



Stockholms
universitet

Effekter av införande av könsneutral prissättning - en fallstudie

Sandra Brännstam

Kandidatuppsats 2015:1
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Effekter av införande av könsneutral prissättning - en fallstudie

Sandra Brännstam*

Juni 2015

Sammanfattning

Anledningen till denna studie är den lagstiftning om att riskpremi-
en inte får påverkas beroende på kön hos försäkringstagaren. Baserat
på data från Wasa försäkringar mellan åren 1994-1998 gör vi en fall-
studie av hur riskpremien (*främst den modellerade skadefrekvensen*)
förändras för män och kvinnor i olika åldrar då man bortser från kön
vid prissättning. Vi har i denna fallstudie kommit fram till att könsne-
utral prissättning leder till att kvinnor höjer sin premie i yngre åldrar
och sänker premien i äldre åldrar. Män ligger någorlunda lika innan
som efter lagstiftning, förutom att premien sänks något i yngre åldrar.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: sandra.brannstam@hotmail.com. Handledare: Martin Sköld, Tom Britton.

Tack

Jag vill inledelsevis tacka Martin Sköld och Tom Britton vid Matematiska institutionen, Stockholms Universitet, för förslag till frågeställning, handledning, och svar på ett otal frågor och funderingar.

Tack även till de kurskamrater jag har diskuterat teori med.

Innehållsförteckning

1	Introduktion	1
1.1	Bakgrund	1
1.2	Syfte	2
1.3	Beskrivning av data	2
2	Teori	5
2.1	Grunder inom prissättning av sakförsäkring	5
2.2	Prissättning med Generaliserade Linjära Modeller	6
2.3	Poissonfördelningen	7
2.3.1	Poissonfördelningen	7
2.3.2	Poisson i GLM	8
2.4	Övriga termer i Generaliserade Linjära Modeller	8
2.5	Splines	9
2.5.1	B-spline	10
2.5.2	Konfidsensband	12
3	Analys	12
3.1	Grafisk analys av skadefrekvens	12
3.2	Splinefunktion för variabeln Ägarålder	13
3.3	Modell	14
3.4	Observerade skillnader i skadefrekvens beroende på kön och ålder	14
3.5	Osäkerhet i predikterade värden	16
4	Slutsatser	19
4.1	Resultat	19
4.2	Förslag på förbättring av analys	19
A	Appendix	20
A.1	Figurer	20
	Referenser	24

1 Introduktion

1.1 Bakgrund

21 December 2012 infördes en lagstiftning i EU om att prissättning av försäkringskontrakt inte längre fick bero på kön. Anledningen till detta ansågs vara att risker som tidigare förknippats med kön i dagsläget inte kunde ges den förklaringen då kvinnor och män, vilket bland annat Daly (u.å) påpekar, lever mer lika livsstilar överlag.

Det har gjorts ett antal undersökningar i hur marknaden påverkats av dessa bestämmelser, vad det kan tänkas få för reaktioner av kunder, hur försäkringsbolagen påverkas och liknande.

Calafa och Bonardi (2014) diskuterar hur bestämmelsen påverkar försäkringsbolaget och kunderna. De tar upp både försäkringar som är obligatoriska (fordons/trafikförsäkringar), samt frivilliga försäkringar (främst livförsäkringar). De nämner även att unga män och kvinnor troligtvis påverkas mest då de förstnämnda är mer riskbenägna.

Schmeiser, Störmer, och Wagner (2014) diskuterar andra faktorer och försäkringstypers, påverkan av könsneutral premie, samt hur marknaden påverkas. De påpekar att en risk med könsneutral prissättning kan vara att prishöjningen för en lägre riskgrupp medför att de slutar teckna försäkringar, och att den högre riskgruppen då överrepresenteras vilket i det långa loppet leder till att basen för prissättningen endast, eller åtminstone till stor del, utgörs av den grupp som har högst förväntad risk. Priserna sätts utefter detta, och kan då bli missvisande. De påpekar även att dessa resultat främst påverkar valfria försäkringar, just fordonsförsäkringar som vi studerar i denna rapport är obligatoriska och påverkas alltså inte på samma sätt. För fordonsförsäkringar tror de däremot att detta kan leda till mer risktagande, vilket skulle leda till högre skadefrekvenser. De nämner även att det också kan leda till positiva reaktioner som att försäkringsbranschen ses som mer etisk och konsumentvänlig.

Sass och Seinfeld (2014) talar bland annat om påverkan av marknaden samt hur en neutral premie gynnar högrisk kunder och missgynnar lågrisk kunder.

Aseervatam, Lex, och Spindler (2014) går däremot in mer på hur premien för bil/motorförsäkring påverkas, vilket vi även studerar i denna rapport. De nämner att det är diskriminerande att prissätta baserat på kön, men att det inte riktigt framgår hur bestämmelserna påverkar förändring i priser. De gör en uppskattning av hur kön påverkar premiesättning i bilförsäkringsmarknaden. De nämner även att då en variabel avlägsnas får de andra kompensera för risken som tidigare förknippats med kön och påpekar som många andra att att kön är en variabel som möjligtvis ersätter andra faktorer, som beteende i trafiken, motorstyrka eller liknande.

1.2 Syfte

Syftet med denna studie är att teoretiskt undersöka hur införandet av lagstiftningen påverkar riskpremien för män och kvinnor i olika åldrar. Rent intuitivt kan tänkas att män i yngre åldrar har dragit nytta av detta, då de brukar beskrivas som mer riskbenägna och därför tidigare har haft en högre premie.

1.3 Beskrivning av data

Data som tillhandahållits kommer från Wasa försäkringar (och illustreras i Tabell 1). Det är 64548st kontrakt för motorcykelförsäkringar mellan åren 1994-1998, där 54695 är manliga försäkringstagare och 9853 är kvinnliga. Wasas tariff är baserad på fyra faktorer (Se Tabell 1), där en tariffcell utgörs av en klass från varje faktor. Se exempel i Tabell 2.

- Ägarålder: Försäkringstagarens ålder, mellan 0 och 99
- Kön: Försäkringstagarens kön, M (Man) eller K (Kvinna)
- Zon: Geografisk zon, standardklassificering av Sveriges alla kommuner/län/ städer, mellan 1 och 7.
- Mcklass: En klassificering av den så kallade EV-kvoten, vilken beskriver förhållandet mellan effekt och vikt, och definieras som $(\text{Motorstyrka i kW} \times 100)/(\text{Fordonsvikt i kg} + 75)$, avrundat nedåt till närmsta heltal. De 75 extrakilona i nämnaren representerar förarens vikt. Dessa kvoter är uppdelade i sju klasser, som visas i Tabell 1.
- Fordonsålder: Fordonsålder, mellan 0 och 99.
- Bonusklass: Bonusklass, antar värden mellan 1 och 7. En ny försäkringstagare startar med bonusklass 1. För varje skadefritt år ökar bonusklassen med 1, efter första skadan avtar bonusklassen med 2.
- Duration: Antal år som försäkringstagare.
- Antal skador: Antal skador försäkringstagaren har råkat ut för.
- Skadekostnad: Den sammanlagda kostnaden av skador för försäkringstagaren.

Vissa av kategorivariablerna har slagits ihop enligt Wasas tariff från 1995, tabell finnes på sida 36 i Ohlsson och Johansson (2010), utöver detta har vi slagit ihop zon 5-7 då de i våra regressionsmodeller inte har visat sig vara signifikant skilda från basklassen, dessutom har vi valt att bortse från försäkringskontrakt där försäkringstagaren är yngre än 16 år då det inte känns rimligt att någon under 16 år kör en motorcykel.

För att bestämma vilka klasser vi vill ha som bas aggregerar vi datamaterialet efter skadekostnad, antal skador, samt duration med avseende på de olika variablerna som ingår. Detta innebär alltså att skadekostnaden, antal skador, och durationen för varje klass inom en variabel har summerats. På

så sätt ser vi vilken klass inom varje variabel som har högst duration och kan välja den som basklass, alltså den klass vars koefficient är 0. Efter att åldersbegränsningarna införts har vi att göra med 62436 försäkringskontrakt, där 52954 är manliga försäkringstagare och 9482 är kvinnliga. Datamaterialet ser nu alltså ut som i Tabell 1, med basklasserna markerade.

