunktion för datasetet över superhundraåringar och (b) Invers integrerad hazardfunktion för datasetet superhundraåringar.

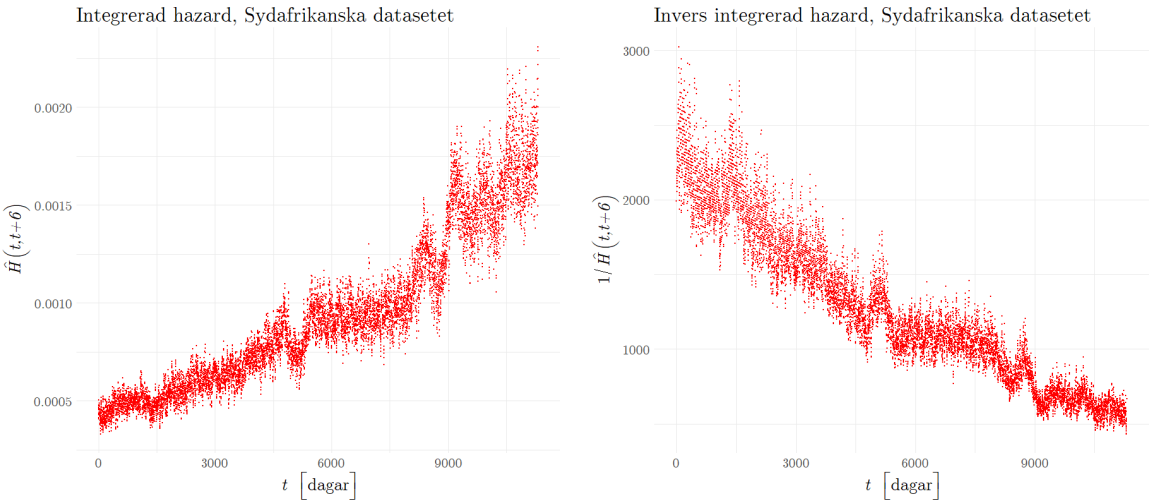
Denna procedur resulterar i figur 8 och 9 för datasetet över superhundraåringar och det italienska datasetet samt figur 10 för det sydafrikanska dataset. De parallella linjerna som förekom i graferna för de inversa hazardfunktionerna (figur 5, 6 och 7) är nu inte lika tydliga, men förekommer fortfarande för de mindre dataseten.



(a) (b)

Figur 9: Delfigurer enligt följande: (a) Integrerad hazardfunktion för det italienska datasetet och (b) Invers integrerad hazardfunktion för det italienska datasetet.

Samtliga figurer kommer jämföras med motsvarande funktions konfidensintervall som förklaras vidare i kaptiel 3.3.



(a) (b)

Figur 10: Delfigurer enligt följande: (a) Integrerad hazardfunktion för datasetet det Sydafrikanska datasetet; (b) Invers integrerad hazardfunktion för datasetet det Sydafrikanska datasetet.

3.2 Bootstrap-metoden

Enligt statistikern Geyner myntades "bootstrap" först i ett statistiskt sammanhang av B. Efron i sin artikel "Bootstrap Methods: Another Look at the Jackknife" [19] [20]. Bootstrapping är en metod för att ta stickprov med återläggning ur en given datamängd, och sedan med hjälp av stickprovet utföra en rad statistiska tester vars resultat sedan kan återkopplas till den originella fördelningen. Med återläggning menas det att en given observation x_i kan plockas flera gånger under ett stickprov, det vill säga att en observation "läggs tillbaka i säcken" efter att den plockats.

Givet en datamängd N med n observationer, det vill säga $N(x_1, x_2, ..., x_n)$, från en okänd fördelning F, tas n nya observationer för att bilda datamängden $\mathbf{N}^*$ så att $\mathbf{N}^*(x_1^*, x_2^*, ..., x_n^*)$. Denna datamängd $\mathbf{N}^*$ kallas för ett bootstrapstickprov.

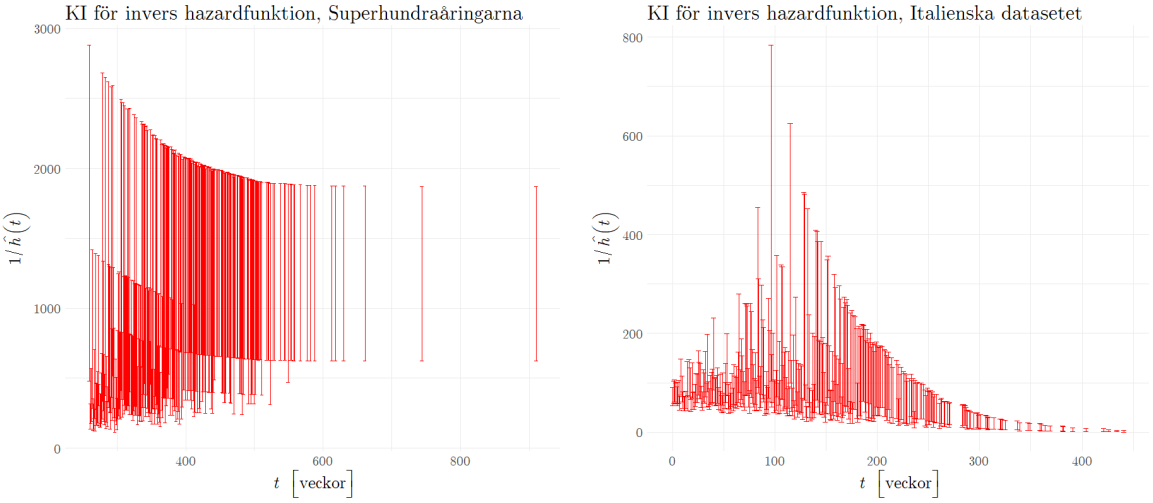
Denna metod för att konstruera bootstrapstickproven repeteras k gånger, så att en serie av stickprov $(\mathbf{N}_1^*, \mathbf{N}_2^*, ..., \mathbf{N}_k^*)$ genereras. På denna serie kan sedan exempelvis medelvärdet beräknas för varje stickprov $\mathbf{N}_i$. Detta kommer ge en serie av medelvärden $(\mu_1^*, \mu_2^*, ..., \mu_k^*)$. Tas sedan andelen $\frac{\alpha k}{100}$ av de största och minsta snitten bort, erhålls ett α-percentilbaserat konfidensintervall för snittet. På samma sätt kan denna metod användas för att konstruera percentilbaserade konfidensintervall med avseende på tidsberoende statistika. Poängen med denna metod är att inget behöver antas om fördelningen som slumpvariabeln härstammar från.

I studien tillämpas bootstrapmetoden för att sedan för varje stickprov $\mathbf{N}_i^*$ beräkna inversen av estimat av hazardfunktionen (16), så att en serie av mängder med inversa estimat $\{\frac{1}{\hat{\mathbf{h}}_1^*}, \frac{1}{\hat{\mathbf{h}}_2^*}, \ldots, \frac{1}{\hat{\mathbf{h}}_k^*}\}$) erhålles och följaktligen kan percentilbaserade konfidensintervall konstrueras.

3.3 Konfidensintervall

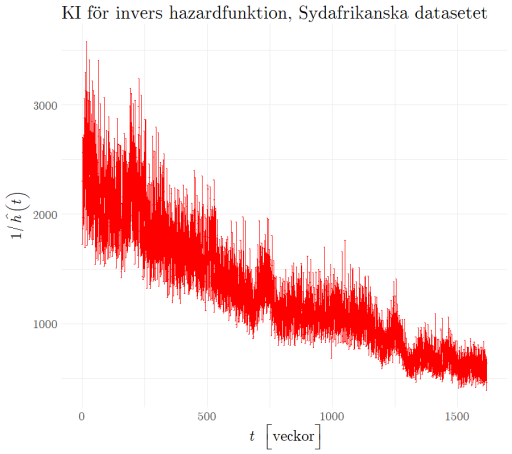
Konfidensintervall skattar en grad av osäkerhet som tillkommer genom statistisk inferens. För att säkerställa att de framtagna funktionerna är pålitliga genereras därför konfidensintervall för de inversa hazardfunktion enligt tidigare beskriven bootstrapmetod i kapitel 3.2. Dessa visas i figur 11(a) för superhundraåringarna, figur 11(b) det italienska datasetet och motsvarande graf för det

sydafrikanska datasetet i figur 12. Resultatet av skattningarna för de två förstnämnda datakällorna skapar konfidensintervall som är så pass stora att en tydlig trend ej kan tydas. Jämför dessa med deras motsvarande inversa hazardfunktioner i figur 5(b) för superhundraåringarna och figur 6(b) för det italienska datasetet. Med denna anledning utförs ingen vidare analys av dessa dataset då varianserna är för stora för att en förskjutningseffekt ska kunna urskiljas. Det sydafrikanska datasetet visar däremot ett snävare intervall och bedömningen görs att dess trend är tillräckligt god för att vidare arbete kan utföras på datakällans hazardfunktion.



