Mikroøkonometri ugeeksamen

15. juni, 2020

Studienr.: 20144070

Antal anslag: 35.517

Indholdsfortegnelse

1 Præsentation af problemstillinger

Opgaven sætter ud for at undersøge, hvad der påvirker at en person starter som selvstændig (hændelsen). I denne sammenhæng anskues især familiemæssige forhold. For at undersøge hændelsen opstilles fire hypoteser, hvilke bekræftes eller afkræftes i analysen[1].

Et relevant familiemæssigt forhold kan være, at personen har småbørn. Småbørn er særligt tidskrævende, hvilket kan have en lempelig effekt på en persons overskud, hvormed de ikke har tid til at starte op som selvstændig. Udover dette kan der være en tendens til at være mere risikoavers som nybagt forælder. Hvis dette er tilfældet, burde personer med småbørn, alt andet lige, have længere til hændelsen end personer uden, dvs. en mindre stejl overlevelsesfunktion[2]. Den første hypotese kan derfor formuleres som:

H_1: Personer med småbørn er mindre tilbøjelige til at starte som selvstændige.

Et andet familiemæssigt forhold, med indflydelse på tiden til hændelsen, antages at være om personen er enlig eller i et forhold. Selvom man ikke nødvendigvis forsørger hinanden i et parforhold, vil nogle måske vælge en mere sikker retning, af hensyn til ens partner, eftersom en fejlet opstart som selvstændig ikke kun påvirker en selv. Udover dette kan der, i større grad for personer i parforhold end enlige, være planer om børn, køb af hus, bryllup eller andre større udgifter, hvorved der potentielt er nogle, der ikke finder råd til de sunkne omkostninger, forbundet ved opstart som selvstændig. Det er også muligt, at der vælges flere personer, der efterstræber stabilitet i dagligdagen, når der udelukkende ses på personer i et parforhold, end der ellers ville være gældende for en repræsentativ stikprøve. Personer i parforhold vil derfor forventes at have mindre risiko for hændelse end personer uden en partner. Derfor bliver den anden hypotese, der undersøges:

H_2: Personer i parforhold har mindre sandsynlighed for at starte som selvstændige end enlige.

Udover familiemæssige forhold viser statistik indenfor området at mænd er mere tilbøjelige til at starte op som selvstændige end kvinder. I 2017 var tre ud af fire nye iværksættere mænd (Egedesø et al. 2018). Det forventes derfor at overlevelsesfunktionen for mænd vil være signifikant lavere end overlevelsesfunktionen for kvinder. Den tredje hypotese der udforskes, er derfor:

H_3: Mænd er væsentligt mere tilbøjelige til at starte som selvstændige end kvinder.

Tal fra Erhvervsstyrelsen i 2014 viste, at størstedelen af danske iværksættere var mellem 36-40 år (Kjempff, 2014). Dette er ikke nødvendigvis tilfældet for opgavens datasæt, med tal fra 1980 til 2005, men der kan findes logiske argumenter for, hvorfor dette alligevel kan være tilfældet. Personer i den aldersgruppe kan forventes at have en del erhvervserfaring, og kan derfor have oplevet nogle områder med plads til forbedring på deres arbejde. De kan have dannet et netværk indenfor branchen, hvilket ville gøre overgangen til selvstændig nemmere. Længerevarig akkumulation af dårlige arbejdsoplevelser kan også have en effekt på

[1] Hypotesernes kritikpunkter udelades derfor fra diskussionen i afsnit 1.

[2] En redegørelse for overlevelsesfunktionen kan findes i afsnit 4.1.

deres motivation til at starte som selvstændig. Samtidig kan det være at personer tættere på deres pension, ikke er lige risikovillige, og derfor blot fortsætter med deres arbejde, indtil de kan pensionere sig. Efter pensionering kan der være flere, der forsøger sig med at blive selvstændig, hvis de er utilfredse med deres pensionssats, eller fordi de endelig har tiden. Såfremt ovenstående er sandt, burde overlevelsestiden generelt være længere for unge og ældre, end midaldrende. Derfor bliver den fjerde hypotese:

H_4: Midaldrende personer har større sandsynlighed for at blive selvstændig end andre alderskategorier.

2 Operationalisering af problemstilling

Første operationaliseringsopgave er at få defineret hændelsen. Dette diskuteres i forbindelse med redegørelsen for populationen i afsnit 3. Her anvendes en overgang fra enhver anden kategori af *pstill*, til kategori 1 eller kategori 11[3], som en hændelse. Dette defineres som variablen *selv*, der tager værdien 1, når personen er blevet selvstændig. Her vil hændelsen noteres, i variablen *event*, i perioden før skiftet fremgår i datasættet. Det vil sige, hvis personen havde en anden primær arbejdsstilling end selvstændig i den j'te periode, og i periode $j+1$ står som værende selvstændig, vil hændelsesvariablen tage værdien 1 i periode j. Årsagen til dette er, at på tidspunkt $j+1$ er hændelsen indtruffet. Det kunne derfor forårsage bias, hvis tidsvariende variable i denne periode inddrages i analysen, eftersom ændringer i disse netop kunne være forårsaget af hændelsen. Variablen *sekvens* bruges til at tælle den enkelte persons forløb, f.eks. identificerer *sekvens*=2 personer, der er i deres andet forløb. Et eksempel på en udeladt faktorer, kunne her være noget der beskriver, hvorfor nogle aldrig bliver selvstændige. Her kunne en variabel der angav, hvorvidt en arbejder var reaktiv eller proaktiv måske være behjælpsom, da det kan antages, at en reaktiv arbejder ville være mere tilpas med at arbejde under andre, end at begynde som selvstændig. Det er ikke alle relevante faktorer, der kan udledes ud fra datasættet, da mange relevante faktorer ville være ekstremt omkostningsrige at indsamle, hvilket besværliggør det til en større kvantitativ undersøgelse.

For at angive, hvornår en person indgår i og udgår fra datasættet, opstilles to dikotome variable, hhv. *pstart* og *pslut*. Disse anvendes bl.a. således personer der er selvstændige første gang de måles, ikke bliver talt af hændelsesvariablen. Til at lave variablen *pstart* og *pslut* anvendes datasættets variable *aar* og *nr*, således *pstart* (*pslut*) ved det tidligste (seneste) år, et bestemt personnummer er i datasættet, tager værdien *1*. Variablene *alder*, *aar* og *dod* anvendes til at opstille variablen *censor*, der udtrykker de forskellige højrecensureringer i datasættet.

Med hændelsesvariablen defineret kan tidstællingen påbegyndes. Dette gemmes i variablen *tid*, der tager værdien 0, ved første tidsperiode en person angives som **ikke** selvstændig, dvs. *selv*=0. Herefter tæller den op indtil personens data udgår, eller personen bliver selvstændig.

For at udforske hypoteserne beskrevet i afsnit 1, dannes forklarende variable til hver af disse. Til at belyse

[3]Henvises fremover som selvstændige.