Faktor	Klass	Beskrivning
Zon	1	Centrala delar av Stockholm, Göteborg, och Malmö
	2	Förorter och medelstora städer
	3	Mindre städer, bortsett från de i 5
	4	Småstäder och landsbygd
	5	Städer i norr, landsbygd i norr, och Gotland
MCKlass	1	EV-kvot: -5
	2	EV-kvot: 6-8
	3	EV-kvot: 9-12
	4	EV-kvot: 13-15
	5	EV-kvot: 16-19
	6	EV-kvot: 20-24
	7	EV-kvot: 25-
Fordonsålder	1	0-1 år
	2	2-4 år
	3	5- år
Bonusklass	1	1-2
	2	3-4
	3	5-7

Tabell 1: Databeskrivning.

Tariffcell				Skadekostnad	Antal skador	Duration
Zon	MCKlass	Fordonsålder	Bonusklass			
1	1	1	1	14533	2	12.4164
1	1	1	2	0	0	3.2603
1	1	1	3	15309	1	11.5699
1	1	2	1	42586	3	29.3808
1	1	2	2	1263	1	14.175341
⋮	⋮	⋮	⋮	⋮	⋮	⋮
5	7	3	2	0	0	0.9397
5	7	3	3	0	0	16.6438

Tabell 2: Exempel på tariffceller.

Efter inledande analys upptäcktes att koefficientskattningarna för bonusklass gick i motsatt riktning än förväntat. Alltså att koefficientskattningarna ökade då man gick upp en bonusklass. Annars är ju tanken att kostnaderna minskar då man gått en längre tid utan skador. En åtgärd kan vara att utesluta den faktorn ur regressionsmodellen och prissätta den separat, alternativt att använda sig av den tariff som redan finns i tabell 2.8 sid 36 i Ohlsson och Johansson (2010) . Vi väljer i denna rapport att gå vidare genom att utesluta bonusklass i regressionsmodellen. Tariffcellerna ser då istället ut som i Tabell 3.

Tariffcell			Skadekostnad	Antal skador	Duration
Zon	MCklass	Fordonsålder			
1	1	1	29842	3	27.2466
1	1	2	43849	4	74.0466
1	1	3	58380	7	451.2301
1	2	1	70746	1	53.6329
1	2	2	3865	1	71.6849
⋮	⋮	⋮	⋮	⋮	⋮
5	7	2	10953	2	4.0767
5	7	3	0	0	26.5562

Tabell 3: Exempel på tariffceller, utan bonusklass.

Då frågeställningen är att undersöka hur riskpremien, eller skadefrekvensen, förändrades för män och kvinnor då man bortser från kön vid prissättning kommer vi att arbeta med, och jämföra, två modeller genom rapporten. En modell där kön inte alls är med som förklarande variabel, samt en modell där både kön och en samspelsterm mellan kön och ålder är med. I regressionsmodellerna betecknas variablerna som i Tabell 4 nedan.

Variabel	Beteckning	Förklaring
Ägarålder	x_1	Kontinuerlig variabel $x_1 \in [16, 92]$
Zon	x_2	Faktorvariabel i 5 nivåer.
MC-klass	x_3	Faktorvariabel i 7 nivåer.
Fordonsålder	x_4	Faktorvariabel i 3 nivåer.
Kön	x_5	Faktorvariabel i 2 nivåer.

Tabell 4: Beteckning av variabler.

Variablerna $x_2 - x_5$ kan även beskrivas som dummyvektorer, en vektor med 1 på den plats som motsvarar den klass variabeln befinner sig i för givet försäkringskontrakt, och 0 för övriga klasser. Befinner sig variabeln i

sin basklass innehåller vektorn endast 0:or. Vektorerna förklaras ytterligare i Tabell 5 nedan.

Faktor	Dummyvektor
x_2	$(Zon_1, Zon_2, Zon_3, Zon_5)$.
x_3	$(Klass_1, Klass_2, Klass_4, Klass_5, Klass_6, Klass_7)$
x_4	$(Fordonsålder_1, Fordonsålder_2)$
x_5	$(Kvinna)$

Tabell 5: Beteckning av variabler.

Låt säga att ett försäkringskontrakt tillhör en man, Zon 3, MC-klass 6, Fordonsålder 1. I detta exempel skulle $x_2 - x_5$ se ut på följande sätt

$$x_2 = (0, 0, 1, 0)$$

$$x_3 = (0, 0, 0, 0, 1, 0)$$

$$x_4 = (1, 0)$$

$$x_5 = (0)$$

Som vi nämnt ovan består den ena modellen av variablerna $x_1 - x_4$ och den andra modellen av $x_1 - x_5$.

2 Teori

Teorin kommer i huvudsak från Ohlsson och Johansson (2010): Non-Life Insurance Pricing with Generalized Linear Models.

2.1 Grunder inom prissättning av sakförsäkring

För varje försäkringskontrakt bestäms premien utifrån värdet på ett antal variabler, prissättningsfaktorer. Prissättningsfaktorerna tillhör generellt någon av följande kategorier:

- Försäkringstagarens egenskaper: Ålder eller kön hos försäkringstagaren om denne är en privatperson, branch om det är ett företag, o.s.v.,
- Det försäkrade objektets egenskaper: Ålder eller modell hos ett fordon, typ av byggnad, etc,
- Det geografiska områdets egenskaper: Inkomst eller populationstäthet i försäkringstagarens bostadsområde, o.s.v.

En skada definieras som en händelse försäkringstagaren har råkat ut för, för vilken denne har begärt ekonomisk kompensation.

Skadefrekvensen är i vårt fall antalet skador dividerat med durationen.

Ibland kan det gälla en specific tidsperiod. Dvs medelvärde av skador per tidsperiod (vanligtvis ett år). Hossack, Polland och Zehnworth (1983) påpekar att täljaren oftast är antal skador, medan nämnaren kan variera beroende på vad man vill mäta. Ofta använder man en enhet som är mer praktisk och genomförbar framför en som kan kännas mer logisk eller teoretiskt önskvärd, men som kanske inte fungerar i praktiken. Medelskadan är den totala skadekostnaden dividerad med antal skador, det vill säga medelkostnaden per skada. Vi bortser från så kallade nollskador, skador som inte ger någon ersättning från försäkringsgivaren, i skadefrekvensen och medelskadan. Det är naturligtvis viktigt att vara konsekvent, och antingen utesluta eller inkludera dessa i båda fallen (skadefrekvens och medelskada). Riskpremien är den totala skadekostnaden dividerat med durationen, alltså medelkostnaden per år (eller annan tidsperiod). Följaktligen är premien produkten av skadefrekvensen och medelskadan.

$$\begin{aligned} \text{Premie} &= \frac{\text{skadekostnad}}{\text{duration}} = \frac{\text{antal skador}}{\text{duration}} \cdot \frac{\text{skadekostnad}}{\text{antal skador}} = \\ &= \text{skadefrekvens} \cdot \text{medelskada} \end{aligned}$$

Premien, skadefrekvensen, och medelskadan ovan kallas för nyckeltal, se även Tabell 6.

Hossack, Polland, och Zehnworth (1983) påpekar hur viktigt det är att skadefrekvens och medelskada betraktas som två separata faktorer. Anledningen till detta är att vissa faktorer påverkar skadefrekvensen och medelskadan på olika sätt. Till exempel kan inflation påverka medelskadan, men det är inte lika troligt att det påverkar skadefrekvensen på samma sätt.

Nyckeltal	Definition
Skadefrekvens	$\frac{\text{Antal skador}}{\text{Duration}}$
Medelskada	$\frac{\text{Skadekostnad}}{\text{Antal skador}}$
Riskpremie	$\frac{\text{Skadekostnad}}{\text{Duration}}$

Tabell 6: Nyckeltal.

2.2 Prissättning med Generaliserade Linjära Modeller

Målet är att undersöka hur Y , i vårt fall skadefrekvens, beror på ett antal förklarande variabler x (se Tabell 4, Avsnitt 1.3) i en multipel linjär regression. Det finns olika faktorer som gör att linjär regression eller den allmänna linjära modellen inte lämpar sig för prissättning inom sakförsäkring.

- (i) En linjär modell antar slumpmässiga, normalfördelade residualer, medan antalet försäkringsskador följer en diskret fördelning över de ickenegativa heltalen, skadekostnaderna är ickenegativa och oftast skevt fördelade åt höger.
- (ii) I linjära modeller är väntevärdet en additiv funktion av de förklarande variablerna, medan multiplikativa modeller är mer rimliga för prissättning.