(a) (b)

Figur 11: Percentilbaserade konfidensintervall för inversa hazardfunkioner framtagna med bootstrapmetoden. Delfigurer med avseende på dataset (a) för superhundraåringar; (b) för det italienska datasetet.



Figur 12: Percentilbaserade konfidensintervall för det sydafrikanska datasetet framtaget med bootstrapmetoden.

3.4 Regressionsmodeller

Efter att variationen av dödsfall har utvärderats med bootstrap-metoden är nästa steg i analysen att upprätta en modell som förklarar trender i den naiva estimatorn av hazardfunktionen för det sydafrikanska datasetet inom intervallet 50 - 80 år. Den inversa hazardfunktionen i figur 7 beaktas, och dess funktion antas vara ett linjärt uttryck. Vi kan då anpassa en enkel regressionsmodell

enligt ekvation (20) till den inversa hazardfunktionen. Hazardfunktionen i figur 7(a) kan snarare beskrivas av en hyperbolisk funktion, vilket stämmer överrens med GP-fördelningen enligt ekvation (19).

För att anpassa en rät linje till den inversa hazardfunktionen utförs en enkel linjär regression, hädanefter benämnt LM, där minsta-kvadratmetoden används för att erhålla värden på koefficienterna β_0^L och β_1^L i en linjär modell på formen enligt

$$\frac{1}{h_{\text{obs}}(t_i)} = \beta_0^L + \beta_1^L t_i + \epsilon_{ih^{-1}}^L \tag{23}$$

där $h_{\text{obs}}(t_i)$ är den observerade hazardintensiteten, t_i representerar dag i och $\epsilon_{ih^{-1}}^L$ är modellen till inversa hazardfunktionens residualer. Det är viktigt att komma ihåg att det fortfarande är avvikelsen av hazardintensiteten som skall analyseras, vilket leder till att de residualer som skall undersökas inte är på formen i ekvation (23) ovan, utan ges av

$$\epsilon_i^L = h_{\text{obs}}(t_i) - \frac{1}{\beta_0^L + \beta_1^L t_i}. \tag{24}$$

För att vidare underbygga ett beslut om en rät linje som en modell av hazardfunktionens invers, testas signifikansen av termen ${\beta_2}^L$ i modellen enligt

$$h_{\text{obs}}(t_i) = {\beta_0}^L + {\beta_1}^L t_i + {\beta_2}^L {t_i}^2 + \epsilon_{ih^{-1}}^L \tag{25}$$

på den inversa hazardfunktionen för det italienska datasetet. T-test (vars procedur beskrivs i kommande kapitel), ger resultatet att termen ${\beta_2}^L$ inte är signifikant skild från 0, så den linjära modellen föredras för vidare analys.

Vidare, för att anpassa en hyperbel till hazardfunktionen väljs en klass av generaliserade linjära modeller (*eng.* Generalized Linear Models) som generaliserar linjär regression till att relateras till en responsvariabel. För alla generaliserade linjära modeller gäller det att:

- responsvariablen y_i antas följa en fördelning från exponentialfamiljen av fördelningar
- $E(y_i) = \mu_i$
- $g(\mu_i) = \mathbf{X}_i B = \eta_i$, där η_i kallas för länkfunktion (*eng.* link function). $\mathbf{X}_i$ representerar matris med värden på förklarande variabler och B är koefficientvektorn.

Förklarande variabel är i detta fall t i dagar. Den modell som ansätts benämns hädanefter GLM. Vidare, för GLM antas det att länkfunktionen är av invers karaktär som i ekvation (19). Därav blir $\eta_i = \frac{1}{\mu_i}$. Eftersom länkfunktionen är på denna form måste en *Gamma*-fördelning tillämpas. En modell av denna karaktär ser ut på formen

$$h_{obs}(t) = \frac{1}{\beta_0^G + \beta_1^G t_i} + \epsilon_i^G \tag{26}$$

där $h_{obs}(t)$ är den observerade hazardintensiteten och t_i representerar dag i. GLM använder sig av maximum-likelihooduppskattning för att estimera dess koefficienter β_0^G och β_1^G.

Modellerna LM och GLM beskrivna ovan kommer brukas vid modellering av hazardfunktionen. Orsaken till att båda tillvägagångssätt tillämpas är för att kunna ge mer underlag till en eventuell slutsats. Tabell 2 visar sammanfattningar över anpassningsgraden för LM respektive GLM anpassade över intervallet [50, 80] år. Notera att de skattade lutningskoeffienterna är signifikanta, vilket bekräftar antagandet om en icke-konstant hazardfunktion.

Modell	Koefficient	Estimat	Standardfel	t-värde	p-värde för koefficientsignifikans
LM	$\beta^L_{0,\text{tot}}$	15756.5932	59.0041	267.04	$2 \cdot 10^{-16}$
	$\beta^L_{1,\text{tot}}$	-1.1168	0.0090	-123.75	$2 \cdot 10^{-16}$
GLM	$\beta^G_{0,\text{tot}}$	14011.8999	50.6381	276.71	$2 \cdot 10^{-16}$
	$\beta^G_{1,\text{tot}}$	-0.9383	0.0058	-160.49	$2 \cdot 10^{-16}$

Tabell 2: Sammanfattning av anpassningsgraden för modellerna. Notera signifikant lutningskoefficient $\beta^L_{1,\text{tot}}$ och $\beta^G_{1,\text{tot}}$.

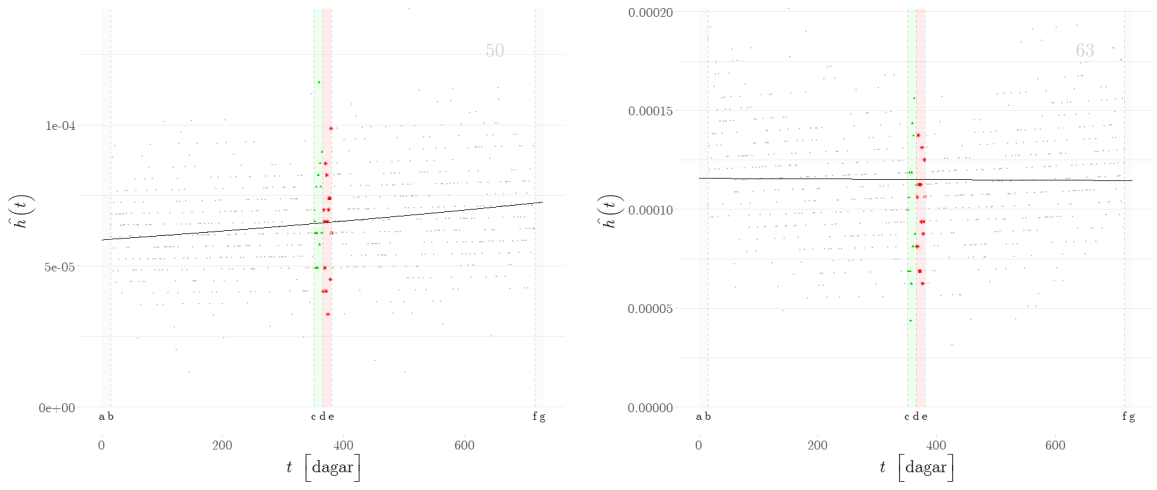
Förklaringsgrad (*eng.* Coefficient of determination) är ett mått på variationen i responsvariabel som förklaras av den förklarande variabeln

$$R^2 = 1 - \frac{SSE_{res}}{SSE_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{27}$$

och används för den linjära modellen. Dock kan inte R^2 enligt ekvation (27) användas som anpassningsgradskriterium för GLM. Dabao Zhang föreslår istället så kallade generaliserade förklaringsgrader, så som R_v^2 [21]. För vår model beräknas ${R_v}^2$ som baseras på variansfunktionen. Båda modellernas p-värden är signifikanta och determinationskoefficienten $R^2 = 0.5749$ för den linjära modellen och ${R_v}^2 = 0.665978$ för GLM. Dessa värden på förklaringsgraderna pekar mot att en stor grad av variationen i hazardfunktionen förklaras av åldern.