H_1 dannes en ny variabel, der angiver, hvorvidt personen har småbørn. Småbørn bliver her begrænset af datasættet til at være årene 0-6, eftersom variablene *boern03* og *boern36* anvendes. Disse omkodes således den største værdi af de to anvendes i perioden. Dette gøres igennem **max** funktionen i SAS. Denne anvendes som en tidsvariende variabel, for at bedre kunne undersøge hypotesen. I denne forbindelse er det netop tidsintervallet op til hændelsen, der er interessant, og ikke udelukkende perioden før hændelsen. Variablen er derfor bedst egnet til at være tidsvarierende, hvilket gøres i form af de 26 kovariater, *smboern1-smboern26*. Der er mange andre familiemæssige forhold, der kunne være interessante at undersøge. F.eks. kunne information om et individs forældre, herunder uddannelse og hvorvidt de har været selvstændige før, anvendes til at forklare hændelsen. Der kunne her opstilles en hypotese om, at der er nedarvede kvaliteter, der har indflydelse på et individs risiko for at opleve hændelsen. En sådan opstilling er ikke er muligt med det gældende datasæt.

I forbindelse med H_2 opstilles en tidsvariende variabel, der angiver, hvorvidt personen er i et parforhold til hvert tidspunkt, de indgår i datasættet. Til dette anvendes variablen *ctype*. Hvis *ctype* tager værdien 1, 2, 31 eller 53 konkluderes det, at personen ikke har en partner i perioden, hvormed *parf* kovariaten, til den tilhørende periode, tager værdien 0. Her tælles personer i kategori 53, som værende enlige og personer i kategori 42 som værende i et parforhold, til det formål at variablene har en værdi til alle udfald af *ctype*. Variablen *koen* omkodes til variablen *sex*, hvilken tager værdien 0 for mænd og 1 for kvinder. Denne anvendes til at udforske H_3. Til H_4 opstilles tre alders kategorier. Dette gøres først ved at lave variablen *startalder*, ved at subtrahere tiden fra alderen, givet ved variablen *alder*. De dikotome variable *under30*, *mellem30og50* og *over50*, tager værdien 1, når personens startalder er indenfor intervallet, og ellers 0. Derfor anvendes maks to af disse i regressionsmodeller, eftersom den tredje er referencegruppen, da den er impliceret af de andre to.
Det der ønskes information om, er hvilken alder personen havde, da de startede deres forløb, for at belyse om aldersgrupperne har forskellige risici for at undergå hændelsen. Datasættet giver ikke et fuldkomment billede af individernes forløb, eftersom forløbene allerede er påbegyndte. I datasættet findes individer, der først begynder i datasættet, når de er over 50 år gamle. Disse personer er under risiko for at opleve hændelsen, men grundet datasættets begrænsninger, er det ikke muligt at tælle hele forløbet. Det forventes derfor at risikoen for hændelse overestimeres, eftersom der, alt andet lige, må være kortere til hændelse.

3 Redegørelse for populationen

Opgavens population ønskes at holdes så bred som muligt, da det ikke kan aflæses direkte af datasættet, hvem der ikke kan blive selvstændig. Der er nogle, der aldrig bliver selvstændig, men karakteristikkerne for dem som aldrig bliver, er lige vigtig, som dem der gør. Opgaven sætter ud for at undersøge *hvad der har*

betydning for, at en person starter som selvstændig, hvilket ikke kan besvares ved kun at se på personer, der oplever hændelsen.

Et problem med det tilgængelige data er, at forløbet næsten altid er igangværende. Her kunne der argumenteres for, at personen skal være voksen for at indgå i populationen under risiko, hvormed populationen under risiko reduceres til 4758 personer. Her er det blot 366 af disse, der oplever hændelsen, hvilket er en reduktion på 77%. Yderligere vil alderseffekter skævvrides, eftersom den maksimale alder i datasættet bliver 43.

For at teste om venstre censureringen har en signifikant effekt, udføres en **lifetest** stratificeret for *vcensor* kovariaten. Dette ses i figur 1.

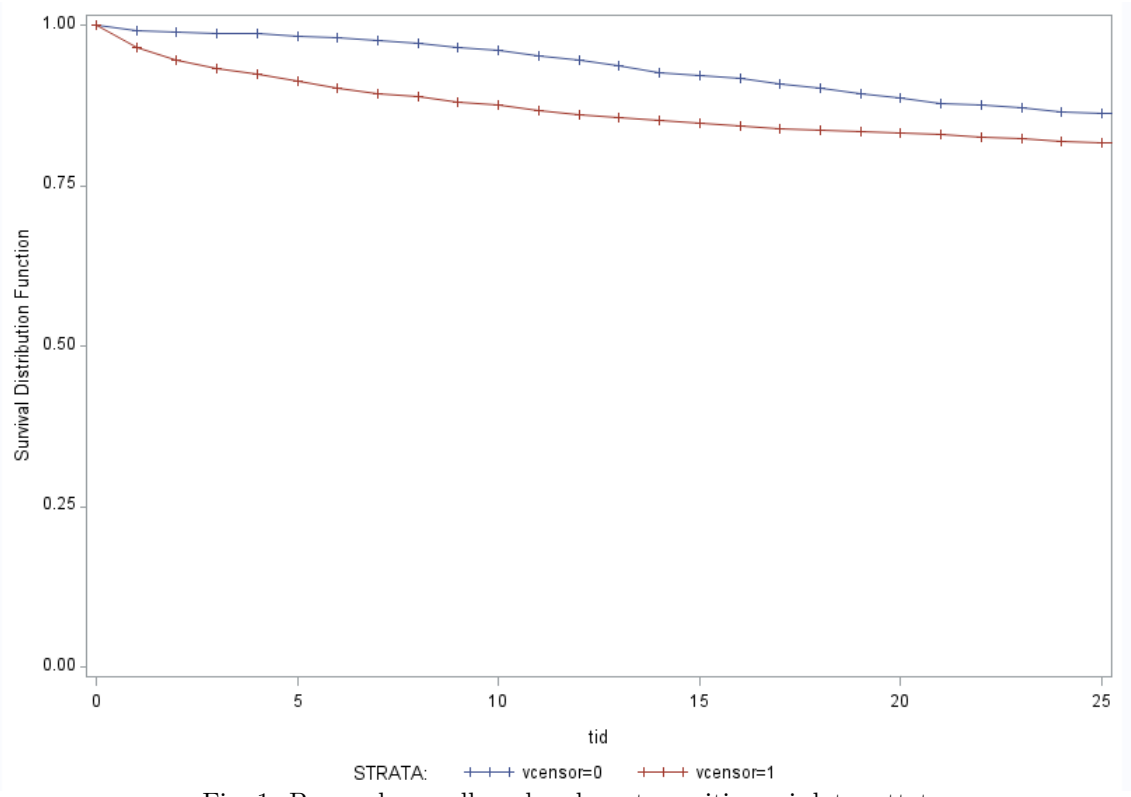


Fig. 1: Bevægelser mellem de relevante positioner i datasættet.