Generaliserade linjära modeller generaliserar den vanliga linjära modellen och tar hand om ovanstående två problem på följande sätt:

- (i) Istället för att anta en normalfördelning, så fungerar GLM med en allmän klass med fördelningar, vilka innehåller ett antal diskreta och kontinuerliga fördelningar som specialfall, speciellt Poisson- och Gammafördelningarna.
- (ii) I linjära modeller är väntevärdet en funktion av de förklarande variablerna, $\mathbf{x}$. I GLM är en monoton funktion f av väntevärdet en linjär funktion av parametrarna, med additiva och multiplikativa modeller som specialfall.

$$f(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} \Rightarrow \mu_i = f^{-1}(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in}),$$

funktionen f kallas för länkfunktion och varierar vanligen mellan:

- Loglänk: $f(\mu_i) = \log(\mu_i)$
- Logitlänk: $f(\mu_i) = \text{logit}(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$
- Identitetslänk: $f(\mu_i) = \mu_i$
- Kanonisk länk: Om sannolikhetsfunktionen kan skrivas på exponentialfamiljform, alltså $p(x|\theta) = A(\theta) \exp(T(x)\eta(\theta))h(x)$ så $f(\mu_i) = \eta(\mu_i)$

I denna rapport använder vi oss av loglänken då vi modellerar en loglinjär Poissonregression för skadefrekvens.

Modellantaganden

- (i) Oberoende försäkringskontrakt: Betrakta n olika kontrakt. Låt X_i beteckna antal skador, och låt Y_i beteckna skadekostnaden för försäkringskontrakt i . Då är $X_1, \dots, X_n$ oberoende, och $Y_1, \dots, Y_n$ oberoende.
- (ii) Oberoende tid: Betrakta n disjunkta tidsintervall. Låt X_i beteckna antal skador, och låt Y_i beteckna skadekostnaden under tidsintervall i . Då är $X_1, \dots, X_n$ oberoende, och $Y_1, \dots, Y_n$ oberoende.
- (iii) Homogenitet: Låt X_i beteckna antal skador, och låt Y_i beteckna skadekostnaden för försäkringskontrakt i . Om två olika kontrakt i samma tariffcell (se exempel på tariffceller i Tabell 3) har samma duration och antal skador, då gäller att X_1 och X_2 är likafördelade samt att Y_1 och Y_2 är likafördelade.

2.3 Poissonfördelningen

2.3.1 Poissonfördelningen

Agresti (2013) förklarar att om responsvariabeln Y , exempelvis är antal skador, så finns det ingen fix övre begränsning n för Y . Eftersom Y an-

tar ickenegativa heltalsvärden bör dess fördelning placeras likaledes. Poissonfördelningen uppfyller detta. Dess sannolikheter beror på en enda parameter, nämligen väntevärdet μ . Poissons sannolikhetsfunktion är

$$p(y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots,$$

vilket uppfyller $E(Y) = \text{Var}(Y) = \mu$. Den är unimodal (endast en maxpunkt), med modalvärde likamedeltalsdelen av μ . Dess skevhet beskrivs av $E(Y - \mu)^3 / \sigma^3 = 1/\sqrt{\mu}$. Poissonfördelningen används till räknedata som inträffar slumpmässigt över tid eller rum när utfall i disjunkta tidsperioder eller regioner är oberoende.

2.3.2 Poisson i GLM

Under antagande om oberoende försäkringskontrakt, oberoende tid, och homogenitet i avsnitt 2.2, så kan antalet skador i ett enskilt försäkringskontrakt anses vara Poissonfördelat. På grund av oberoende försäkringskontrakt är då även den totala summan av alla skador i en viss tariffcell Poissonfördelat. Vi är dock mer intresserade av skadefrekvensen $Y_i = X_i/w_i$, där X_i är antal skador för försäkringskontrakt i , och w_i är durationen hos samma försäkringskontrakt. Detta är en slags transformation av Poissonfördelningen, och därför är det vanligt att den kallas för en relativ Poissonfördelning. Länkfunktionen är $\log(\mu)$, och regressionsmodellen kallas för en loglinjär modell.

$$\log(\mu_i) = \alpha + \sum_{j=1}^n \beta_j x_{ij}, \quad \text{eller}$$

$$\mu_i = e^\alpha \cdot \prod_{j=1}^n e^{\beta_j x_{ij}},$$

vilket innebär att om x_j ökar med en enhet, så har det en multiplikativ inverkan med faktor e^{β_j} : $\mu(x_j + 1) = \mu(x_j) e^{\beta_j}$

2.4 Övriga termer i Generaliserade Linjära Modeller

Interaktion I en multiplikativ modell för ett nyckeltal har alla faktorer samma relativa inverkan, oberoende av värdena på de andra faktorerna. Ett fall där multiplikativitet inte är en rimlig modell är för (till exempel) ålder och kön i trafikförsäkringar. Unga män har märkvärt högre skadefrekvens än yngre kvinnor, medan könsskillnaden generellt är mindre bland medelålders förare. Detta betyder att det finns ett samspel mellan prissättningsfaktorerna kön och ålder. För att inkludera denna effekt i modellen introducerar vi en ny faktor $x_{i1} \cdot x_{i5}$, en samspelsterm som beskriver relationen mellan kön och ålder.

Offset Agresti (2013) påpekar att det ofta kan vara så att en respons Y_i har en faktor t_i , som är känd på förhand, sådan att dess väntevärde är proportionellt mot t_i . Denna faktor kan till exempel, som i vårt fall, vara en tidsperiod. Då responsen Y_i i vårt fall är antal skador, och t_i är duration, får vi att skadefrekvensen Y_i/t_i har väntevärde μ_i/t_i . Med $\mathbf{x}$ som förklarande variabler har en loglinjär modell för den förväntade kvoten följande form

$$\log(\mu_i/t_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_n x_{in}$$

eller ekvivalent

$$\log(\mu_i) - \log(t_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_n x_{in} \Rightarrow$$

$$\Rightarrow \log(\mu_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_n x_{in} + \log(t_i).$$

$\log(t_i)$ justerar loglänken av väntevärdet och kallas för en offset. Skattningen ovan är ekvivalent med att använda sig av $\log(t_i)$ som en förklarande variabel i högerledet av ekvationen, men då en offset står i direkt relation till responsvariabeln ska en koefficient inte skattas till denna. Detta behöver specificeras i R med hjälp av det inbyggda kommandot "offset".

AIC Maindonald och Braun (2010) beskriver AIC, eller Akaike's Information Criteria, som en statistika utformad för att bland ett begränsat antal givna modeller välja den vars prediktionsförmåga är bäst. Agresti (2013) förklarar vidare att AIC bedömer en modell efter hur nära dess skattade värden tenderar att ligga det sanna väntevärdet. Trots att en enklare modell ligger längre ifrån den sanna relationen än vad en mer komplex modell har en benägenhet att göra kan den i vissa fall vara att föredra, ty den kan ge bättre skattningar av vissa egenskaper, som cellsannolikheter. Alltså är den optimala modellen den som tenderar att ha skattningarna närmst de sanna värdena. Kriteriet väljer den modell som minimerar

$$AIC = 2k - 2\ln(L) = -2(\ln(L) - k),$$

där k är antal parametrar och L är det maximerade likelihoodvärdet. Ett mindre värde på AIC är bättre, så vi ser att AIC straffar modeller med ett högre antal parametrar, som ju generellt har en bättre anpassning.

2.5 Splines

Teorin om B-splines kommer i huvudsak från J.R. Racine (2014), samt S. Sinha (u.å.).

2.5.1 B-spline

Maindonald och Braun (2010) förklarar att en splinekurva är en funktion som sammanfogar två eller fler polynomkurvor och kallas vanligtvis styckvis polynom. Punkterna där de sammanfogas kallas ”knutar”.

Den typ av spline vi kommer att använda oss av är en såkallad B-spline, basis spline. En B-splinekurva $B(x)$ av grad p är en linjärkombination av dess basfunktioner $B_{i,p}(x)$ av grad p som ges av

$$B(x) = \sum_{i=0}^{N+p} \beta_i B_{i,p}(x),$$

där $N \geq 0$ är antalet inre knutpunkter, p är graden, m är ordningen ($m = p + 1$), och antal basfunktioner är $K = N + m = N + p + 1$, där β_i ibland kallas för kontrollpunkter. Se Figur 1 för exempel av en splinekurva (nederst) och dess basfunktioner (överst). Funktionen ovan inkluderar intercept för splinekurvan, $B_{0,p}$, men då funktionen är tänkt att användas i en regressionsmodell som redan skattar ett intercept riskerar vi att designmatrisen inte får full rang, vilket innebär att de två separata intercepten blir oidentifierbara. Funktionen justeras alltså till

$$B(x) = \sum_{i=1}^{N+p} \beta_i B_{i,p}(x),$$

utan intercept. Bortsett från de inre knutpunkterna, N , finns även två gränsknutar, (en övre och en undre) t_0 och t_{N+1} så att

$$t_0 \leq t_1 \leq \dots \leq t_N \leq t_{N+1}.$$

Den totala mängden med knutpunkter fås genom att addera $(p + 1)$ av den undre gränsknuten, och $(p + 1)$ av den övre gränsknuten. Den ordnade mängden av knutpunkter är alltså denna totala mängd av knutar i ordning. Om avståndet mellan knutarna är lika långt kallas de likformiga. R använder sig av en metod som kallas kvantilknutssekvens vid placering av knutar. Detta innebär att längden på intervallen mellan knutarna kan variera, men att antalet observationer i varje intervall är lika många. Om vi till exempel har tre inre knutar så placerar R ut dem på följande sätt

$$(t_0, t_1, t_2, t_3, t_4),$$

så att 25% av observationerna är i intervallet (t_i, t_{i+1}) , för $i = 0, 1, 2, 3$.