3.5 Residualanalys

För det sydafrikanska datasetet, där hazardintensitetet beräknats för alla inom åldersintervallet 50-80 år, skapas 30 stycken överlappande två-årsintervall. Starten på två-årsintervall i är just efter den $(i + 49) : e$ födelsedagen. Med hjälp av GLM beskrivet i kapitel 3.4 illustreras två av dessa i figur 13.



(a) (b)

Figur 13: GLM av intervallet för 50-åringar till 51-åringar i (a) och för 63-åringar till 64-åringar i (b) med observerade värden som punkter, och relevanta intervall runt födelsedagen markerade.

För varje intervall om två år, det vill säga 730 dagar, anpassas modeller med hjälp av regressioner enligt avsnitt 3.4 med avseende på det sammanslagna intervallet $\boldsymbol{D_o}$. Intervallet $\boldsymbol{D_o}$ utgörs av $(d_{15}, \ldots, d_{351})$ tillsammans med $(d_{380}, \ldots, d_{716})$ där d_i motsvarar dag i efter början av intervallet. Detta motsvarar intervall $[b, c] \cup [e, f]$ i figur 13. Denna modell interpoleras sedan för att även

täcka $\boldsymbol{D_i}$ där $\boldsymbol{D_i}$ utgörs av $(d_{352}, \ldots, d_{379})$ som motsvarar intervall $[c, e]$ i figur 13. $\{\boldsymbol{D_o}, \boldsymbol{D_i}\}$ utgör således hela intervallet $[b, f]$. R^2 låg typiskt mellan 1% och 5% för varje tvåårsperiod, däremot hade samtliga modeller signifikanta koefficienter, därför fortskred analysen.

Dagarna $(d_1, \ldots, d_{14})$ och $(d_{717}, \ldots, d_{730})$ (motsvarande $[a, b]$ respektive $[f, g]$ i figur 13) exkluderas helt för varje intervall, med syftet att ej låta data från 14 dagar efter födelsedag $i + 49$ och data från 14 dagar innan födelsedag $i + 51$ påverka vår analys.

Om vi låter väntevärdet för residualerna två veckor innan födelsedagen av intresse ges av $E(\epsilon_f)$ och väntevärdet för residualerna två veckor efter födelsedagen ges av $E(\epsilon_e)$ kan två ensidiga hypotestest upprättas för varje väntevärde enligt (28) och (29)

$$\begin{aligned} &1. \begin{cases} H_0 : E(\epsilon_f) = 0 \\ H_a : E(\epsilon_f) < 0 \end{cases} && 2. \begin{cases} H_0 : E(\epsilon_e) = 0 \\ H_a : E(\epsilon_e) < 0 \end{cases} \\ &\text{”Dip” före födelsedag} && \text{”Dip” efter födelsedag} \end{aligned} \tag{28}$$

$$\begin{aligned} &1. \begin{cases} H_0 : E(\epsilon_f) = 0 \\ H_a : E(\epsilon_f) > 0 \end{cases} && 2. \begin{cases} H_0 : E(\epsilon_e) = 0 \\ H_a : E(\epsilon_e) > 0 \end{cases} \\ &\text{”Peak” före födelsedag} && \text{”Peak” efter födelsedag} \end{aligned} \tag{29}$$

där ”dip” och ”peak” tolkas som nedgångar respektive uppgångar i hazardintensiteten. Enligt förskjutningshypotesen borde en ”dip” observeras innan födelsedagen följt av en ”peak”.

Ensidiga t-test är lämpliga test för utvärdering av hypotestesten (28) och (29) då stickprovstorleken är liten, eftersom antalet residualer är 14 för varje test. T-testets antagande att väntevärdet av stickprovet är normalfördelat och antagandet om att residualerna är oberoende är uppfyllda genom modelkonstruktionen. Testvariabeln t ges av

$$t = \frac{\bar{\epsilon} - \epsilon_0}{s/\sqrt{n}} = \frac{\bar{\epsilon}}{s/\sqrt{14}} \sim t_{n-1}, \tag{30}$$

där $n - 1$ är antal frihetsgrader. Tillämpas test enligt (28) och (29) på $E(\epsilon_f)$ och $E(\epsilon_e)$ kan vi urskilja en signifikant fluktuation åt något håll. Dessa test utförs med testvariabler för t-test definierade enligt ekvation (30). Med en linjär modell och ϵ_i^L framtagna enligt ekvation (24) utförs testen ovan på de 30 två-årsintervallen varpå det erhölls att 13 av 30 $E({\epsilon_f}^L)$ före födelsedagen och 12 av 30 $E({\epsilon_e}^L)$ efter födelsedagen visade en signifikant ”peak”. Bara 1 av 30 $E({\epsilon_f}^L)$ visade en signifikant ”dip”, och för $E({\epsilon_e}^L)$ återfanns ingen signifikant ”dip”.

Test på de 30 två-årsintervallen ovan utförs även med GLM och residualer på formen som ekvation (26), varpå det erhålls att 6 av 30 $E({\epsilon_f}^G)$ före födelsedagen och 6 av 30 $E({\epsilon_e}^G)$ efter födelsedagen visar en signifikant ”peak”. 4 av 30 $E({\epsilon_f}^G)$ visar en signifikant ”dip”, samt för $E({\epsilon_e}^G)$ återfinns också 4 av 30 signifikanta ”dips”. För alla dessa hypotestest sätts signifikansnivån α till 0.1

Det gäller dock att ett hypotestest med signifikansnivån α har sannolikheten α att förkasta nollhypotesen trots att den stämmer. När en stor mängd hypotestest utförs, blir det således förväntat att nollhypotesen förkastas ett antal gånger även om den är sann. För att avgöra om frekvensen av signifikanta fluktuationer av hazardintensiteten åt något håll är högre än vad som är väntat, kan resultaten av signifikansen av t-testen representeras i en mängd av två binära slumpvariabler. Dessa ges av Z_{dip} och Z_{peak} där $\{Z_{\text{dip}}, Z_{\text{peak}}\} : \{\text{Icke signifikant, Signifikant}\} \to \{0, 1\}$. Det innebär att Z antar värdet 0 vid ett icke-signifikant testresultat och att Z antar 1. Om det ej finns någon svängning runt födelsedagarna, bör dessa slumpvariabler Z_{peak} och Z_{dip} båda följa en binomialfördelning, det vill säga att $\{Z_{\text{peak}}, Z_{\text{dip}}\} \sim \text{Bin}(30, \alpha)$ där sannolikheten α är signifikansnivån på hypotestesten (28) och (29) ovan. Därmed kan test på formen enligt (31) utföras på både Z_{dip} och Z_{peak}

$$H : \begin{cases} H_0 : p \leq \alpha \\ H_1 : p > \alpha. \end{cases} \tag{31}$$

Binomialfördelningens kumulativa fördelningsfunktion ges av

$$P(Z \geq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}. \tag{32}$$

P-värde för testet enligt (31) ges av $1 - P(Z \geq k)$. Om detta test returnerar ett signifikant p-värde kan slutsatsen om en genomgående fluktuation dras. Den statistiska signifikansen för testet avgörs med p-värden och signifikansnivån $\alpha = 0.1$. Resultat för binomialtesten baserat på väntevärdestesten enligt (28) och (29) sammanfogas och presenteras i tabell 3. Enligt tabellen kan det ses att människor verkar dö i större utsträckning tiden kring deras födelsedag.

	Före födelsedag				Efter födelsedag			
modell	**p-värde "peak"**	**Signifikant "peak"**	**p-värde "dip"**	**Signifikant "dip"**	**p-värde "peak"**	**Signifikant "peak"**	**p-värde "dip"**	**Signifikant "dip"**
LM	2.3e-06	Ja	0.9576	Nej	1.528e-05	Ja	1	Nej
GLM	0.07319	Ja	0.3762	Nej	0.07319	Ja	0.3762	Nej

Tabell 3: P-värden och signifikans av binomialtest med båda modellerna. Både LM och GLM ger ett "peak-peak"-mönster.