I figur 1 er den blå kurve estimeret ud fra forløb, der ikke er venstrecensurerede[4], og til den røde kurve er der udelukkende brugt forløb, hvor personen ikke allerede er selvstændig og ikke er 19 år gammel endnu. Baseret på figuren tyder det på, at forskellen er størst i de første år. Dette kan skyldes, at gennemsnitsalderen er væsentligt lavere ved applikation af den nævne venstrecensurering, hvormed mange af personerne stadig er igang med uddannelse. Hvor cut-off punktet vælges, vil have stor indflydelse på resultaterne af undersøgelsen. Derfor, samt at problemformuleringen er bred, og den væsentlige reduktion af observationer, vurderes det at analysen anvender de ucensurerede forløb. Analysen udledes derfor velvidende om, at der kan være bias, grundet forløb der starter før det angives i datasættet. Der tælles derfor tid på samtlige personer i datasæt-

[4]Hvis det antages at populationen er personer i alderen 18-43.

tet, som ikke allerede er selvstændige.

Selvstændige mv. på orlov inddrages som selvstændige eftersom datasættet opdateres årligt. Her er hændelsen i princippet interval censoreret, dvs. det vides ikke præcis hvornår den indtræffer, men blot at den indtræffer efter målingen et år, og før målingen det næste år. Af denne grund er det muligt, at et individ oplever hændelsen, og herefter går på orlov, hvormed de placeres i kategorien *pstill = 1*. Personen har stadig oplevelset hændelsen, og det skal derfor tælles som et *failure* for forløbet.

4 Analyse

Et vigtigt element af overlevelsesanalyse er at identificere, hvorledes det anvendte data er censureret. I modsætning til naturvidenskab, vil samfundsvidenskab sjældent have fuldkomne datasæt, og det er derfor nødvendigt at kunne bearbejde ufuldkomment data. Datasættet indeholder mange personer under risiko, som udgår fra datasættet inden en hændelse indtræffer. Et overblik over, hvordan individer kan forlade populationen under risiko, kan ses i figur 2.

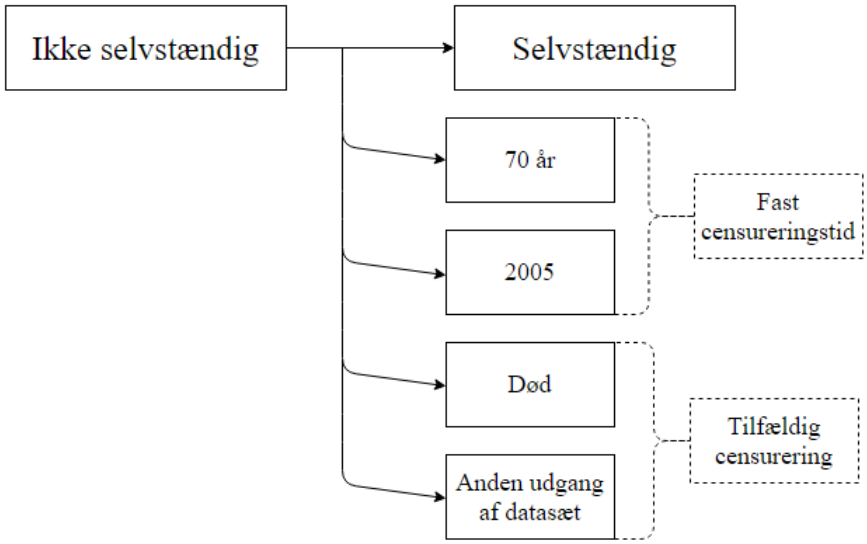


Fig. 2: Bevægelser mellem de relevante positioner i datasættet.

I figur 2 repræsenteres populationen under risiko af de *ikke selvstændige*. Disse kan, i teorien, til enhver tid *risikere* at blive selvstændige, hvormed hændelsen indtræffer. Hvis dette ikke er tilfældet, kan de udsættes for type I højrecensurering, dvs. hvor der er en fast censureringstid. I det gældende datasæt sker dette, når personen bliver 70 år gammel, eller i år 2005. Baseret på den information, der er tilgængelig, er disse personer stadig under risiko for at opleve hændelsen, men det er ikke muligt at vide, hvornår hændelsen sker, og hvis den overhovedet gør dette. Type I censur påvirker ikke maximum likelihood metoden med væsentlig bias, og er derfor ikke problematisk. Tilfældig censur kan derimod skabe problemer med estimationsbias, hvis denne er informativ. Hvis personer der frameldes cpr-registreret, og derfor ikke indgår i datasættet

længere, generelt har anden sandsynlighed for at blive selvstændig, end de der ikke gør, ville dette kunne forårsage biased resultater (Allison, 2010, s. 12-14). Det antages, at den tilfældige censurering i datasættet ikke er informativ, og at opgavens parameterestimationer derfor ikke påvirkes væsentligt af dette.

4.1 Overlevelsesfunktionen

Overlevelsesfunktionen for populationen er en sandsynlighedsfordeling, der viser sandsynligheden for at *overleve* til en periode senere end periode t, hvor tiden typisk figureres på x-aksen. Formelt kan overlevelsesfunktionen udledes via variablens kumulative fordelingsfunktion, $F(t)$, hvilken angiver sandsynligheden, ved et givent tidspunkt, for at variablen, T, tager en værdi, der er mindre eller lig t. Overlevelsesfunktionen findes her ved $1 - F(t)$, og udtrykker derfor sandsynligheden for, at T tager en værdi højere end t (ibid. s. 15). Her anvendes T som eventtidspunkt, således sandsynligheden, der aflæses på y-aksen, angiver sandsynligheden for at personen ikke er blevet selvstændig i tidspunkt $t + 1$. Figur 3 afbilder overlevelsesfunktionen for datasættets personer, der ikke er selvstændige.