$$(t_{0\%}, t_{25\%}, t_{50\%}, t_{75\%}, t_{100\%}),$$

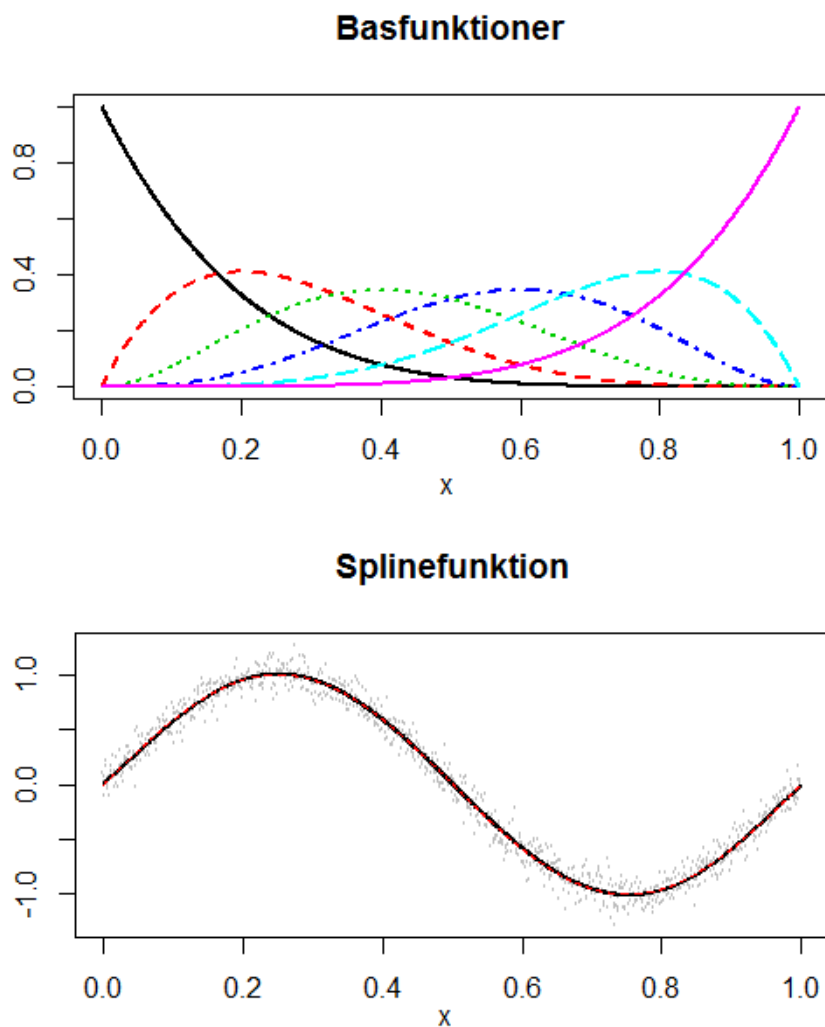
Längden på dessa intervall behöver alltså inte vara lika långa.

Baskurvorna konstrueras med hjälp av en rekursiv formel, vilket innebär att man för att konstruera en basfunktion av exempelvis grad 3 måste konstruera alla basfunktioner av lägre grad.

$$B_{i,0}(x) = I(u_i \leq x < u_{i+1}),$$

$$B_{i,n}(x) = \frac{u_{i+p} - u_i}{x - u_i} B_{i,p-1}(x) + \frac{u_{i+p+1} - x}{u_{i+p+1} - u_{i+1}} B_{i+1,p-1}(x),$$

där u_i är de ordnade knutpunkterna, och splinekurvans grad, p , antar värdena 1, 2, och 3. För ovanstående beräkning definierar vi $\frac{0}{0}$ som 0.



Figur 1: Den övre bilden illustrerar 7 basfunktioner, den undre bilden är den splinekurva de bildar.

Bilder från J. S. Racine (2014): A primer on regression splines.

I båda regressionsmodellerna ersätts den förklarande variabeln ”Ägarålder” med en splinekurva, och i modellen som även inkluderar kön som förklarande variabel ersätts även interaktionstermen mellan ägarålder och kön med ytterligare en splinekurva, se Avsnitt 3.3. Då antalet knutpunkter och placeringarna av dessa är lika för alla kurvor är även baskurvorna för de olika splinekurvorna lika. Det som skiljer sig är koefficientskattningarna för varje baskurva.

2.5.2 Konfidensband

Konfidensband En grafisk metod att illustrera osäkerheter i modellen är att konstruera ett konfidensband runt kurvan för de predikterade värdena. Alltså två kurvor som stänger in den skattade kurvan i ett slags konfidensintervall. Konfidensband av grad $1 - \alpha$ kan enligt Agresti (2013) konstrueras på följande sätt

$$y_{pred}(x) \pm \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)},$$

där $y_{pred}(x)$ är det predikterade värdet i punkten x , $\hat{\sigma}_{y_{pred}(x)}$ är dess skattade standardavvikelse, och λ är lämplig kvantil. Vi får två nya kurvor som för varje x ger ett konfidensintervall för det skattade värdet i given punkt. Den undre kurvan blir alltså

$$y_{pred}(x) - \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)},$$

och den övre blir

$$y_{pred}(x) + \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)}.$$

Bredden på konfidensbandet säger oss hur osäker skattningen är i en viss punkt, ju smalare bandet är desto säkrare är skattningen.

Då vi har en loglänk, $f(\mu_i) = \log(\mu_i)$, är det vanligt att till en början skatta ett konfidensband till $\log(\mu_i)$. Låt säga att det skattade konfidensbandet till $\log(\mu_i)$ är

$$(y_{pred}(x) - \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)}, \quad y_{pred}(x) + \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)}),$$

där $y_{pred}(x)$ är det predikterade värdet i punkten x för $\log(\mu_i)$. Då fås ett konfidensband till μ_i av att exponentiera ovanstående konfidensband

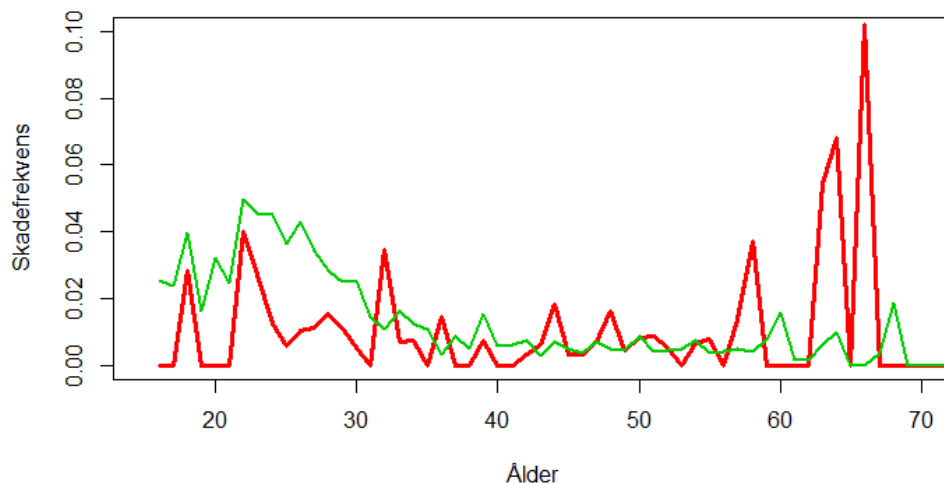
$$(e^{y_{pred}(x) - \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)}}, \quad e^{y_{pred}(x) + \lambda_{\alpha/2} \cdot \hat{\sigma}_{y_{pred}(x)}}).$$

3 Analys

3.1 Grafisk analys av skadefrekvens

För att få en översikt av hur skadefrekvensen ser ut för olika åldersgrupper och kön plottas medelvärden av skadefrekvenserna mot ålder i ett dataset

där både män och kvinnor är med (Figur 9 i Appendix), ett enbart för män, och ett enbart för kvinnor (se Figur 2). Man kan som förväntat se att män har högst skadefrekvens i åldrarna 20-30. Kvinnorna har däremot högsta skadefrekvens i åldrarna 60-70, värt att notera är dock det låga antalet kvinnor som försäkringstagare i detta fall, vilket även nämndes i beskrivningen av datamaterialet. Framförallt antal kvinnor med registrerad skada, de är bara 60st, jämfört med 606st män. Detsamma görs för medelskadan (se Figur 8 i Appendix). Till skillnad från skadefrekvensen ser medelskadan inte ut att ha något samband mellan varken ålder eller kön. Det intressanta blir alltså att studera skadefrekvensen närmre.