För hela datasetet av residualer utförs en tvåsidig hypotesprövning i syfte att kontrollera att GLM över hela datasetet producerade $E(\epsilon) = 0$ på formen

$$H : \begin{cases} H_0 : E(\epsilon) = 0 \\ H_a : E(\epsilon) \neq 0. \end{cases} \tag{33}$$

Storleken av det här datasetet är 11323 och i detta fall är testvariabeln enligt centrala gränsvärdessatsten approximativt normalfördelad. Därför tillämpas ett z-test med testvariabeln på formen

$$z = \frac{\bar{\epsilon} - \epsilon_0}{s} \tag{34}$$

där s är stickprovsstandardavvikelse som approximerar den riktiga standardavvikelsen σ. Detta test resulterar i att p-värde= 1. Att $p = 1$ innebär mest troligen att beräkningsprogrammet avrundar p till 1 för att det sanna värdet låg nära nog.

4 Diskussion

I detta kapitel diskuteras våra resultat och vilka slutsatser som kan dras utifrån dessa. Arbetets studerade parameter lyfts samt de begräsningar som bedöms vara relevanta. Resultaten jämförs med tidigare studier inom samma område och vidare förslag läggs fram till fortsatt arbete om hypotesen.

4.1 Arbetets utfall

Resultaten som baserades på det sydafrikanska datasetet och sammanställdes i tabell 3 pekar mot att någon eventuell förskjutningstverkan ej är markant nog för att signifikant kunna påvisas i den analys som genomförts i kapitel 3.5. Däremot påvisade resultaten för båda modellerna signifikant ökade hazardintensiteter i tidsperioden före och efter födelsedagen. Det måste dock påpekas att signifikansnivån valdes till 10%. Ifall 5% valdes istället skulle p-värdet för GLM model bli icke-signifikant. Men i och med att båda modellerna påvisade samma fenomen stärker det slutsatsen av att fenomenet existerar.

I kaptitel 3.5 noterades det att R^2 låg mellan 0.01-0.05 för samtliga modeller av två-årsintervallen

som hade konstruerats, trots att modellerna hade signifikanta p-värden. Det föreligger att R^2 är känslig mot en stor varians, vilket är något som återfanns i alla dataset. Däremot var R^2 för modeller anpassade över hela intervallet [50, 80] år 57% för LM och 66% för GLM. I två-årsintervall ändras den förklarande variabeln (åldern) så pass lite att modellerna inte riktigt kan fånga variationen i responsvariabeln. Figur 13 visar att den anpassade regressionslinjen är nära hazardfunktionens medelvärde, vilket leder till att kvoten $\frac{SSE_{res}}{SSE_{tot}}$ blir nästan 1, vilket resulterar i ett lågt R^2. Sammanfattningsvis kan det konstateras att åldern förklarar en genomgriplig trend väl, medan variationen i små tidsspann hade kunnat förklarats bättre med hjälp av utökade modellparametrar.

Inhämtningen av ny data visade sig vara i sig ett utmanade moment med begränsad framgång. Att erhålla mikrodata är inget trivialt ärende, denna typ av information inhämtas och lagras hos statistiska institutioner som ej lämnar ut dessa uppgifter utan rätt befogenhet. Mikrodata är dessutom kostsamt både monetärt och tidsmässigt. Projektets uppfyllde inte kraven för att kunna erhålla data hos majoriteten av statistiska institutioner. Storleken av datamängden är också avgörande vid analys och en större datamängd än vad som var tillgängligt är troligtvis att föredra. Anledningen till detta är att analysen av diskret data blir enklare om det åtminstone dör ett betydande antal personer vid varje tidsteg som undersöks. Detta leder till att kravet på datastorlek fort blir stort, givet att steglängden som valts är relativt liten, vilket krävs för att kunna undersöka förskjutningshypotesen.

Mikrodata, som behandlas i detta arbete, för ofta med sig känslig information om individerna som utgör datamängden. I linje med att respektera deras anonymitet gjordes valet att om sådan data erhölls så skulle detaljer som ej bedömdes vara nödvändiga för att utföra analyserna i arbetet raderas. I övrigt bedömdes det att inga andra etiska eller samhälleliga aspekter behövde beaktas. Ingen typ av känslig data stöttes på under arbetets gång.

4.2 Resultat jämfört med tidigare studier och framtida rekommendationer

Kapitel 1.2 presenterade tidigare studier som undersökt förskjutningshypotesen. Till skillnad från de som producerade resultat som pekade på en förskjutningseffekt, ett ”dip-peak”-mönster, fann vi en ökad dödlighet både före och efter bemärkelsedagen. Detta kan endast sägas tangera en av de föregångna arbetenas resultat. Phillips studie ’The Birthday: Lifeline or Deadline?’, fann att män dog i större utsträckning en tidsperiod innan födelsedagen och kvinnor dog i större utsträckning en tidsperiod efter födelsedagen [5]. Detta visar att det föreligger tidigare slutsatser om att en del av populationen upplever en större risk att avlida före födelsedagen. Någon könsuppdelning gjordes ej i vår studie, vilket leder till att det ej går att säga huruvida resultatet i vår studie reflekterade Phillips studie helt, eller om resultatet är säreget.

Vid en fortsatt analys hade en uppdelning med vidare särskiljning kunnat vara av intresse, exempelvis med avseende på kön i linje med Phillips studie [5]. På så vis hade en eventuell divergens mellan hazardintensiteten kunnat etableras. Som nämnt fann vi att de individuella modellerna i det undersökta intervallet hade låga förklaringsgrader, ålder var alltså en dålig variabel för att prediktera hazardintensiteten på kortare sikt. Därför är det högst relevant att vidareutveckla en regressionsmodell som bättre kan förklara mortaliteten på kortare sikt. Detta kan möjligtvis göras genom tillämpning av multivariata modeller som innefattar flera förklarande variabler. Det borde inte vara orimligt att anta att parametrar som socioekonomisk ställning och beteende mönster runt födelsedagen är faktorer som spelar in i dödsrisken runt händelsen. Som förklarat är mikrodata en typ av informationskälla som är svårtåtkomlig. Vid ett vidare arbete med nytt dataset kan med fördel en förtidsplacerad order placeras för att erhålla mikrodata från fler källor. På så vis kan fler simultana analyser genomföras eller kan olika dataset sammanslås för att erhålla ett större gemensamt sådant.

Att använda födelsedagen som en positivt betydelsefull händelse är därför något som rimligen bör revideras. Det är inte garanterat att födelsdagen är något en människa ser fram emot. Det kan snarare vara en milstolpe som åsamkar stress - fysisk stress på grund av firande vid födelsedagen,

och psykisk stress inför den. Andra har istället valt högtider som meningsfull händelse, men samma problem kan uppstå där. Dessutom påverkar högtider hela populationen så att säsongsberoende effekter kan leda till periodiskt bias. Mer intressant hade varit att undersöka mer relationsbaserade händelser, så som ett barnbarns födelse. Detta är en resurskrävande form av studie som framtvginar mer känsliga uppgifter än detta projekt innefattar.

Referenser

[1] David P. Phillips och Elliot W. King. "DEATH TAKES A HOLIDAY: MORTALITY SURROUNDING MAJOR SOCIAL OCCASIONS". I: *The Lancet* 332.8613 (1988). Originally published as Volume 2, Issue 8613, s. 728 –732. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(88)90198-5. URL: http://www.sciencedirect.com/science/article/pii/S0140673688901985.

[2] J. Lane och R. Lane. "Postponing death: another failure to replicate." I: *Psychosomatic medicine* 54 (nov. 2004), s. 973–4. DOI: 10.1097/01.psy.0000146794.41017.8e.

[3] Gary Smith. "Asian-American Deaths Near the Harvest Moon Festival". I: *Psychosomatic medicine* 66 (maj 2004), s. 378–81. DOI: 10.1097/01.psy.0000127875.38685.ba.

[4] David P. Phillips och Daniel G. Smith. "Postponement of Death until Symbolically Meaningful Occasions". I: *JAMA : the journal of the American Medical Association* 263 (maj 1990), s. 1947–51. DOI: 10.1001/jama.263.14.1947.

[5] David P. Phillips, Camilla A. Van Voorhees och Todd E. Ruth. "The Birthday: Lifeline or Deadline?" I: *Psychosomatic medicine* 54 (sept. 1992), s. 532–42. DOI: 10.1097/00006842-199209000-00001.