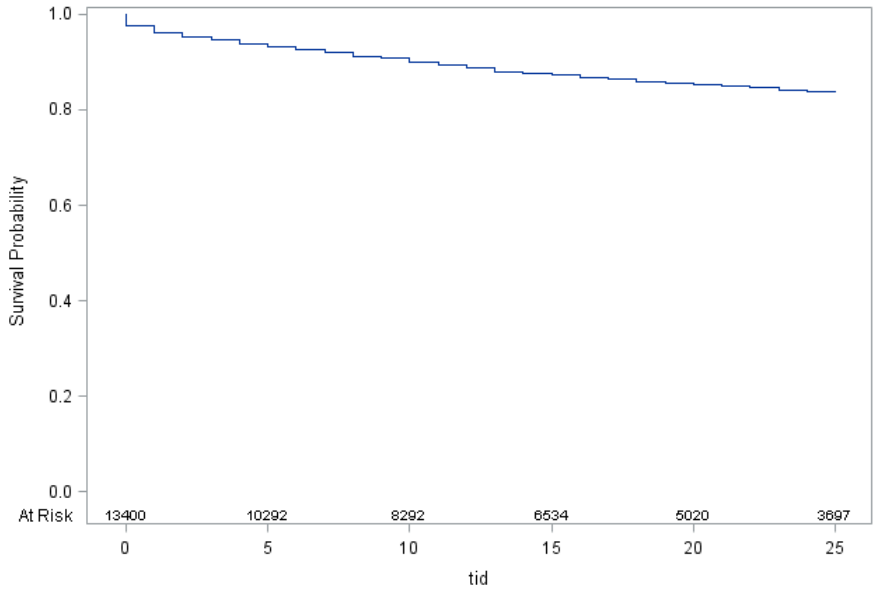


Fig. 3: Overlevelsesfunktionen for populationen under risiko.

Over figurens x-akse, ses antallet af individer i risikosættet til forskellige tidspunkter. Det ses at overlevelsesfunktionens niveau generelt er højt. Dette skyldes at mange i populationen aldrig oplever hændelsen, hvormed sandsynligheden for at *overleve* estimeres til at være højt.

Til at estimere sandsynlighedsfordelingen anvendes aktuarmetoden. Her er det ikke et krav med et præcist hændelsestidspunkt, hvilket er passende for datasættets hændelsesinterval. Ydermere indgår de censurerede værdier i beregningen, og metoden er bedre egnet til datasæt med mange observationer. Overlevelsesfunk-

tionen kan med aktuarmetoden estimeres ved

$$\hat{S}(t) = \prod_{j=1}^{i-1} (1 - q_j) \tag{1}$$

hvor $q_j = \frac{d_j}{n_j}$. q_j angiver et estimat af sandsynligheden for *failure* i et tidsinterval, såfremt der overleves til starten af intervallet. d_j angiver antallet af hændelser i intervallet, og n_j er populationen under risiko ved starten af intervallet, fratrukket halvdelen af antallet af censurerede observationer i det gældende tidsinterval (ibid. s. 51-52). En periode med mange censureringer vil derfor, alt andet lige, resultere i en mindre nævner, og eftersom tælleren ikke påvirkes af censureringerne, tager q_j en værdi tættere på 0. Sandsynligheden for at overleve i dette interval vil derfor være højere, alt andet lige. Produktfunktionen i ligning 1, ganger overlevelsessandsynlighederne fra tidligere tidsintervaller sammen. Resultatet af dette er den estimerede sandsynlighed for at overleve til t eller længere, dvs. $\hat{S}(t)$.

En anden måde at beskrive risikoen for populationen over tid er hazardfunktionen. I kontinuerlig tid kan denne gives ved

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{2}$$

I kontinuerlig tid er sandsynligheden for at hændelsen indtræffer til præcis tidspunkt t lig 0. Derfor anvendes tidsintervaller af en størrelse så tæt på 0, som det er muligt. Tælleren af ligning 2 udtrykker sandsynligheden for, at hændelsen sker i tidsintervallet $[t, t + \Delta]$, såfremt hændelsen ikke er sket før t. Selve hazardfunktionen kan ikke fortolkes som en sandsynlighed, eftersom størrelsen på nævneren i funktion 2, kan medføre værdier der er større end 100 %. Hazardfunktionens værdier, dvs. hazardraten til hvert tidspunkt, angiver antallet af hændelser per tidsinterval (ibid. s. 16-17). Hazardfunktionen er per definition uobserveret, og det er derfor nødvendigt at estimere den. Hertil anvendes igen aktuarmetoden, der estimerer den i'te persons hazardfunktion ved

$$h(t_{im}) = \frac{d_i}{b_i \left(n_i - \frac{w_i}{2} - \frac{d_i}{2}\right)} \tag{3}$$

hvor t_m er midten af interval t, d henviser til antallet af hændelser, b er bredden på intervallet, og w er antallet af individer, der censureres i intervallet (ibid. s. 52). Her anvendes bredden på tidsintervaller fra datasættet, dvs. et år. Således estimeres hazardfunktionen, der ses i figur 4.

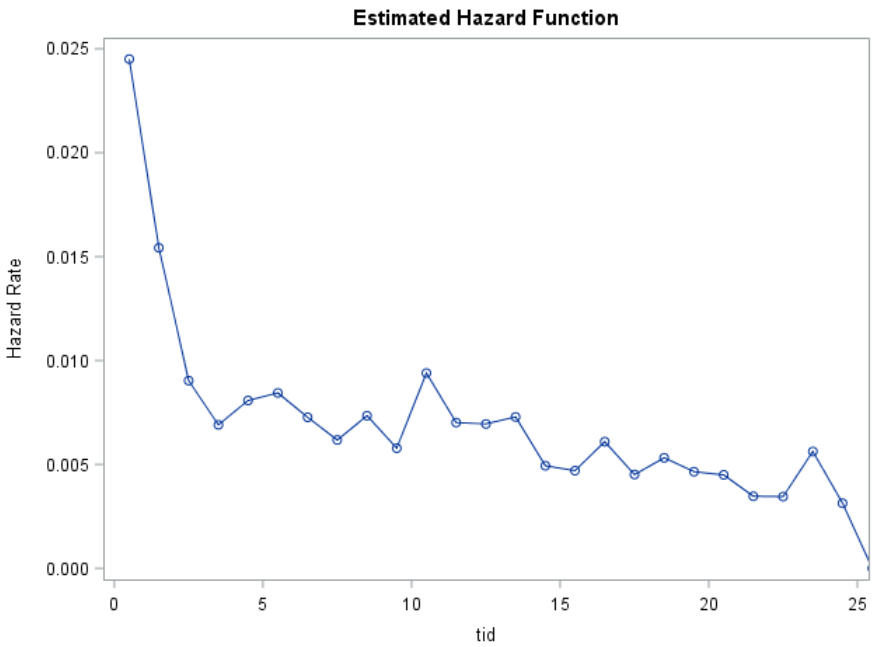


Fig. 4: Hazardfunktion for populationen under risiko.

Hazardraten illustrerer tydligere, at populationens risiko for at opleve hændelsen er en smule højere i de første perioder. Her kan hazardraten for første periode, aflæses til at være lige under 2.5%. Dette indikerer, at 2.5% af populationen under risiko oplever hændelsen indenfor første år. Hazardfunktionen tydeliggør også, at risikoen for at opleve hændelsen har en negativ sammenhæng med tiden. Dette kan være forårsaget af at andelen af populationen, der *overlever* de første perioder, aldrig oplever hændelsen.

