Matematisk statistik
Stockholms universitet

Prissättning av en fordonsförsäkring med R

Christian Savemark

Examensarbete 2012:6

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.math.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2012:6**,
http://www.math.su.se/matstat

Prissättning av en fordonsförsäkring med R

Christian Savemark*

December 2012

Sammanfattning

Prissättning av sakförsäkringar kan göras med en mängd olika statistiska modeller. Vi utgår från två olika modeller och implementerar dem med hjälp av programmeringsspråket R och skapar prislistor - s.k. tariffer - för fordonsförsäkringar. Arbetet är upplagt med utgångspunkt från ett och samma datamaterial från ett försäkringsbolag. Vi baserar den första tariffen på faktorvariabler och en generaliserad linjär modell (GLM). Därefter använder vi oss av en generaliserad additiv modell (GAM) med kurvanpassning (B-splines) för kontinuerliga variabler.

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

PRISSÄTTNING AV EN FORDONSFÖRSÄKRING MED R

Abstract

Pricing of non-life insurance can be made with a variety of statistical models. We assume two different models and implement them using the programming language R and create price lists - so called tariffs - for vehicle insurance. The work is organized on the basis of the same data set from an insurance company. We base the first tariff on factor variables and a generalized linear model (GLM). Then, we use a generalized additive model (GAM) with curve fitting (B-splines) for continuous variables.

Tack

Jag vill speciellt tacka till Björn Johansson vid Länsförsäkringar för att ha formulerat problemställningen, svarat på frågor och försett mig med datamaterial. Stort tack riktas även till Jan-Olov Persson vid Matematiska insitutionen, Stockholms universitet för handledning och all konstruktiv kritik.

Innehåll

1 Introduktion

Vid bestämmande av priset för ett försäkringsavtal är det nödvändigt att utföra beräkningar på stora datamaterial. Dessa datamaterial är skapade av försäkringsbolaget och innehåller bland annat information om försäkringstagarna, försäkringsobjektens egenskaper, försäkringsavtalens giltighetsperioder och historik över försäkringstagarnas skador och kostnader. Även med dagens kraftfulla datorer och serversystem kan det ta lång tid att beräkna priset för försäkringsavtalet i den vanligt förekommande programmeringsmiljön SAS[1]. I vissa fall vill man använda sig av teori och funktionalitet som ännu inte implementerats eller som finns i en senare version vilken man inte har betalat licensavgift för. R är ett annat[2] programmeringsspråk vars huvudsakliga användningsområde – liksom SAS – är statistisk analys. Till skillnad från SAS är R gratis och bygger på principen om att alla är välkomna att utveckla det genom skapandet av tilläggsbibliotek. Det är därmed ett potentiellt alternativ för ett försäkringsbolag som önskar minska sina licenskostnader.

Vi ämnar undersöka om vi med framgång kan bestämma en tariff – lista med priser för en försäkring – på ett stort datamaterial från Länsförsäkringars motorförsäkringsportfölj med R. Bestämmandet av priser gör vi genom att modellera hur stor ekonomisk risk varje försäkringstagare utgör i ett kundbestånd eftersom det är denna ekonomiska risk som ligger till grund för hur mycket vi förväntar oss att behöva betala försäkringstagaren. Rent teoretiskt är det bäst att låta varje försäkringstagare betala för exakt så stor förväntad ekonomisk risk han eller hon utgör – varken mer eller mindre. Ty om vi låter försäkringstagaren betala för mycket riskerar vi att förlora försäkringstagaren till en konkurrent och om vi låter försäkringstagaren betala för lite förlorar vi inkomst. I verkligheten måste man även ta hänsyn till annat, såsom försäkringsbolagets omkostnader, reservsättning, marknadsläget, prisdifferentiering, vinstmål etc. Försäkringstagarens ekonomiska risk skattar vi genom att dela upp försäkringstagare med liknande egenskaper i så homogena celler som möjligt, för att sedan skatta cellens gemensamma ekonomiska risk.