[6] "Death has a preference for birthdays – an analysis of death time series." I: *Annals of Epidemiology* 8 (2012), s. 603. ISSN: 1047-2797. URL: http://proxy.lib.chalmers.se/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsgao&AN=edsgcl.296471363&site=eds-live&scope=site.

[7] Holger Leerhoff och Ulrike Rockmann. *Death Won't Wait. Cancer Deaths Around Birthdays and Religious Holidays*. URL: https://www.statistics.gov.hk/wsc/CPS204-P41-S.pdf.

[8] Gerontology Research Group. *About the Gerontology Research Group*. [Online]. 2019. URL: http://www.grg.org/index.html.

[9] Elisabetta Barbi m. fl. "The plateau of human mortality: Demography of longevity pioneers". I: *Science* 360.6396 (2018), s. 1459–1461. ISSN: 0036-8075. DOI: 10.1126/science.aat3119. eprint: https://science.sciencemag.org/content/360/6396/1459.full.pdf. URL: https://science.sciencemag.org/content/360/6396/1459.

[10] Datafirst. *South Africa - Mortality and Causes of Death 2015 Study Description*. [Online]. 2019. URL: https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/610/study-description.

[11] Datafirst. *About DataFirst*. [Online]. 2019. URL: https://www.datafirst.uct.ac.za/about-us.

[12] Statistics South Africa. *Mortality and causes of death in South Africa, 2015: Findings from death notification*. Tekn. rapport. [Report]. URL: https://www.statssa.gov.za/publications/P03093/P030932015.pdf.

[13] David G. Kleinbaum och Mitchel Klein. *Survival Analysis A Self-Learning Text, Third Edition*. [Online]. Springer, New York, NY, 1998, s. 4–15. ISBN: 978-1-4419-6646-9. DOI: 10.1007/978-1-4419-6646-9.

[14] Rupert G. Miller Jr. *Survival Analysis, Second Edition*. Wiley Classics Library. [Online]. John Wiley och Sons, 2011, s. 2. ISBN: 9781118031063. URL: https://books.google.se/books?id=Mv09I8g3zxAC.

[15] D.R. Cox och D. Oakes. *Analysis of Survival Data, Monographs on Statistics and Applied Probability 21*. Boca Raton London New York Washington. D.C.: Chapman och Hall/CRC, 1998, s. 13–15. ISBN: 0-412-224490-X.

[16] Paul Meier E. L. Kaplan. *Nonparametric Estimation from Incomplete Observations*. 53:282, 457-481. Journal of the American Statistical Association, 1958. URL: https://doi.org/10.1080/01621459.1958.10501452.

[17] Holger Rootzén och Dmitrii Zholud. "Human life is unlimited – but short". I: *Extremes* 20.4 (2017), s. 713–728. ISSN: 1572-915X. DOI: 10.1007/s10687-017-0305-5. URL: https://doi.org/10.1007/s10687-017-0305-5.

[18] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values. [electronic resource].* Springer Series in Statistics. Springer London, 2001, s. 75–76. ISBN: 9781447136750. DOI: 10.1007/978-1-4471-3675-0.

[19] B. Efron. *Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7.* no. 1, 1–26. 1979.

[20] C.J. Geyner. *5601 Notes: The subsampling bootstrap.* Minneapolis: University of Minnesota, 2013.

[21] Dabao Zhang. "A Coefficient of Determination for Generalized Linear Models". I: *The American Statistician* 71 (dec. 2016). DOI: 10.1080/00031305.2016.1256839.

A Ytterligare tabeller för analys

	Före födelsedag		Efter födelsedag		
Ålder	**LM**	**GLM**	**LM**	**GLM**	
50	Peak	-	-	-	
51	Peak	-	Peak	-	
52	Peak	Peak	Peak	Peak	
53	Peak	-	-	-	
54	-	-	-	-	
55	-	-	Peak	-	
56	Peak	-	-	-	
57	-	-	-	Dip	
58	-	Dip	-	-	
59	-	-	-	-	
60	-	-	Peak	-	
61	Peak	Peak	-	-	
62	-	-	-	Dip	
63	-	Dip	Peak	Peak	
64	Peak	Peak	-	-	
65	-	-	Peak	-	
66	Peak	-	-	-	
67	-	-	-	-	
68	-	-	Peak	Peak	
69	Peak	Peak	-	-	
70	-	-	-	-	
71	-	-	Peak	Peak	
72	Peak	Peak	-	Dip	
73	Dip	Dip	Peak	-	
74	Peak	-	-	Dip	
75	-	-	Peak	Peak	
76	Peak	-	-	-	
77	-	Dip	Peak	Peak	
78	Peak	Peak	-	-	
79	-	-	Peak	-	

Tabell 4: Signifikanta avvikeler redovisas för modellerna för varje år i intervallet 50 - 79 år. ”Dip” representerar en nedgång i hazardintensiteten och ”peak” en uppgång.

	Dagar efter födelsedag	Ålder i dagar	Ålder efter startålder i dagar	Ålder i år
1	81	42084	23822	115
2	16	42019	23757	115
3	147	42151	23888	115
4	3	42006	23744	115
⋮	⋮	⋮	⋮	⋮
84193	217	27611	9348	75
⋮	⋮	⋮	⋮	⋮
248880	115	18377	115	50
248881	131	18393	131	50

Tabell 5: Exempeltabell baserad på det sydafrikanska datasetet för att illustrera hur informationen i dataseten är organiserad.

B Kod

I denna bilaga presenteras den kod som används för diverse bräkningar och grafer som används i arbetet, samtliga program har utvecklats i programspråket R.

B.1 Histogram

Histogram som visar dödsfrekvenser över datakällans period.

```
hist_days <- function(data,DaysAfterStartDays,interval){
# data: SouthAfrica or equivalent,
# DaysAfterStartDays: Number of days lived in a studied period, e.g people who died between 50 - 80 years old
# interval: Sets the width of each partition in the

library(ggplot2)
library(extrafont)

windowsFonts("CMU Serif" = windowsFont("CMU Serif"))

q <- as.data.frame(SouthAfrica)
# Computes a histogram over the number of days for displaying frequencies over the total number of days for the dataset for a chosen bin size
p <- ggplot(data=q[,c(3,1)],aes(q[,3]))+ geom_histogram(breaks=seq(from=min(days50),to=interval*ceiling(max((days50)-min(days50))/interval)+min(days50),
                  by=interval),
                  col="red",
                  fill="green",
                  alpha = 0.2) +
      ggtitle("interval dagar levda efter 50 r") +
      labs(x="Dagar", y="Frekvens") +theme_minimal() +
      theme(text=element_text(family="CMU Serif", size=16))

p
}
```

B.2 Periodiskt histogram

Histogram som visar dödsfrekvenser över perioden ett år.

```r
hist_daysBD <- function(data){
# data: SouthAfrica or equivalent
library(ggplot2)
library(extrafont)

windowsFonts("CMU Serif" = windowsFont("CMU Serif"))

q <- as.data.frame(SouthAfrica)

# Computes a histogram over a year with the smaller bins representing a
    week and the larger, hollow ones a month
ggplot(data=q)+geom_histogram(aes(x=q[,1],y=..density..),breaks=seq(0,365,
    length=53), col="red", fill="green", alpha = 0.2)+
  geom_histogram(aes(x=q[,1],y=..density..),breaks=seq(0,365,length=13),
                 lty=1, size=0.8, col="darkgreen",
                 fill="white", alpha = 0)+
  ggtitle("Antal dagar levda efter fdelsedag sista ret") +
  labs(x="Dagar", y="Frekvens") +theme_minimal() +
  theme(text=element_text(family="CMU Serif", size=16))
}
```