4.2 Cox regression

Cox regressionsmetoden er anvendelige til at undersøge, hvad der har indflydelse på overlevestiden i forskellige grupper. Den er derfor relevant at inddrage i forbindelse med opgavens hypotesetests. Cox regressioner er mere robuste, sammenlignet med ventetidsmodeller[5], eftersom det ikke er nødvendigt at specificere, hvilken sandsynlighedsfordeling baseline hazardfunktionen, $\lambda_0(t)$, følger. Denne robusthed kommer på bekostning af statistisk præcision, eftersom der, ved at vælge den passende sandsynlighedsfordeling, kan findes et bedre fit på data. Typisk anvendes Cox regression, når sandsynlighedsfordelingen ikke kendes, da det her er bedre at anvende en robust model. Udover dette kan der anvendes højrecensureret data, og gøres brug af tidsvarierende variable.

Hazardfunktionen med k tids uafhængige kovariater, x_j, for det i'te individ, kan skrives op ved

$$h_i(t) = \lambda_0(t) \exp(\beta_1 x_{i1} + ... + \beta_k x_{ik}) \tag{4}$$

[5]Opgaven anvender ikke ventetidsmodeller, hvorfor disse ikke vurderes relevant at redegøre for.

hvor β_j angiver den j'te kovariats parameterværdi. Det ses fra ligningen, at når alle kovariater tager værdien nul, vil hazardfunktionen være lig baseline hazardfunktionen, hvoraf navnet. Denne baseline hazardfunktion indeholder også modellens skæringspunkt med y-aksen og fejlled.

Regressionsmetoden antager, at der altid er proportionalitet imellem to valgfrie individer i populationens hazardrater, dvs. der må kun være forskel i funktionernes niveauer. Dette kan også udtrykkes ved at udregne ratioen mellem hazardraten for det i'te og det j'te individ

$$\frac{h_i(t)}{h_j(t)} = \exp\Big\{\beta_1(x_{i1} - x_j) + ... + \beta_k(x_{ik} - x_k)\Big\} \tag{5}$$

Her ses det at baseline hazardfunktionen forsvinder, hvormed ratioen imellem hazardraterne er konstant over tid (ibid. s. 127-128). I forhold til analytisk arbejde er det ofte interessant at undersøge ratioen mellem hazardrater. Antaget proportionale hazardrater, udtrykker hazardratioen niveauet af den sammenlignte hazardrate, i forhold til referencegruppens hazardrate. F.eks. hvis risikoratioen er 0.7, har gruppen der stilles op imod referencegruppen, en hazardrate der er 30% lavere i niveau end referencegruppen. Her kendes de faktiske niveauer på de to funktioner ikke, men blot forholdet imellem dem.

I forbindelse med H_3, er det interessant at undersøge proportionalitetsantagelsen. Dette kan gøres ved at lave en **lifereg**, stratificeret for køn, i SAS. Her testes H_0, at der ingen forskel er mellem overlevelsesfunktioner. Til at teste antagelsen, anvendes test statistikkerne Log-Rank og Wilcoxon[6]. For den i'te gruppe gives Log-rank-statistikken ved

$$\sum_{j=1}^{r}(d_{ij} - e_{ij}) \tag{6}$$

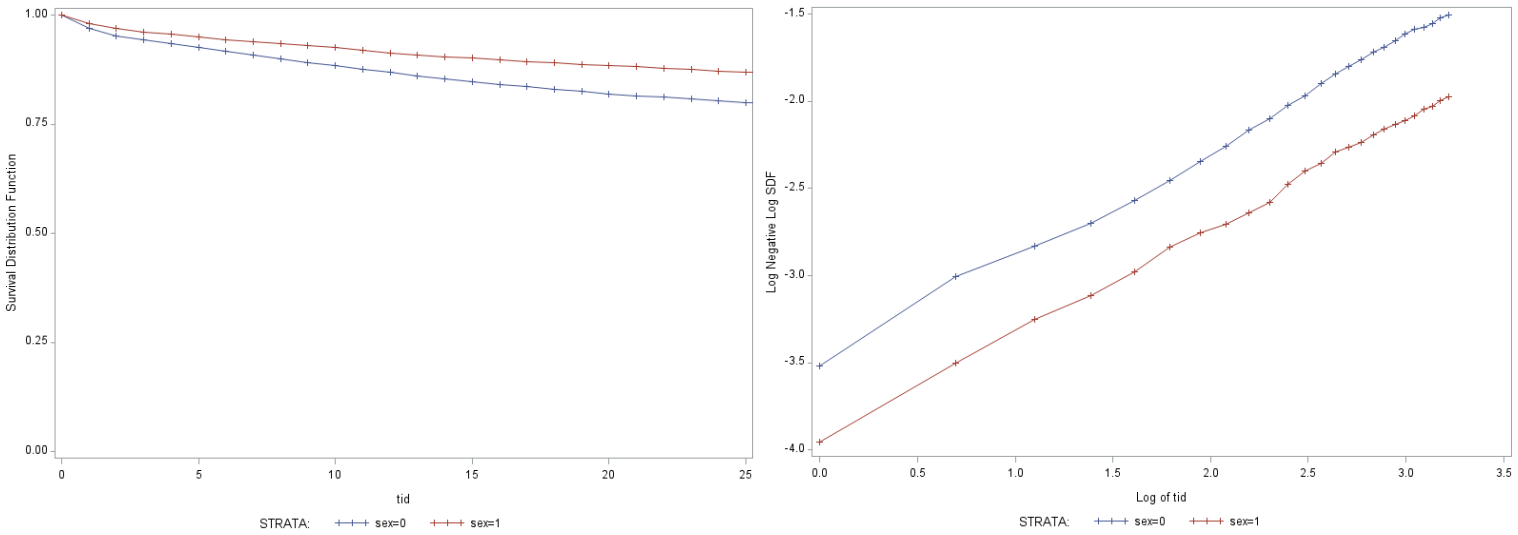
hvor r angiver antallet af unikke hændelsestidspunkter over begge grupper, d_{ij} er antallet af døde i tidspunkt j, og e_{ij} er det forventede antal hændelser på tidspunkt j. Det forventede antal hændelser er et gennemsnit udregnet i perioden før tidspunkt j. Test statistikken er derfor en sum af forskellen mellem de forventede antal hændelser, og det observerede antal hændelser. Wilcoxon-statistikken afviger fra denne formulering, eftersom den kan gives ved

$$\sum_{j=1}^{r} n_j(d_{ij} - e_{ij}) \tag{7}$$

hvor n_j er det samlede antal under risiko til tidspunkt j. Såfremt der er hændelser i datasættet, vil n_j være aftagende med tiden. Wilcoxon-statistikken vægter derfor tidligere observationer højere end de senere, hvormed forskelle i grupperne, ved senere observationstidspunkter, har mindre indflydelse på selve teststatistikken (ibid. s. 41-42).

Overlevelsesfunktionerne for populationen under risiko, stratificeret for køn, kan ses i figur 5a.