Figur 2: Den tjockare, röda linjen illustrerar medelvärdet av skadefrekvensen för kvinnor. Den något tunnare, gröna linjen är medelvärdet av skadefrekvensen för män.

3.2 Splinefunktion för variabeln Ägarålder

För att bestämma hur många knutar som är optimalt för våra modeller analyserar vi deras AIC-värden, då antal knutar varierar mellan 2-30, och övriga faktorer hålls fixa.

Bäst anpassning för modellen utan kön som förklarande variabel fås för skadefrekvensen då vi har 4 knutar, i modellen då vi har med både kön och samspelsterm mellan kön och ålder blir AIC lägst då vi har 3 knutar. Då syftet är att jämföra dessa olika modeller passar det bäst att vi väljer samma antal knutar för de olika modellerna. Jämför vi splinekurvorna för de predikterade värdena med kurvan för de observerade medelvärdet av

skadefrekvensen per ålder (Figur 9 i Appendix), kan man se att 5 knutar (Figur 11 i Appendix) motsvarar det faktiska datamaterialet bättre än 4 knutar (Figur 10 i Appendix), trots att 4 knutar gav lägst AIC.

En förklaring till detta kan vara att knutarnas placering flyttas då antalet ändras, eftersom R placerar ut knutarna baserat på kvantiler, och att det är en stor anledning till förändrat AIC-värde. Det kan då tänkas att en modell med knutar på samma ställen som, eller i närheten av, där knutarna befann sig i splinekurvan med 4 knutar, men med en eller ett par knutar fler kan ge bättre anpassning.

Den bästa modellen av de prövade fås då vi placerar ut tre inre knutar vid $t_1 = 36$, $t_2 = 49$, och $t_3 = 52$, ändknutarna är vid första och sista observerade x -värdena, $t_0 = 16$ och $t_4 = 92$.

3.3 Modell

Baserat på teorin i Avsnitt 2, samt ovanstående motivering gällande placering av splineknutarna har vi kommit fram till följande två regressionsmodeller.

Regressionsmodell utan kön som förklarande variabel

$$M_1 : \log(\mu_i) = \alpha + \sum_{k=1}^6 \beta_{1k} B_{k,p}(x_{i1}) + \sum_{j=2}^4 \beta_{2j} x_{ij} + \log(t_i),$$

samt regressionsmodell med kön som förklarande variabel

$$M_2 : \log(\mu_i) = \alpha + \sum_{k=1}^6 \beta_{1k} B_{k,p}(x_{i1}) + \sum_{j=2}^5 \beta_{2j} x_{ij} + x_{i5} \sum_{k=1}^6 \beta_{3k} B_{k,p}(x_{i1}) + \log(t_i),$$

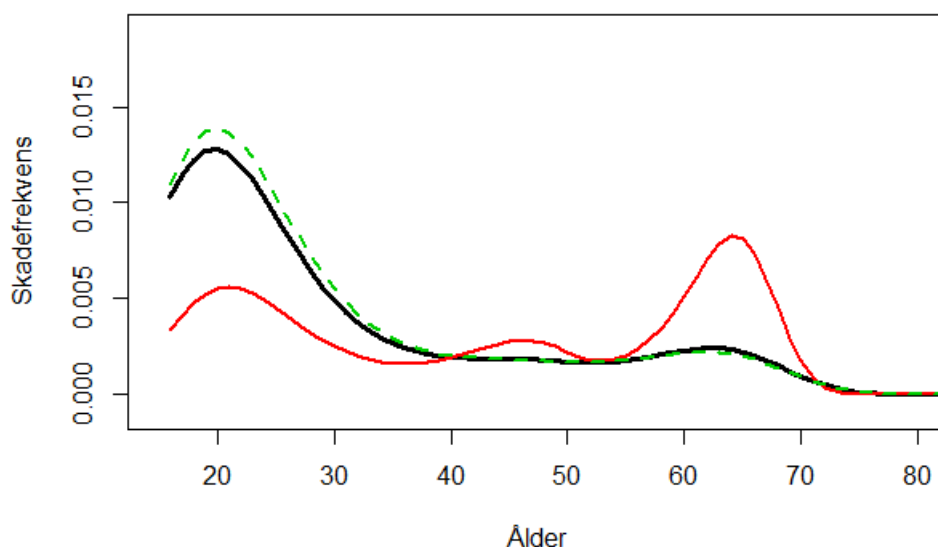
där x :en betecknas som i Tabell 4, Avsnitt 1.3, μ_i är antal skador, och t_i är durationen för försäkringskontrakt i .

Varje försäkringskontrakt befinner sig i en viss tariffcell (se Tabell 3), där koefficientskattningen för $x_2 - x_5$ beror på vilken tariffcell försäkringskontraktet befinner sig i. Detta illustreras med hjälp av x_{ij} som dummyvektorer, alltså en vektor med 1 på den plats som representerar vilken klass variabeln för givet försäkringskontrakt tillhör (se även Tabell 4), samt β_{2j} som en parametervektor med koefficientskattningarna, $2 \leq j \leq 5$. Vektorn $x_{ij} = 0$ om faktorn befinner sig i sin basklass, vilket exempelvis innebär att $x_{i5} = 1$ då försäkringstagaren är en kvinna, men då försäkringstagaren är en man befinner sig variabeln i sin basklass och $x_{i5} = 0$.

3.4 Observerade skillnader i skadefrekvens beroende på kön och ålder

För att få en bild av hur skadefrekvensen varierar beroende på kön och ålder plottar vi de predikterade värdena i var och en av de tre slutliga

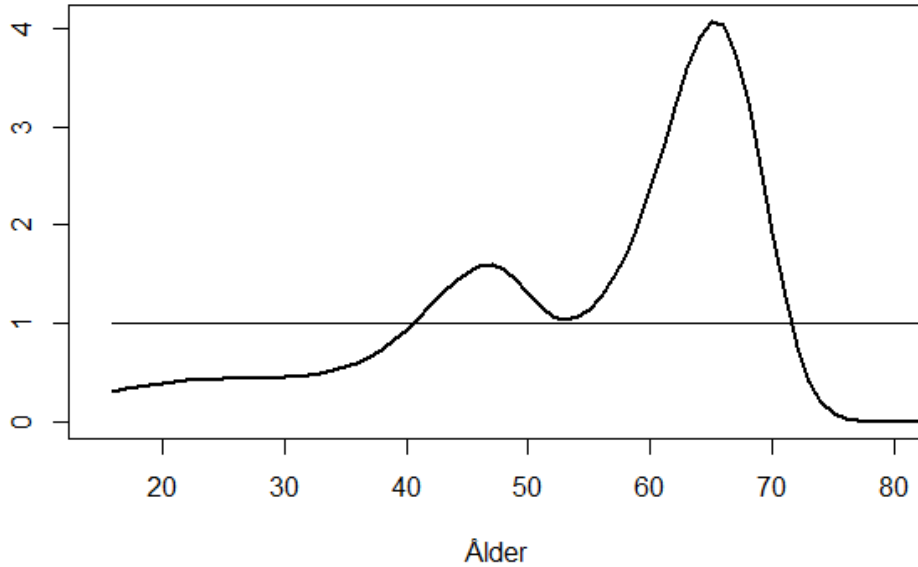
regressionsmodellerna, M_1 , M_2 för män, samt M_2 för kvinnor, mot ålder. Vi ser i Figur 3 att män i alla åldrar kommer att sänka sin premie, medan kvinnors premie höjs i yngre åldrar, och sänker premien i äldre åldrar. Figur 12 och 13 i Appendix visar den relativa förändringen för män och kvinnor. Detta stämmer ganska bra överens med de antaganden vi gjort efter att vi plottat skadefrekvenserna mot ägarålder i början av stycket. Förhållandet mellan kvinnor och män kan ses i Figur 4, som illustrerar kvoten mellan de predikterade värdena för kvinnor och män. Här ser vi att skadefrekvensen för kvinnor till en början är ungefär en tredjedel av den för män och att de som mest har ca fyra gånger högre skadefrekvens än män.)