B.3 Grafer

Används för grafisk hantering av kod som ämnar att producera diagram.

```r
#i  mporting libraries for plots
library(ggplot2)
library(extrafont)

# latex-alike font for plots (need to download font CMU)
windowsFonts("CMU Serif" = windowsFont("CMU Serif"))

# importing dataset
# SouthAfrica <-as.matrix(read.csv("SouthAfricaSorted.csv", header=TRUE,
    sep=";"))
# stringen = "Sydafrikanska datasetet"
#
# datasetet <- SouthAfrica
# age_max <- 80

# datasetet <- as.matrix(read.csv("SuperCSorted.csv", sep=";"))
# stringen = "Superhundraringarna"
# age_max <- 120

datasetet <- as.matrix(read.csv("TrueItalianSorted.csv",sep=";"))
stringen = "Italienska datasetet"
age_max <- 120

#------------------------------------------------------------------------

##script for running hazard.R and making ggplots

hfun <- hazard(datasetet,age_max)

plot1 <- ggplot(data=as.data.frame(hfun$h),aes(x=hfun$x,y=hfun$h))+geom_
    point(size=0.5, shape=20, col="red") +
  ggtitle(paste("Hazardfunktion,",stringen))+
  xlab(italic(t) ~" "~ bgroup("[",dagar, "]"))+
  ylab(italic(hat(h))~italic(bgroup("(",t,")")))+
  theme_minimal()+
  theme(text=element_text(family="CMU Serif", size=20))
```

```
35
36 plot2 <- ggplot(data=1/as.data.frame(hfun$h),aes(x=hfun$x,y=1/hfun$h))+geom
       _point(size=0.5, shape=20, col="red") +
37   ggtitle(paste("Invers hazardfunktion,",stringen))+
38   xlab(italic(t) ~" "~ bgroup("[",dagar, "]"))+
39   ylab(italic(1/hat(h))~italic(bgroup("(",t,")")))+
40   theme_minimal()+
41   theme(text=element_text(family="CMU Serif", size=20))
42
43 #-----------------------------------------------------------------------
44
45 ##  script for running Integrated_hazard.R and making ggplots
46
47 # number of days to compute integrated hazard over
48 interval <- 7
49
50 Hfun <- Integrated_hazard(datasetet,age_max,interval)
51
52 plot3 <- ggplot(data=as.data.frame(Hfun$H),aes(x=Hfun$x,y=Hfun$H))+geom_
       point(size=0.5, shape=20, col="red") +
53   ggtitle(paste("Integrerad hazard,",stringen)) +
54   xlab(italic(t) ~" "~ bgroup("[",dagar, "]"))+
55   ylab(italic(hat(H))~italic(bgroup("(","t,t+6",")")))+
56   theme_minimal() +
57   theme(text=element_text(family="CMU Serif", size=20))
58
59 plot4 <- ggplot(data=1/as.data.frame(Hfun$H),aes(x=Hfun$x,y=1/Hfun$H))+geom
       _point(size=0.5, shape=20, col="red") +
60   ggtitle(paste("Invers integrerad hazard,",stringen)) +
61   xlab(italic(t) ~" "~ bgroup("[",dagar, "]"))+
62   ylab(italic(1/hat(H))~italic(bgroup("(","t,t+6",")")))+
63   theme_minimal() +
64   theme(text=element_text(family="CMU Serif", size=20))
```

B.4 Hazardfunktion

Används vid beräkning av hazardfunktion för dataset. Även inversa hazardfunktionen beräknas genom att beräkna inversen av funktionen.

```
1  hazard <- function(data,age_max){
2  # data: SouthAfrica or equivalent,
3  # age_max: in years, 80 for SouthAfrica
4
5  # computes for total number of individuals in dataset
6    days_tot <- as.numeric(data[,3])
7    deathbyday_tot <- as.data.frame(table(days_tot))
8    deathbyday_tot[,1] = as.numeric(as.character(deathbyday_tot[,1]))
9    tot <- sum(deathbyday_tot[,2])
10
11 # makes a subset of our data from 50-age_max years of age
12   data <- (data[data[,4]<=age_max,])
13   days = as.numeric(data[,3])
14
15 # creats table of days vs frequency
16   deathbyday <- as.data.frame(table(days))
17   deathbyday[,1] = as.numeric(as.character(deathbyday[,1]))
18
19
20 # computes hazard function
21   h <- deathbyday[,2]/(tot-(cumsum((deathbyday[,2]))))
22
```

```
  return(list(hazard=h,x=deathbyday[,1]))
}
```

B.5 Integrerad hazarftunktion

Beräknar en integrerad hazardfunktion.

```
Integrated_hazard <- function(data,age_max,interval){
# data: SouthAfrica or equivalent,
# age_max: in years, 80 for SouthAfrica

# computes for total number of individuals in dataset
  days_tot <- as.numeric(data[,3])
  deathbyday_tot <- as.data.frame(table(days_tot))
  deathbyday_tot[,1] = as.numeric(as.character(deathbyday_tot[,1]))
  tot <- sum(deathbyday_tot[,2])

# makes a subset of our data from 50-80 years of age
  data <- (data[data[,4]<=age_max,])
  days = as.numeric(data[,3])

# creats table of days vs frequency
  deathbyday <- as.data.frame(table(days))
  deathbyday[,1] = as.numeric(as.character(deathbyday[,1]))

# importing library zoo, needed for rollapply
  library(zoo)

# computing integrated hazard for interval with start from every position
    in vector
  H <- log(1+rollapply(deathbyday[,2],interval,sum)/(tot-cumsum(deathbyday
    [,2])[-(1:interval-1)]))

  return(list(H=H,x=deathbyday[1:(length(deathbyday[,1])-(interval-1)),1]))
}
```

B.6 Bootstrap och konfidensintervall

Program som skapar bootstrap och konfidensintervall

B.6.1 Bootstrap

Skapar övre- och undre gränser för ett percentilbaserat konfidensintervall.

```
manboot=function(d,N,it,repl,alpha,stats,k){
# Returns a quantile based confidence interval based on 'it' number of
    bootstrap samples
# d is the data or sample
# N is the fraction of the sample size of s, with replacement N == 1
    typically, subsampling 0.5 <= N <= 0.632
# it is the number of iterations or bootstrap samples to generate
# repl is a logical for replacement, 0 is FALSE and 1 is TRUE
# 1-alpha is the confidence level of the interval
# stats is the statistic you want to compute
# k is a logical for which type of indata your stats function has, 1 for
    TABLED data, 0 for RAW data
  if (k == 1){
    frame = as.data.frame(table(d))
    l <- rep(NA,length(frame$d))
    s <- matrix(data = 0, nrow = length(frame$d), ncol = it)
    for (i in 1:it){
```

```
      l <- as.data.frame((table(sample(d, floor(N*length(d)), replace =
    repl, prob = NULL))),stringsAsFactors = FALSE)
      v  = stats(l)
      for (j in 1:length(l[,2])){
        s[which(frame$d == as.numeric(l$Var1[j])),i] <- v[j]
      }
    }
    s[s == 0] <- NA
    manbootm <- matrix(data = NA, nrow = length(frame$d), ncol = 2 )
    for (i in 1:length(frame$d)){
      manbootm[i,1] <- quantile(s[i, ], alpha/2, na.rm = TRUE)
      manbootm[i,2] <- quantile(s[i, ], 1-alpha/2, na.rm = TRUE)
    }
    manbootm <- na.omit(manbootm)
    return(manbootm)
  }

  else if (k==0) {
    frame = as.data.frame(d)
    colnames(frame) <- "d"
    l <- rep(NA,length(frame$d))
    s <- matrix(data = 0, nrow = length(frame$d), ncol = it)
    for (i in 1:it){
      l <- as.data.frame((sample(d, floor(N*length(d)), replace = repl,
    prob = NULL)),stringsAsFactors = FALSE)
      v  = stats(l)
      l <- as.data.frame(table(l))
      colnames(l) <- c("Var1")
      l$Var1 <- as.numeric(as.character(l$Var1))
      for (j in 1:length(l[,1])){
        s[l$Var1[j],i] <- v[j]
      }
    }
    s[s == 0] <- NA
    manbootm <- matrix(data = NA, nrow = length(frame$d), ncol = 2 )
    for (i in 1:length(frame$d)){
      manbootm[i,1] <- quantile(s[i, ], alpha/2, na.rm = TRUE)
      manbootm[i,2] <- quantile(s[i, ], 1-alpha/2, na.rm = TRUE)
    }
    manbootm <- na.omit(manbootm)
    return(manbootm)
  }
}
```

B.6.2 Vektorbaserad inverse hazard

Beräknar invers hazard med en vektor som indata.

```
invHtb <- function(data){
# Computes an inverse hazard based a vector as indata
# data is the data in as a dataframe with a column of days lived after BD,
    a col with age in whole years.
  k = 248881
  deathbyday <- as.data.frame(table(data))
  ih <- (k-(cumsum((deathbyday$Freq))))/deathbyday$Freq
  deathbyday$ih <- ih
  return(ih)
}
```