[6]SAS rapporterer også test-statistikken for likelihood-ratio testen, men den rapporteres ikke, eftersom denne kræver antagelsen at hazardfunktionen er konstant i hver gruppe, hvilket ikke er tilfældet.



(a) Overlevelsesfunktionen for mænd og kvinder.

(b) LLS plot af overlevelsesfunktioner.

Fig. 5: Visuel test for proportionalitetsantagelse.

Her noteres det, at funktionerne i figur 5a divergerer langs x-aksen, og med en væsentlig hastighed givet størrelsesordenen. Et andet plot der er relevant at inddrage her, er log af negative log overlevelsesfunktionen. Denne ses i figur 5b. Her burde afstanden også være konstant over tid, hvilket den ikke synes at være.
En mere konkret test er at anskue de nævnte teststatistikker. Teststatistikkerne, der tester H_0, at de to figurer er lig hinanden, kan ses i tabel 1.

Test	χ^2	DF	p-værdi
Log-Rank	83.1291	1	< 0.0001
Wilcoxon	74.4209	1	< 0.0001

Tabel 1: Teststatistikker for test af lighed af køns overlevelsesfunktioner.

Wilcoxon-statistikken udregnes til at være lavere en Log-rank-statistikken, hvilket skyldes, at der er mindre forskel på de to funktioner i de tidlige observationer, end der er i de sene. χ^2-værdierne er for ekstreme, for en χ^2-fordeling med én frihedsgrad, ved de to tests, til at godkende H_0, og det forkastes med næsten al sandsynlighed, at overlevelsesfunktionen er den samme for mænd og kvinder. Eftersom det er usandsynligt, at de to kurver har samme hældning over tid, brydes proportionalitetsantagelsen.

4.2.1 Ikke proportionale Cox regressionsmodeller

Et brud af proportionalitetsantagelsen er ensbetydende med, at kovariater interagerer over tid, dvs. de ikke er uafhængige af hinanden. Dette implicerer også at hazard ratioen er afhængig af tiden. Der anskues en periode på 25 år, og det er derfor de fleste kovariater, der vil ændres over tid. Derfor giver det mest mening at anvende disse som tidsvariende variable, hvormed hazard ratioen ikke kan holdes konstant.

Baseret på testen fra forrige afsnit, samt inklusionen af tidsvarierende variable, må antagelsen om proportionale hazardratioer være brudt. Estimater fra en Cox regressionen, uden proportionale hazardrater, fortolkes som gennemsnitlige effekter, hvilket kan være problematisk, hvis der er stærke tids-afhængige effekter. Den gennemsnitlige effekt kan medføre, at man underestimerer på nogle tidspunkter og overestimerer på andre, hvilket resulterer i et dårligt fit.

En test for om hazard ratioen er afhængig af tid, er at inddrage en lineær fremskrivning af variablens effekt. Dette gøres under antagelsen, at tidseffekten er lineær over tid. Til eksemplet anvendes ligning 4, hvor logaritmen er taget på begge sider af lighedstegnet, og der er en enkelt kovariat, for simplicitet. Log hazardfunktionen gives derfor ved

$$logh(t) = \alpha(t) + (\beta_1 + \beta_2 t)x \tag{8}$$

hvor $\alpha(t)$ angiver log af baseline hazardfunktionen. Parametren β_1 angiver derfor udgangspunktet, dvs. effekten af kovariaten x til tidspunkt 0, og β_2 angiver effekten af variablen over tid. Hvis estimatet af β_2 er signifikant, kan det konkluderes at kovariaten har en tidsafhængig effekt på hazardfunktionen. Selve effekten af kovariaten skal derfor ses som en sum af dens effekt i tid 0 og den tidsafhængige effekt. Denne afhænger af størrelsen på β_2 og fortegnet (ibid. s. 177-178). Eftersom småbørn og parforhold allerede er tidsvarierende variable, testes kovariaterne, der angiver alderen ved starten af forløbet, dvs. *under30*, *mellem30og50* og *over50*. Dette resulterer i signifikante estimater af deres tidsafhængige effekt, hvorfor tidsinteraktionseffekterne inkluderes i den endelige model i afsnit 4.5, iform af *under30tid* og *over50tid*.

4.3 Gentagne hændelser

Et problem med gentagne hændelser er, at der kan være ikke observeret heterogenitet. Dette kan stamme fra en uobserveret faktor, der medfører, at de enkelte observationer ikke er uafhængige af hinanden. Dette kunne være en *entrepeneural spirit*, som nogle af individerne besidder, der gør at de hurtigere, og måske oftere, oplever hændelsen end andre. Hvorvidt dette er gældende undersøges ved en **lifetest**, der sammenligner personer i deres første forløb og personer i deres andet, dvs. stratificeret for *sekvens(1,2)*. Dette kan ses i figur 6, hvor den blå overlevelsesfunktion er estimeret fra individer i populationen, der maks oplever én (observeret) hændelse, og den røde er fra individer, der allerede har oplevet hændelsen en gang.

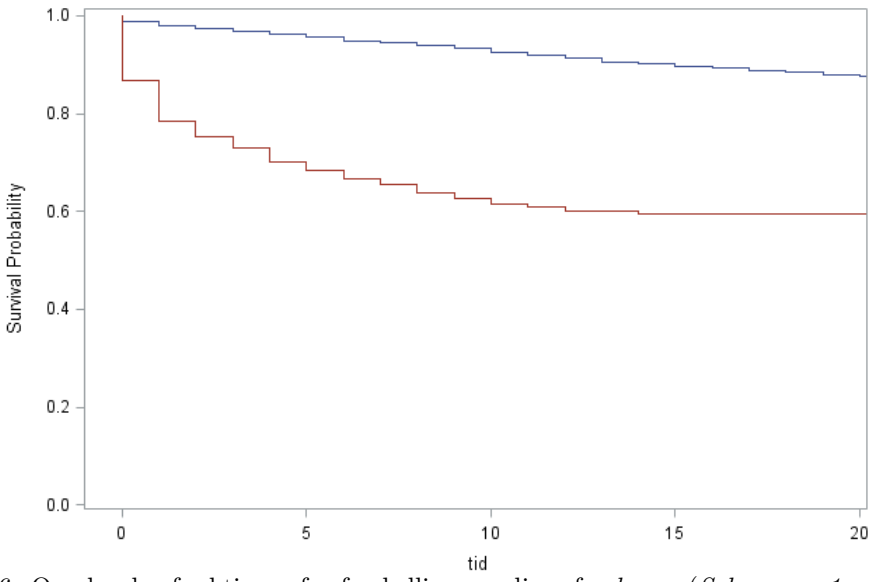


Fig. 6: Overlevelsesfunktioner for forskellige værdier af *sekvens* (*Sekvens = 1* er blå).