Figur 3: Splinekurvor som representerar relationstalen för gemensamt material (svart, tjockare kurva), kvinnor (röd, tunnare kurva), och män (grön, streckad kurva). Respektive kurva har 5 knutar, med placeringarna anpassade så att AIC blev lägst. Övriga faktorer (Zon, MC-klass, Fordonsålder) är i sin basklass.

Relationstalen för skadefrekvens, som används för att prissätta en tariffklass fås genom att exponentiera regressionsmodellens koefficientskattningar, se Tabell 7.

När vi nu har utfört en regression med en splinekurva får vi en koefficientskattning för varje baskurva som ingår i vår spline. Vi har nu alltså fått en koefficientskattning för varje knut som ingår i vår spline. Dessa multipliceras med respektive basfunktioner för att sedan summeras till den splinefunktion vi vill ha. När detta är gjort har vi en kurva som motsvarar koefficientskattningarna för åldrarna 16-92, det vi egentligen var ute efter var ju relationstalen som fås genom att exponentiera koefficienterna, eller i vårt fall, splinekurvan. Detta ger kontinuerliga relationstal för skadefrekvens.



Figur 4: Illustration av kvoten mellan de predikterade värdena för kvinnor och män.

3.5 Osäkerhet i predikterade värden

För att skatta osäkerheten i kurvorna konstrueras ett 95%-igt konfidsensband runt dessa. I och med att datamaterialet är så pass stort kan vi normalapproximera och får följande konfidsensintervall för skattningarna

$$y_{pred}(x) \pm 1.96 \cdot \hat{\sigma}_{y_{pred}(x)},$$

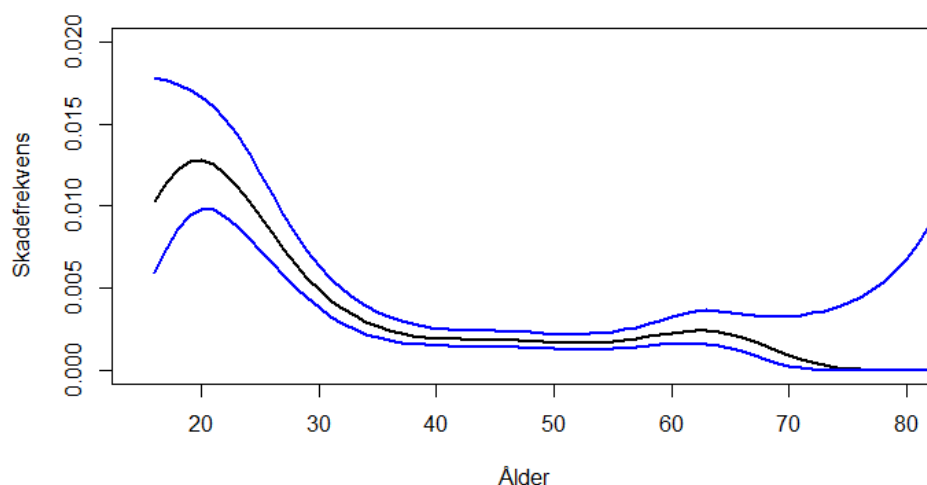
där $y_{pred}(x)$ är det predikterade värdet i punkten x för $\log(\mu_i)$, och $\hat{\sigma}_{y_{pred}(x)}$ är dess skattade standardavvikelse. Konfidsensbandet för μ_i ges då av

$$(e^{y_{pred}(x)-1.96 \cdot \hat{\sigma}_{y_{pred}(x)}}, e^{y_{pred}(x)+1.96 \cdot \hat{\sigma}_{y_{pred}(x)}}).$$

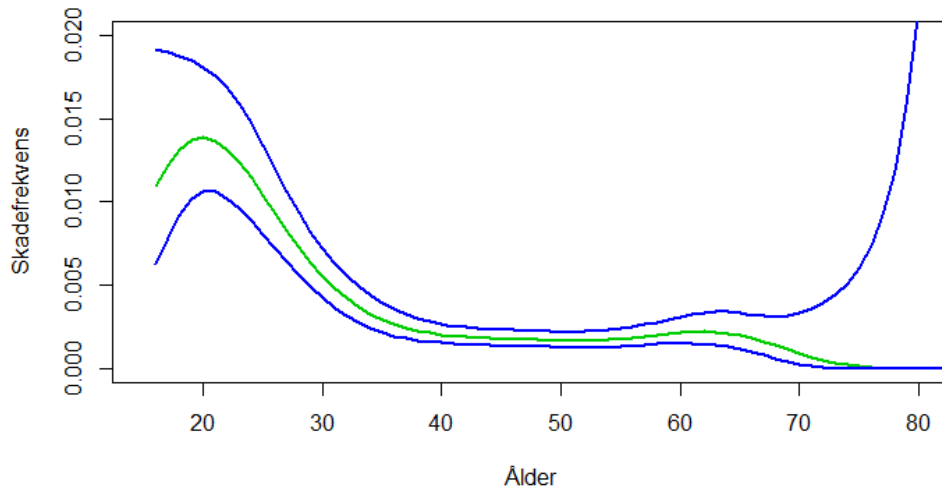
Gemensamt för alla tre funktioner (gemensamt, män, och kvinnor, se Figur 5-7) är att konfidsensbandet är väldigt brett i början, samt i slutet av kurvan. Vilket tyder på att det är en stor osäkerhet i dessa skattningar. Alla konfidsensband är snävast mellan åldrarna (ca) 35-60, vilket kan tänkas bero på att vi har mest data i de åldrarna. För kvinnor (Figur 7) är konfidsensbandet överlag relativt brett, om vi jämför med hur det ser ut för det gemensamma materialet (Figur 5), eller för män (Figur 6). Anledningen till att konfidsensbandet är såpass brett i början och slutet är att vi har få observationer där, för kvinnor har vi till exempel bara observationer mellan 22-66 medan vi för män har observationer mellan åldrarna 18-92.

Faktor	Klass	Skadefrekvens	
		Utan Interaktion	Med Interaktion
μ		0.0103	0.01099
Zon	1	4.5768	4.6216
	2	2.6174	2.6406
	3	1.5630	1.5724
	4 basklass	1	1
	5	0.9538	0.9694
MCKlass	1	1.2674	1.2128
	2	1.609519	1.5853
	3 basklass	1	1
	4	1.1203	1.1029
	5	1.7130	1.6593
	6	3.0569	2.9124
	7	1.8818	1.7722
Fordonsålder	1	3.4364	3.4023
	2	1.9212	1.8914
	3 basklass	1	1
Kön	Man, basklass	-	1
	Kvinna	-	0.3064

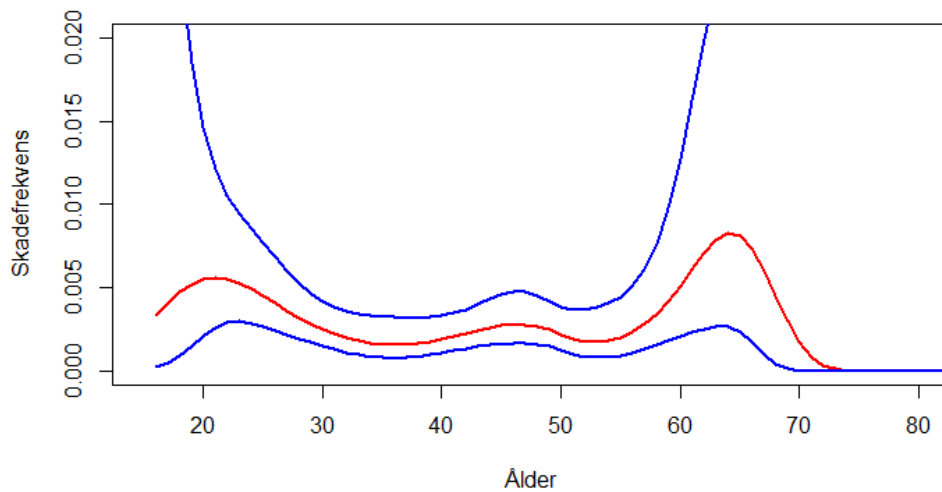
Tabell 7: Relationstal för skadefrekvens i modell utan interaktionsterm, samt med interaktionsterm.



Figur 5: Splinekurva som representerar relationstalen för skadefrekvens utan kön som faktor, samt konfidsband med konfidsgrad 95%. Zon, MC-klass, och Fordonsålder i respektive basklass.



Figur 6: Splinekurva som representerar relationstalen för skadefrekvens för män, samt konfidsensband med konfidsensgrad 95%. Zon, MC-klass, och Fordonsålder i respektive basklass.