B.6.3 Plot för bootstrappad konfidensintervall

Anropar bootstrapfunktionen tillsammans med den vektorbaserade inversa hazardfunktionen och plottar konfidensintervallen.

```
#Computes and plots a confidence interval based on the boostrapmethod

#importing libraries for plots
library(ggplot2)
library(extrafont)

#creates bootdata based on the function manboot
bootdata <- manboot(data,1,1000,1,0.1,invHtb,0)

#latex-alike font for plots (need to download font CMU)
windowsFonts("CMU Serif" = windowsFont("CMU Serif"))

lowerCI <- bootdata[,1]
upperCI <- bootdata[,2]

#for sc och I
#weeks <- bootdata[,3]

#for SA
weeks <- seq(1,length(lowerCI))

#stringen = "Italienska datasetet"
#stringen = "Superhundraringarna"
stringen = "Sydafrikanska datasetet"

plot1 <- ggplot(data=as.data.frame(bootdata),aes(x=weeks,y=(lowerCI+upperCI
    )/2))+
  geom_errorbar(aes(ymin=lowerCI,ymax=upperCI),size=0.3,
                width=5, col="red")+
  ggtitle(paste("KI fr invers hazardfunktion,",stringen))+
  xlab(italic(t) ~" "~ bgroup("[",veckor, "]"))+
  ylab(italic(1/hat(h))~italic(bgroup("(",t,")")))+
  theme_minimal()+
  theme(text=element_text(family="CMU Serif", size=20))
```

B.7 Regressionsmodeller

Kod relaterad till regressionsmodeller.

B.7.1 Regression med linjär modell

```
#read libraries
library(gridExtra)
library(grid)
library(ggplot2)
library(lattice)
library(cowplot)
library(extrafont)

#import font
windowsFonts("CMU Serif" = windowsFont("CMU Serif"))

#data
SouthAfrica <- as.data.frame(SouthAfrica)
period <- 0
Y <- invH(SouthAfrica, period)
```

```
16
17 #intervals
18 k1 <- 31
19 k2 <- 335
20 k3 <- k1 + 365
21 k4 <- k2 + 365
22
23 bb = matrix(data = 0, nrow = dim(Y)[2], ncol = 2)
24 br <- matrix(data = 0, nrow = dim(Y)[2], ncol = k1-1)
25 btp <- matrix(data=0,nrow=dim(Y)[2],ncol=1)
26 btn <- matrix(data=0,nrow=dim(Y)[2],ncol=1)
27 ar <- matrix(data = 0, nrow = dim(Y)[2], ncol = k1-1)
28 atp <- matrix(data=0,nrow=dim(Y)[2],ncol=1)
29 atn <- matrix(data=0,nrow=dim(Y)[2],ncol=1)
30
31 xspan1 <- c(k1:k2)
32 xspan2 <- c(((k2+1):365))
33 xspan3 <- c(k3:k4)
34 xspan4 <- c((k4+1):730)
35
36 lv <- matrix(data=0, nrow=31,ncol=1)
37
38 for (i in 1:dim(Y)[2]){
39   z1 = lm(c(Y[xspan1,i],Y[xspan3,i])~c(xspan1,xspan3),weights=c(xspan1,
       xspan3))
40   #each row in bb contains coefficients b0 and b1 of the linear regressions
       per year
41   bb[i,] <- z1[[1]]
42   br[i,] <- Y[xspan2,i]-(bb[i,1]+bb[i,2]*c(xspan2))
43   #btp[i] is the sum of all postive residuals between days 336-365 after
       last birthday for year i+49
44   btp[i] <- sum(br[i,] > 0)
45   #btn is the sum of all negative residuals between days 336-365 after last
       birthday for year i+49
46   btn[i] <- sum(br[i,] < 0)
47   ar[i,] <- Y[xspan4,i]-(bb[i,1]+bb[i,2]*c(xspan4))
48   #atp[i] is the sum of all postive residuals between days 0-30 after last
       birthday for year i+49+1
49   atp[i] <- sum(ar[i,] > 0)
50   #atn[i] is the sum of all negative residuals between days 0-30 after last
       birthday for year i+49+1
51   atn[i] <- sum(ar[i,] < 0)
52 }
53
54 #creats one plot for every two year interval with ggplot
55 makeplot <- function(i){
56   year=i+49
57   if(period == 0){
58   stryear = sprintf("%d",year)
59   stringen2 = sprintf("Dagar efter fdelsedag")
60   }else{
61     stryear = sprintf("%d.5",year)
62     stringen2 = sprintf("Dagar efter halvrsdag")
63   }
64   ggplot()+
65     geom_point(data = as.data.frame(cbind(k2:365,Y[k2:365,i])),aes(k2:365,Y
       [k2:365,i]),size=2, shape=1, col="green")+
66     geom_point(data = as.data.frame(cbind(366:k3,Y[366:k3,i])),aes(366:k3,Y
       [366:k3,i]),size=2, shape=8, col="red")+
67     geom_point(data = as.data.frame(cbind(k3:k4,Y[k3:k4,i])),aes(k3:k4,Y[k3
       :k4,i]),size=0.5, shape=19, col="lightgray")+
```

```
    geom_point(data = as.data.frame(cbind(k1:k2,Y[k1:k2,i])),aes(k1:k2,Y[k1:k2,i]),size=0.5, shape=19, col="lightgray")+
    geom_line(data = as.data.frame(cbind(0:730,bb[i,1]+bb[i,2]*c(0:730))),aes(0:730,bb[i,1]+bb[i,2]*c(0:730)),size=0.1)+
    geom_segment(aes(x = k2, y = -10, xend = k2, yend = max(Y[,i])), col = "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = k3, y = -10, xend = k3, yend = max(Y[,i])), col = "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = 365.5, y = -10, xend = 365.5, yend = max(Y[,i])), col = "lightgray", lty=2, data = NULL)+
    annotate("rect", xmin = k2, xmax = 365, ymin = -10, ymax = max(Y[,i]), fill="green",
             alpha = .08)+
    annotate("rect", xmin = 366, xmax = k3, ymin = -10, ymax = max(Y[,i]), fill="red",
             alpha = .08)+
    geom_text(aes(label = stryear, family ="CMU Serif",x=650,y=0.9*max(Y[,i])), size = 12,col="lightgray") +
    labs(x=stringen2, y="") + theme_minimal()+
    theme(panel.grid.minor.x = element_blank(),panel.grid.major.x = element_blank(),text=element_text(family="CMU Serif", size=16),axis.ticks = element_line(colour = "lightgray", size=(0.5)))
}


nums <- 1:dim(Y)[2]

#put created plots in grid as subplots
plots<- lapply(nums,makeplot)
p1 <- plot_grid(plotlist=plots[1:6],ncol=2)
p2 <- plot_grid(plotlist=plots[7:12],ncol=2)
p3 <- plot_grid(plotlist=plots[13:18],ncol=2)
p4 <- plot_grid(plotlist=plots[19:24],ncol=2)
p5 <- plot_grid(plotlist=plots[25:dim(Y)[2]],ncol=2)
pspec <- plot_grid(plotlist=plots[c(1,5,11,15,21,25)],ncol=2)
pspec

pspec1 <- plot_grid(plotlist=plots[c(1,5)],ncol=2)
pspec1

pspec2 <- plot_grid(plotlist=plots[c(11,15)],ncol=2)
pspec2

pspec3 <- plot_grid(plotlist=plots[c(21,25)],ncol=2)
pspec3
```