I arbetsgången, som föreslogs av Björn Johansson vid Länsförsäkringar, implementerar vi två modeller som är tänkbara vid skattning av sådana ekonomiska risker. Vår första tariff är en tillämpning av en generaliserad linjär modell (förkortat GLM). Denna modell formulerades av John Nelder och Robert Wedderburn på 70-talet. Den andra modellen generaliserar en beståndsdel i en GLM

[1] http://www.sas.com
[2] http://www.r-project.org/

i något som kallas för en generaliserad additiv modell (GAM). Eftersom det tar lång tid att implementera modellerna fokuserar vi på själva prissättningen och utför ingen modelldiagnostik eller större jämförelse av modellerna.

2 Datamaterial

Till vårt förfogande har vi ett datamaterial som består av ungefär två miljoner observationer. Datamaterialet kommer från Länsförsäkringars motorförsäkringsportfölj. Varje observation utgör en försäkring med eventuella förnyelser vilket innebär att vissa har gällt länge men även att olika observationer kan vara samma kund vid olika tidpunkter. Observationerna innehåller uppgifter om försäkringstagarens kön och ålder, fordonets tillverkare, ålder, vikt och vikt per hästkraft, hur länge försäkringen gällt, antal skador och total skadekostnad för skadorna. Det är inte känt under vilka datum dessa försäkringskontrakt har gällt. I tabell 1 ger vi en överblick av tillgänglig data och i appendix (tabell 7) ger vi ett utdrag av observationer.

Tabell 1: Överblick av datamaterialet.

Kategoriska variabler	
Variabel	Värden
Kön	M för man
	K för kvinna
Bilmärkeskod	001 till 151
Körsträckecell	1 = 0 till 1000 mil per år
	2 = 1000 till 1500 mil per år
	3 = 1500 till 2000 mil per år
	4 = 2000 till 2500 mil per år
	5 = 2500 eller fler mil per år

Kontinuerliga/Diskreta variabler	
Variabel	Värden
Fordonsägarens ålder	Hela år
Fordonets ålder	Hela år
Fordonets vikt	Hela kg
Fordonets vikt/effekt (kg/hk)	Kg per hästkraft
Antal försäkringsår	År
Antal skador	
Skadekostnad	Hela kronor

Sammanfattning	
Antal försäkringar	1 749 358
Antal skador	157 253
Medelvärde bland skadekostnader	20 096 kronor
Minimum bland skadekostnader	1 kronor
Maximum bland skadekostnader	395 705 kronor

3 Generaliserad linjär modell

3.1 Allmänt

Vår första tariff kommer att baseras på en generaliserad linjär modell[3]. I detta avsnitt beskriver vi kortfattat vad det är för att sedan tillämpa den på datamaterialet.

En generaliserad linjär modell är en generalisering av en linjär modell på så vis att den tillåter responsvariabeln Y_i att ha en annan fördelning från exponentialfamiljen än normalfördelningen men som vanligt med en linjär struktur mellan de oberoende variablerna $\boldsymbol{X}$ (kallad designmatris) och väntevärdena för Y_i enligt

$$E[\boldsymbol{Y}] = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X}\boldsymbol{\beta})$$

där g är den s.k. länkfunktionen och $\boldsymbol{\beta}$ är okända parametrar. Allmänt kan $\boldsymbol{X}$ innehålla värden för de oberoende variablerna i varje kolonn och dummyvariabler om en variabel är en faktor i modellen.

Sannolikhetsfunktionen (eller täthetsfunktionen) för Y_i kan skrivas

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right]$$

för några funktioner $b(\cdot)$ och $c(\cdot)$. θ_i är den s.k. kanoniska parametern och kan bero på i medan dispersionsparametern $\phi > 0$ är samma för alla i. Här antas att data är given på listform vilket innebär att i motsvaras av en rad i en tabell med samma form som tabell 7. En rad i en sådan tabell innehåller även exponeringsvikten w_i och responsvariabeln y_i.

Vi kommer att använda oss av relativ Poissonfördelning, varför vi nu ger detta som exempel på en medlem i exponentialfamiljen. Anledningen till att vi använder oss av den relativa Poissonfördelningen är att vi senare antar att $w_iY_i \sim Po(w_i\mu_i)$.