Figur 7: Splinekurva som representerar relationstalen för skadefrekvens för kvinnor, samt konfidsensband med konfidsensgrad 95%. Zon, MC-klass, och Fordonsålder i respektive basklass.

4 Slutsatser

4.1 Resultat

Vi har visat att könsneutral prissättning ger störst skillnad för kvinnor, som i unga år förlorar på könsneutral prissättning, men vinner på det i äldre åldrar. För männen blir skillnaden mindre, men de tjänar något på det i yngre åldrar. Resultatet för männen känns ganska naturlig, i och med att det är känt att unga män generellt sett inte är så försiktiga i trafiken. Resultatet för äldre kvinnor är mer förvånande. Återigen kan den skeva fördelningen i antal män och kvinnor i datamaterialet nämnas. Framförallt de med registrerade skador, där det är 10 gånger fler män än kvinnor. Det kan även spekuleras i huruvida äldre kvinnor faktiskt är de som kör den registrerade motorcykeln, det finns en risk att det är yngre män som har valt att registrera sin motorcykel på sin mamma, farmor, eller mormor.

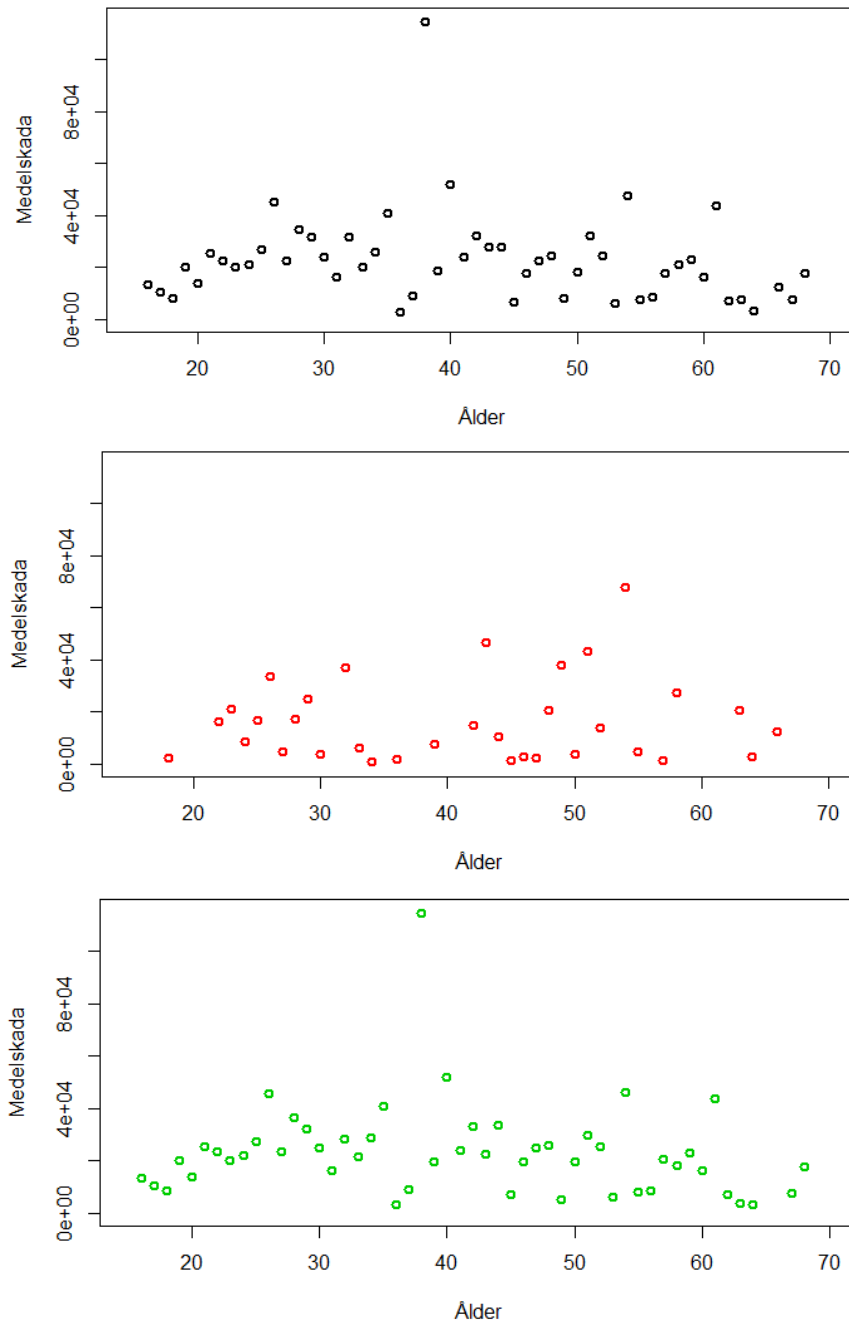
4.2 Förslag på förbättring av analys

Någonting som kan noteras är hur lika kurvorna för män och det gemensamma materialet var. En tanke är att det beror på att de (män) är överrepresenterade i detta datamaterial. Samtidigt går det inte att bortse från att män är överrepresenterade gällande motorcykelförsäkringar, frågan blir då i vilken utsträckning? Dessutom hade ett större datamaterial kanske gjort att vi hade fått bättre kurvor för medelskadan, vilket i sin tur hade lett till att vi hade kunnat studera den faktiska riskpremien istället för att enbart analysera skadefrekvensen. Det hade även varit av intresse att samlat in data från efter det att lagen införts för att jämföra det teoretiska med hur det faktiskt ser ut i dagsläget.

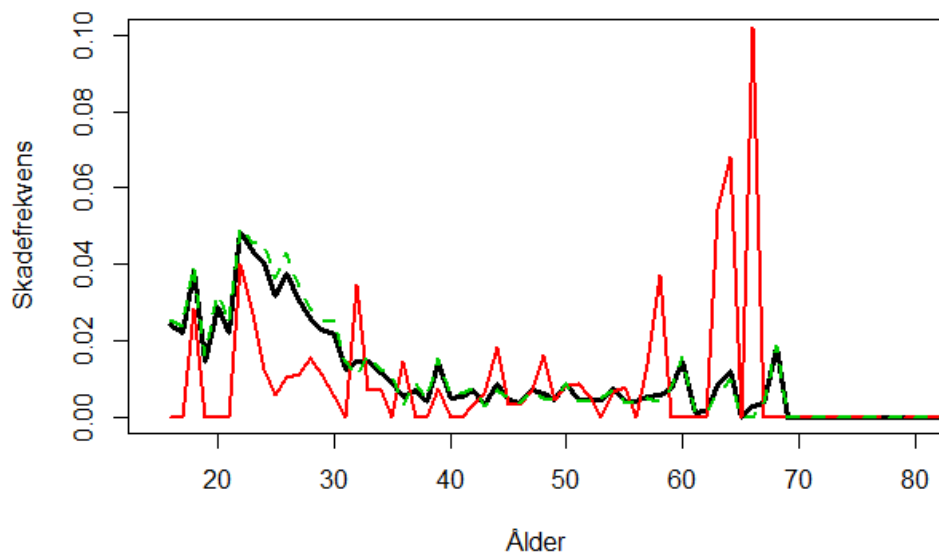
En förbättring hade även kunnat vara att göra en mer noggrann data- och modellanalys. Att exempelvis slå ihop fler av de klasser som inte är signifikant skilda från respektive basklass.