Til figuren anvendes **maxtime** = 20, eftersom individer med flere hændelser per definition har kortere tid tilbage, før de udgår af datasættet. De to overlevelsesfunktioner er baseret på væsentligt forskellige populationsstørrelser[7], hvilket kan have en effekt på resultatet. Det ses, at individer der har oplevet hændelsen én gang, har væsentligt lavere sandsynlighed for at overleve, hvilket indikerer, at disse personer har uobserverede egenskaber.

En mere formel test af dette, er at inkludere *lagtid* som en forklarende variabel. Denne angiver længden af det forrige forløb, dvs. *sekvens*-1. Der testes for afhængighed mellem forløbene i en Cox regression med alle modellens forklarende variable, dvs. småbørn, parforhold, køn og alderskovariaterne (ibid. s. 264-265). Variablene inkluderes for at undersøge, om der er uforklaret afhængighed mellem tidspunkter i modellen, men deres parameterestimater og hazardratioer, er ikke vigtige for testen. Testværdierne kan ses i tabel 2.

Parameter	Parameter estimat	Standardfejl	P-værdi
Lagtid	-0.0469	16.1685	< 0.0001

Tabel 2: Resultat af test for uafhængighed mellem forskellige værdier af *sekvens*.

Her ses det at estimatet af *Lagtid*, er signifikant anderledes fra 0, med næsten al sandsynlighed, hvormed det **ikke** kan afvises at længden af forløb 2, er uafhængig af længden af forløb 1.

Uafhængigheden mellem forløbslængder indenfor samme person, kan håndteres med forskellige fremgangsmåder. Eftersom forskningsspørgsmålet, der ønskes besvaret, er hvorfor individer *starter* som selvstændig, vurderes det passende at udelade senere forløb. Her er det første forløb særligt relevant, da individer der er i deres

[7] 12175 individer med *sekvens=1* og 924 individer med *sekvens=2*. Dette forværres kun, hvis højere værdier af *sekvens* anvendes, hvorfor dette udelades.

andet forløb, allerede har startet som selvstændig. Eksklusionen af *sekvens* $>$ *1* medfører, at antallet af hændelser falder til 1191, hvilke analysen arbejder videre med.

4.4 Forventninger til den endelige model

Variabel	Referencegruppe	Parameter estimat	Hazard ratio
Sex	Mænd	-	< 1
Smaborn	Uden børn	-	< 1
Pfh	Enlige	-	< 1
Under30	*Mellem30og50*	-	< 1
Under30tid	*Mellem30og50*	+	
Over50	*Mellem30og50*	-	< 1
Over50tid	*Mellem30og50*	+	

Tabel 3: Skema over forventede effekter af kovariater.

Tabel 3 viser de forventede effekter af variable i analysen, baseret på hypoteserne i afsnit 1. Negative (positive) parameterestimater er ensbetydende med hazard ratioer under (over) 1, eftersom hazardratioen, for det enkelte estimat, kan udregnes ved $\exp(\beta)$.

4.5 Den endelige regressionsmodel

Baseret på ovenstående afsnit, kan modellen af hazardfunktionen, for det i'te individ i populationen under risiko til tid t, opstilles på log form, som

$$logh_i(t) = \alpha(t) + \beta_1 sex_i + \beta_2 smaborn_i(t) + \beta_3 pfh_i(t) + (\beta_4 + \beta_5 t)under30_i + (\beta_6 + \beta_7 t)over50_i \qquad (9)$$

Denne genereres ved udelukkende at anvende individer, der ikke er observeret som værende selvstændige endnu, og tælle tid på deres forløb. Resultaterne fra Cox regressionen, med ikke proportionale hazard rater, kan ses i tabel 4.

Parameter	DF	Estimat	Std. fejl	χ^2	P-værdi	Hazard ratio
Småbørn	1	0.3653	0.0798	20.9748	<0.0001	1.441
Parforhold	1	0.2594	0.0757	11.7501	0.0006	1.296
Køn	1	-0.4750	0.0619	58.9385	<0.0001	0.622
Under 30	1	-1.6290	0.1197	185.1732	<0.0001	0.196
Under 30 tid	1	0.1587	0.0128	152.9638	<0.0001	1.172
Over 50	1	-0.0732	0.1380	0.2814	0.5958	0.929
Over 50 tid	1	-0.1119	0.0306	13.3699	0.0003	0.894

Tabel 4: Beskrivende statistikker for endelige Cox regression.

Resultaterne i tabel 4, anvendes i følgende afsnit til at diskutere hypotese 1-4 og de forventede effekter.

Hypotese 1

Estimatet for *smaborn* er positivt, mod forventningen. Derfor bliver hazardratioen også over 1, hvormed modellen forventer en 44% højere gennemsnitlig hazardrate[8], sammenlignet med referencegruppen, dvs. personer uden småbørn. Estimatet er højsignifikant, hvormed H_1 afvises. Det forventede resultat var, at personer med småbørn ikke havde tid til at starte op som selvstændige, men det er muligt, at individer med småbørn har forøget risici for at opleve hændelsen, da de netop gerne vil have mere tid derhjemme. Som selvstændig vil de bedre kunne tilrettelægge deres arbejdsdag for at tage hensyn til barnet, således det ikke er nødvendigt med børnepasning. Ideen om at være fri og selvstændig kan virke bedre, når man allerede går hjemme fra arbejdet, og gerne vil bruge tid med sit barn, hvilket måske er hvorfor flere bliver selvstændige. Det er også en mulighed, at det skyldes mellem-person variation, dvs. at der er forskelle mellem de to grupper i kovariaten. Hvis der anvendes mere end en hændelse per person, kan det undersøges ved f.eks. en fixed effects model, hvor der tilføjes et fejlled, der absorberer den uobserverede heterogenitet imellem personer.
Ved at omkode *smaborn*, til at kun registrere når personen har et barn i 0-2 års alderen, bliver variablen væsentligt mindre signifikant, og er her kun anderledes fra 0 på et 5% signifikansniveau. Effekten er stadig positiv, men det er muligt, at den forventede effekt indefinder sig i et meget kortere tidsinterval. Det er måske primært iløbet af barnets første år, at småbørn har en lempelig effekt på risikoen for at blive selvstændig, hvorefter det bidrager positivt, hvilket kan være forårsaget af en selektionseffekt.

Hypotese 2

Igen afviger resultatet fra forventningerne og hypotesen, eftersom estimatet og hazardratioen er positiv. Estimatet er meget signifikant på et 0.1% signifikantniveau, og hazardratioen angiver at den gennemsnitlige hazardrate for personer i et parforhold er 29.6% højere end personer, der ikke er i et parforhold. Dette

[8]Eftersom PH antagelsen er brudt.

betyder, at H_2 afvises, og at den alternative hypotese, at personer i et parforhold har større risiko for at begynde som selvstændig, accepteres. Her er det muligt, at en partner kan bidrage med økonomisk stabilitet, mens at en person starter som selvstændig, ved at partneren fortsætter med at arbejde, og at det derfor er nemmere at starte op.