B.7.2 Regression med generaliserad linjär modell

```
#Creates generalized linear models of the hazard in two-year intervals and computes t-ttests for interpolated residuals around the birthday.
Y <- invH(data, 0, pop)
Y <- Y[1:730,]
k1 <- 15
k2 <- 351
k3 <- k1 + 365
k4 <- k2 + 365
x1 = c(1:730)
xspan1 <- c(k1:k2)
xspan2 <- c(((k2+1):365))
xspan3 <- c(k3:k4)
xspan4 <- c((k4+1):730)
```

```
bb = matrix(data = 0, nrow = dim(Y)[2], ncol = 2)
br <- matrix(data = 0, nrow = dim(Y)[2], ncol = k1-1)
ar <- matrix(data = 0, nrow = dim(Y)[2], ncol = k1-1)
ttestvec1 <- matrix(data=0, nrow=dim(Y)[2],ncol=1)
ttestvec2 <- matrix(data=0, nrow=dim(Y)[2],ncol=1)
ttestvec3 <- matrix(data=0, nrow=dim(Y)[2],ncol=1)
ttestvec4 <- matrix(data=0, nrow=dim(Y)[2],ncol=1)
reglist = list()
for (i in 1:dim(Y)[2]){
  #constructs a generalized linear model for the given two-year interval
  z1 = glm(1/c(Y[xspan1,i],Y[xspan3,i])~c(xspan1,xspan3), family = Gamma(
    link = "inverse"))
  reglist[[i]] <- z1
  bb[i,] <- z1[[1]]
  #stores interpolated residuals before birthday
  br[i,] <- 1/Y[xspan2,i]-1/(bb[i,1]+bb[i,2]*c(xspan2))
  #stores interpolated residuals after birthday
  ar[i,] <- 1/Y[xspan4,i]-1/(bb[i,1]+bb[i,2]*c(xspan4))
  #computes t-tests for interpolated residuals right before and after
    birthdays and saves the p-values
  ttestvec1[i] <- t.test(br[i,], alternative = "less", conf.level = 0.9)
    [[3]]
  ttestvec2[i] <- t.test(br[i,], alternative = "greater", conf.level = 0.9)
    [[3]]
  ttestvec3[i] <- t.test(ar[i,], alternative = "greater", conf.level = 0.9)
    [[3]]
  ttestvec4[i] <- t.test(ar[i,], alternative = "less", conf.level = 0.9)
    [[3]]

}
## saves p-value of binomial tests
bintestvec <- matrix()
bintestvec[1] <- binom.test(sum(ttestvec1 < 0.1), length(ttestvec1), p =
    0.1, alternative = "less")
bintestvec[2] <- binom.test(sum(ttestvec2 < 0.1), length(ttestvec2), p =
    0.1, alternative = "greater")
bintestvec[3] <- binom.test(sum(ttestvec3 < 0.1), length(ttestvec3), p =
    0.1, alternative = "greater")
bintestvec[4] <- binom.test(sum(ttestvec4 < 0.1), length(ttestvec4), p =
    0.1, alternative = "less")
#legend("topright",legend=agestr, col=unique(c(1,lv)),cex=0.75, lty=1:2)
library(gridExtra)
library(grid)
library(ggplot2)
library(lattice)
library(cowplot)
library(extrafont)

#plot around birthday
period <- 0

# code for making ggplots
makeplot <- function(i){
  year=i+49
  stryear = sprintf("%d",year)

  ggplot()+
    geom_point(data = as.data.frame(cbind(k2:365,1/(Y[k2:365,i]))),aes(k2
    :365,(1/Y[k2:365,i])),size=1, shape=17, col="51")+
    geom_point(data = as.data.frame(cbind(366:k3,1/(Y[366:k3,i]))),aes(366:
    k3,1/(Y[366:k3,i])),size=1, shape=8, col="red")+
```

```
    geom_point(data = as.data.frame(cbind(k3:k4,1/(Y[k3:k4,i]))),aes(k3:k4
    ,1/(Y[k3:k4,i])),size=0.33, shape=19, col="gray75")+
    geom_point(data = as.data.frame(cbind(k1:k2,1/(Y[k1:k2,i]))),aes(k1:k2
    ,1/(Y[k1:k2,i])),size=0.33, shape=19, col="gray75")+
    geom_line(data = as.data.frame(cbind(0:730,1/(bb[i,1]+bb[i,2]*c(0:730))
    )),aes(0:730,1/(bb[i,1]+bb[i,2]*c(0:730))),size=0.1)+
    geom_segment(aes(x = k2, y = 0, xend = k2, yend = max(1/Y[,i])), col =
    "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = k3, y = 0, xend = k3, yend = max(1/Y[,i])), col =
    "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = 365.5, y = 0, xend = 365.5, yend = max(1/Y[,i])),
    col = "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = k1, y = 0, xend = k1, yend = max(1/Y[,i])), col =
    "lightgray", lty=2, data = NULL)+
    geom_segment(aes(x = k4, y = 0, xend = k4, yend = max(1/Y[,i])), col =
    "lightgray", lty=2, data = NULL)+
    annotate("rect", xmin = k2, xmax = 365, ymin = 0, ymax = max(1/Y[,i]),
    fill="green",
             alpha = .07)+
    annotate("rect", xmin = 366, xmax = k3, ymin = 0, ymax = max(1/Y[,i]),
    fill="red",
             alpha = .07)+
    annotate("rect", xmin = 0, xmax = k1, ymin = 0, ymax = max(1/Y[,i]),
    fill="lightgray",
             alpha = .1)+
    annotate("rect", xmin = k4, xmax = 730, ymin = 0, ymax = max(1/Y[,i]),
    fill="lightgray",
             alpha = .1)+
    geom_text(aes(label = stryear, family ="CMU Serif",x=650,y=0.9*max(1/Y
    [,i])), size = 8,col="lightgray") +
    geom_text(aes(label = "a", family ="CMU Serif",x=0,y=-0.02*max(1/Y[,i])
    ), size = 5) +
    geom_text(aes(label = "b", family ="CMU Serif",x=k1,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    geom_text(aes(label = "c", family ="CMU Serif",x=k2,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    geom_text(aes(label = "d", family ="CMU Serif",x=365,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    geom_text(aes(label = "e", family ="CMU Serif",x=k3,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    geom_text(aes(label = "f", family ="CMU Serif",x=k4,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    geom_text(aes(label = "g", family ="CMU Serif",x=730,y=-0.02*max(1/Y[,i
    ])), size = 5) +
    xlab(italic(t) ~"  "~ bgroup("[",dagar, "]"))+
    ylab(italic(hat(h))~italic(bgroup("(",t,")")))+
    theme_minimal()+
    theme(panel.grid.minor.x = element_blank(),panel.grid.major.x = element
    _blank(),text=element_text(family="CMU Serif", size=20),axis.ticks =
    element_line(colour = "lightgray", size=(0.5)))
}


nums <- 1:dim(Y)[2]

#put created plots in grid as subplots
plots<- lapply(nums,makeplot)
```

B.7.3 Inverse hazard med upp av två-årsintervall

Funktion som beräknar invers hazard för två-årsintervall och returnerar en matris där varje kolonn motsvarar inversa hazardintensiteter för varje intervall.

```
invH <- function(data,period, pop){
  #data is the data in as a dataframe with a column of days lived after BD,
     a col with age in whole years.
  #period is the shift in days after birthday from the beginning of a year
  #y denotes the interval in whole years of age that the analysis is done
    on
  #Pop is the total number of people in the population you analyze
    y <- max(data$AgeYear)-min(data$AgeYear)
    if (period != 0){
      y = y-1
    }
    l <- 735
    M <- matrix(data=0, nrow = l, ncol = y) #M is a placeholder matrix
    for (i in 1:y){
      z <- i + 1

      #j1, j2, j3 subsets with respect to both age in whole years and age
    measured in days after last birthday
      j1 <- subset(data,AgeYear == 49+i & AgeAfterStartDays >= period)
      j2 <- subset(data,AgeYear == 49+i+1)
      j3 <- subset(data,AgeYear == 49+i+2 & AgeAfterStartDays < period)

      j <- rbind(j1,j2,j3)

      #sets the AgeAfterStartDays column to a time series starting from 0
    days
      # observe that it no longer is days after birthday after this has
    been computed
      j$AgeAfterStartDays[j$AgeYear == 49+i] <- j$AgeAfterStartDays[j$
    AgeYear == 49+i] - period
      j$AgeAfterStartDays[j$AgeYear == 49+z] <- j$AgeAfterStartDays[j$
    AgeYear == 49+z] + 365 - period
      j$AgeAfterStartDays[j$AgeYear == 49+z+1] <- j$AgeAfterStartDays[j$
    AgeYear == 49+z+1] + l - period

      #tabling the data for computation
       deathbyday <- as.data.frame(table(j$AgeAfterStartDays))

      #computes inverse hazard
      ih <- (pop-(cumsum((deathbyday$Freq))))/deathbyday$Freq
      #stores number of total deaths in the year i
      totdeath <- dim((subset(data, AgeYear == i+49)))[1]
      pop <- pop-totdeath
      #print(k)
      for (u in 1:length(ih)){
        M[u,i] <- ih[u]
      }

    }
  return(M)

  #plot(deathbyday$data,invHM)
}
```

Beräknar R_v^2 för GLM

```
library(rsq)
#Computes variance based R-squared for GLM
R_squared <- rsq(modell,adj = TRUE,type='v')
```