A Appendix

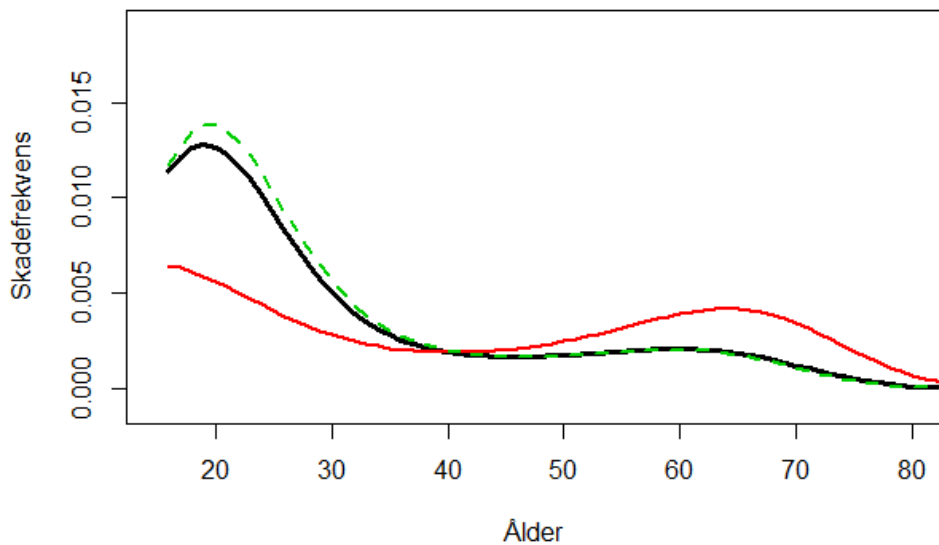
A.1 Figurer



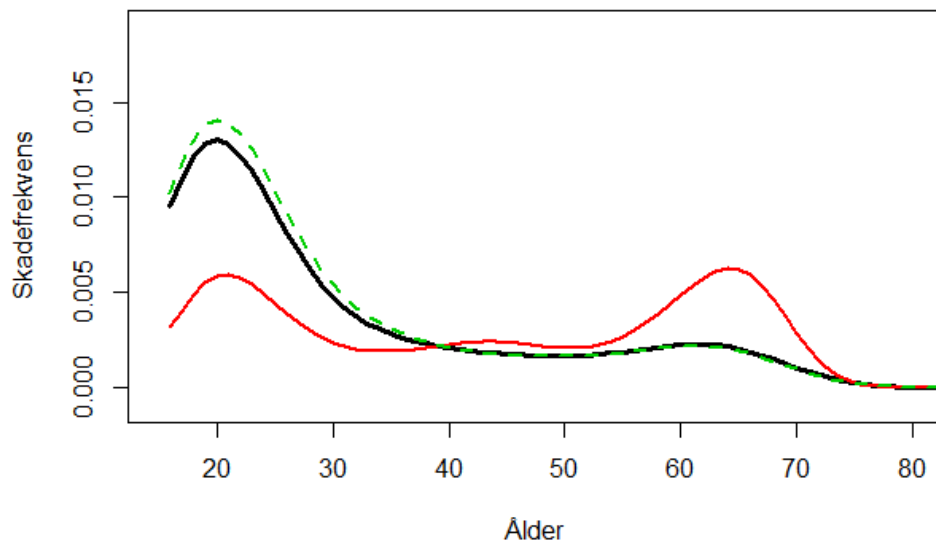
Figur 8: Scatterplots för medelskada för gemensamt material (överst), för kvinnor (mitten), och för män (nederst).



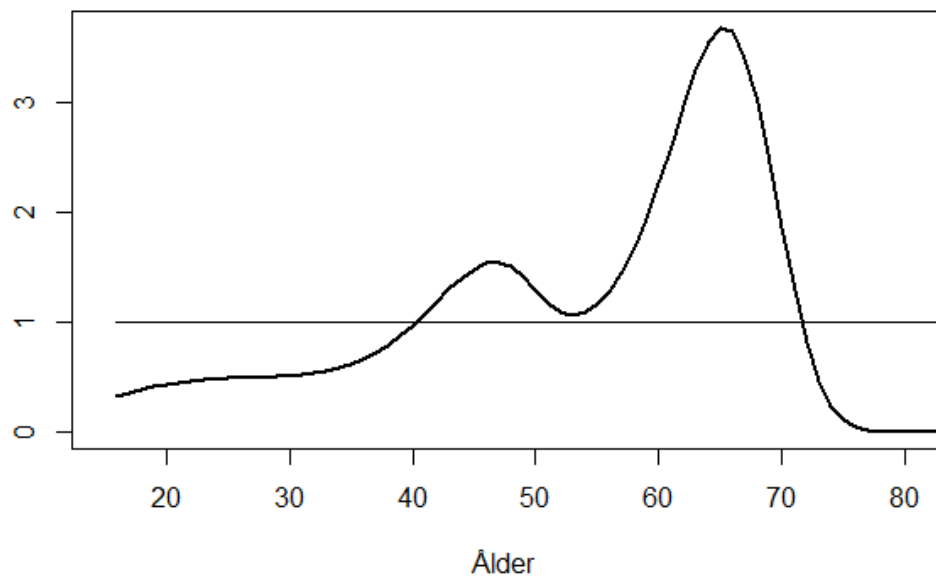
Figur 9: Medelvärde av den observerade skadefrekvensen för gemensamt material (tjockare, svart kurva), för kvinnor (tunnare, röd kurva), och för män (streckad grön kurva).



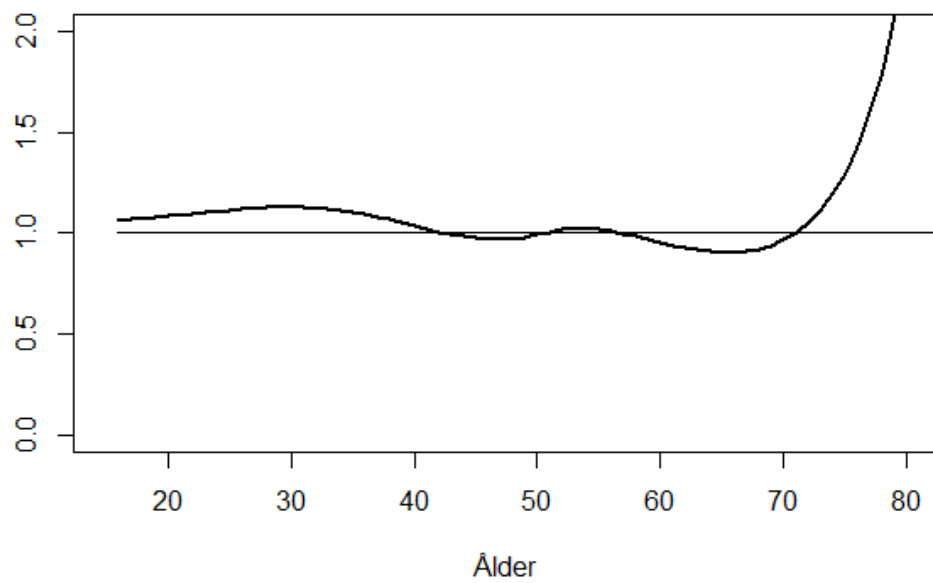
Figur 10: Splinekurvor som representerar relationstalen för gemensamt material (tjockare, svart kurva), för kvinnor (tunnare, röd kurva), och för män (streckad, grön kurva) där respektive kurva har 4 knutar. Övriga faktorer (Zon, MC-klass, Fordonsålder) är i sin basclass.



Figur 11: Splinekurvor som representerar relationstalen för gemensamt material (tjockare, svart kurva), för kvinnor (tunnare, röd kurva), och för män (streckad, grön kurva) där respektive kurva har 5 knutar. Övriga faktorer är i sin basclass.



Figur 12: Predikterade värden för kvinnor dividerat med predikterade värden för gemensamt material, för att illustrera förändringen för kvinnor.



Figur 13: Predikterade värden för män dividerat med predikterade värden för gemensamt material, för att illustrera förändringen för män.

Referenser

- A. Agresti, (2013). *Categorical Data Analysis*. Wiley.
- V. Aseervatam, C. Lex, M. Spindler, (2014). *How do unisex rating regulations affect gender differences in insurance premiums?*
- L. Calafa, O. Bonardi, (2014). *The use of gender in insurance pricing*. Fondazione Giacomo Brodolini.
- J. Daly, (u.å). *Impacts of Gender Neutral Pricing on the Life Insurance Industry*. IQ Financial.
- I.B Hossack, J.H. Pollard, B.Zehnwirth, (1983). *Introductory statistics with applications in general insurance*. Cambridge University Press.
- J. Maindonald, W.J. Braun, (2010). *Data Analysis and Graphics Using R*. Cambridge University Press
- E. Ohlsson, B. Johansson, (2010). *Non-Life Insurance Pricing with Generalized Linear Models*. Springer.
- J.S. Racine, (2014). *A primer on regression splines*.
- R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: <http://www.R-project.org/>.
- J. Sass, F. T. Seinfried, (2014). *Insurance markets and unisex tariffs: is the European Court of Justice improving or destroying welfare?* Scandinavian Actuarial Journal, No3 2014
- H. Schmeiser, T. Störmer, J. Wagner, (2014). *Possible Market Implications of Unisex Insurance Pricing*. Newsletter Insurance Economics, The Geneva Association No. 70 July 2014
- H. Schmeiser, T. Störmer, J. Wagner, (2014). *Unisex Pricing: Consumers' Perception and Market Implications*. The Geneva Papers, 2014.39.2 The International Association

for the Study of Insurance Economics.

S. Sinha, (u.å.). *A very short note on B-splines*. Department of Statistics, Texas A&M University.