Hypotese 3

Resultatet for køn-kovariaten passer med det forventede, eftersom dennes parameterestimat er negativ og højsignifikant. Der estimeres en hazardratio på 0.622, hvilket indikerer, at hazardraten for kvinder i gennemsnit er 62.2% af mændenes hazardrate, dvs. 37.8% lavere. Grundet dette, og ikke-proportionaliteten mellem overlevelsesfunktionerne for mænd og kvinder nævnt i afsnit 4.2, kan H_3 ikke afvises.

Hypotese 4

Estimaterne af dummy variablene for alder stemmer overens med forventningerne, pånær variablen der angiver tidsinteraktionen for *over50*. Både *under30* og *under30tid* er signifikant anderledes fra nul med næsten al sandsynlighed. Estimatet af *Under30* kovariatens hazardratio angiver, at hazardraten for denne gruppe er meget lav ift. referencegruppen, men *under30tid* indikerer, at hazardraten konvergerer mod referencegruppen over tid. En årsag til dette kan være, at *under30* indeholder data for personer i alderen 13-18, hvoraf nogle oplever hændelsen, men der er væsentligt lavere risiko for det. Når tiden øges, og derved alderen, kommer individet tættere på alderen i referencegruppen, hvor der er markant højere risiko for at opleve hændelsen. Estimatet af *over50* er insignifikant, hvilket indikerer at hazardraten i gennemsnit er på samme niveau som referencegruppen, men *over50tid* er negativ og højsignifikant, hvorfor hazardraten falder med tiden. Det dårlige fit, der findes til *over50*, kan være forårsaget af, at der stadig er mange, der begynder som selvstændig de første tidsperioder i *over50*, men at udviklingen i hazardfunktionen herefter ikke kan forklares godt af en lige linje.

Selvom referencegruppen ikke har passet perfekt på datasættet, vurderes det, at H_4 ikke kan afvises. Yderligere, hvis *over50* udelades fra regressionen, er *under30* stadig højsignifikant med et estimat på -1.51, og tidsinteraktionseffekten er også signifikant, og kun marginalt anderledes. Eksklusionen af *over50* ændrer referencegruppen til at være individer i en alder over 30, hvilket anvendes i følgende test.

Afslutningvist testes der for hypotesen, at interaktionseffekten mellem startalder og tid er 0 ved forskellige tidspunkter. Her testes tidsperioderne 1-20, hvoraf et udkast af resultaterne ses i tabel 5.

Variabelnavn	χ^2	DF	P-værdi
tidstest5	82.6336	1	< 0.0001
tidstest9	0.6776	1	0.4104
tidstest10	1.2515	1	0.2633
tidstest15	49.5088	1	< 0.0001
tidstest15	86.9068	1	< 0.0001

Tabel 5: Fem eksempler på test af interaktionseffekt over tid.

Baseret på denne test, tyder det på, at den positive interaktionseffekt er signifikant i næsten alle perioder, undtagen efter 9 og 10 år, hvilket kan skyldes at hazardraterne for *under30* og referencegruppen, er lig hinanden i disse perioder.

Opsummering

Samlet set indikerer regressionen at familiemæssige forhold, herunder hvorvidt en person har en partner eller at de har småbørn, har betydning for, at personen starter som selvstændig, og begge har positiv effekt på risikoen for at opleve hændelsen. Ydermere er det sandsynligt, at mænd har større risiko for at starte som selvstændige, og risikoen for hændelse er højere for individer over 30. Det er ikke muligt at se, hvornår hvert individs forløb er begyndt, hvormed det er sandsynligt, at risikoen for hændelse vurderes større end det er gældende for populationen, eftersom de observerede forløb er kortere end realiteten. Det er også en mulighed, at analysen tæller tid på personer, der allerede har oplevet hændelsen, hvilket også skævvrider resultaterne.

5 Litteraturliste

Allison P.D. (2010). *Survival Analysis using SAS: A Practical Guide*. SAS Institute and Wiley Interscience.

Egedesø, C. Hansen K. & Nielsen, P. (2018). *Iværksætteri i Danmark*. Danmarks Statistik. Hentet d. 8/06-2020 fra https://www.dst.dk/Site/Dst/Udgivelser/nyt/GetAnalyse.aspx?cid=31441

Kjempff, M. (17. november 2014). *Alder og erfaring kan være guld værd for iværksættere*. Hentet d. 9/06-2020 fra https://www.dr.dk/nyheder/regionale/oestjylland/alder-og-erfaring-kan-vaere-guld-vaerd-ivaerksaettere

6 Oversigt over nye variable

Variabelnavn	Type	Beskrivelse
pstart	Indikatorvariabel	Første periode personnr. optræder i datasættet = 1, ellers = 0.
pslut	Indikatorvariabel	Sidste periode personnr. optræder i datasættet = 1, ellers = 0.
selv	Tilstandsvariabel	Person angives som selvstændig (1) når *pstill* = 1 eller 11, når *pstart*$\neq$ 1, ellers = 0.
censor	Censureringsvariabel	Ingen censur = 0, person bliver 70 = 1, datastop i 2005 = 2, udvandring fra datasæt = 3, dør = 4.
vcensor	Censureringsvariabel	Ingen censur = 0, hvis *pstart* = 1 og alder $>$ 18 = 1, . hvis *pstart* = 1 og *selv* = 1 sættes *vcensor* = .
tid	Tidsvariabel	Tæller tid i hvert forløb.
event	Hændelsesvariabel	Hvis personen starter som selvstændig i næste periode =1, ellers = 0.
sekvens	Sekvensvariabel	Angiver forløbsgang 1 - 6.
lagtid	Tidsvariabel	Tid i forrige sekvens.
smboern1-26	Hjælpevariable	26 variable der tager værdien af **max**(boern02, boern36), i deres tilsvarende periode.
parf1-26	Hjælpevariable	26 variable der angiver forholdsstatus (0,1), i deres tilsvarende periode.
smaborn	Forklarende variabel	Array indeholdende *smboern1-smboern26*.
pfh	Forklarende variabel	Array indeholdende *parf1-parf26*.
sex	Forklarende variabel	Hvis kvinde = 1, ellers = 0.
under30	Forklarende variabel	Hvis personen er under 30 ved forløbets start = 1, ellers = 0.
mellem30og50	Forklarende variabel	Hvis personen er mellem 30 og 50 ved forløbets start = 1, ellers = 0.
over50	Forklarende variabel	Hvis personen er over 50 ved forløbets start = 1, ellers = 0.
under30tid	Forklarende variabel	Lineær interaktion med tid for personer under 30.
over50tid	Forklarende variabel	Lineær interaktion med tid for personer over 50.
$startalder_{gr}$	Hjælpevariabel	Tager værdien 1, hvis person er under 30 ved forløbets start, 2 hvis mellem 30 og 50 ved start, og 3 hvis over 50 ved forløbets start.