Exempel 3.1. *Relativ Poissonfördelning.* Med $wy \in \mathbb{N}^+$ där w är en vikt, är täthetsfunktionen för en relativt Poissonfördelad stokastisk variabel Y

$$f_Y(y; \mu) = e^{-w\mu}\frac{(w\mu)^{wy}}{(wy)!} = \exp[w(y\log(\mu) - \mu) + c(y, \phi, w)], \qquad \mu > 0$$

Med den nya parametriseringen $\theta = \log(\mu)$ får vi

$$f_Y(y; \theta) = \exp\left[\frac{y\theta - e^\theta}{\phi/w} + c(y, \phi, w)\right]$$

[3]På engelska: *generalized linear model*, se exempelvis [1] eller originalartikeln av Nelder och Wedderburn på http://www.jstor.org/stable/view/2344614.

Detta är alltså en täthetsfunktion ur exponentialfamiljen med $\phi = 1$ och $b(\theta) = e^{\theta}$.

3.2 Offset

En *offset* $\boldsymbol{O}$ är en vektor med konstanter i den linjära strukturen

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{O}$$

Det kan vara så att man redan känner till en effekt på $g(\boldsymbol{\mu})$ från en variabel som man inte vill skatta parametrar för och som inte ingår i designmatrisen men ändå vill inkludera den. Vi kommer att använda oss av offsets i senare avsnitt.

3.3 Skattningar av parametrar

Mellan θ_i och μ_i gäller att

$$b'(\theta_i) = \mu_i \tag{1}$$

och mellan μ_i och $\boldsymbol{\beta}$ att

$$\mu_i = g^{-1}\left(\sum_{j=1}^{m} x_{ij}\beta_j\right) \tag{2}$$

för de m parametrarna β_j. För att få den mest sannolika skattningen av $\boldsymbol{\beta}$ (och därmed även θ_i), tittar vi på logaritmen av produkten $\prod_i P(Y_i = y_i) = \prod_i f_{Y_i}(y_i; \theta_i)$. Denna, som funktion av $\boldsymbol{\theta}$, är log-likelihoodfunktionen

$$l(\boldsymbol{\theta}; \phi, \boldsymbol{y}) = \frac{1}{\phi}\sum_i w_i(y_i\theta_i - b(\theta_i)) + \sum_i c(y_i, \phi, w_i)$$

Tillsammans med sambanden (1) och (2) ovan och kedjeregeln deriverar vi den m.a.p. β_j. Vi sätter derivatan till 0 och får då Maximum Likelihood-ekvationerna vilket ger skattningar för $\boldsymbol{\beta}$. Resultatet är

$$\frac{\partial l}{\partial \beta_j} = \sum_i w_i \frac{y_i - \mu_i}{b''(b'^{-1}(\mu_i))g'(\mu_i)} x_{ij} = 0, \qquad j = 1, 2, \ldots, m$$

där även

$$\mu_i = g^{-1}\left(\sum_{j=1}^{m} x_{ij}\beta_j\right)$$

för givna x_{ij} måste vara uppfyllt. Lösningen maximerar alltså likelihoodfunktionen och räknas vanligtvis ut numeriskt med hjälp av dator. Se [1] för en mer detaljerad uträkning.

3.4 Cellindelning

Vissa variabler i datamaterialet med försäkringstagarens och fordonets uppgifter är kategoriska medan andra är antingen diskreta eller kontinuerliga. Vi kommer att skapa s.k. faktorer av variablerna. Varje faktor har ett antal *nivåer* som motsvaras av antingen värden för variabeln eller intervall av värden. Kategoriska variabler har ofta en naturlig nivåuppdelning medan övriga delas upp i disjunkta mängder och varje sådan mängd motsvarar en nivå. *Celler* som försäkringstagare delas in i är en kombination av en nivå från var och en av faktorerna och eventuellt ett värde på var och en av övriga förklarande variabler i modellen – alla variabler i designmatrisen behöver i allmänhet inte vara faktorer. Olika rader i datamaterialet (som är på listform) med samma kombination av nivåer och värden på variabler som inte är faktorer, slår vi ihop till en enda cell. Antalet skador, skadebelopp och hur länge försäkringarna har gällt summerar vi i och med detta.

Vår modell, som kommer att baseras på en generaliserad linjär modell enligt föregående avsnitt, låter vi innehålla de förklarande variablerna kön-ägarålder, fordonsålder, årlig körsträcka, fordonsvikt och fordonets vikt per effekt. Kön-ägarålder är ett kopplat argument med en indelning i yngre män, äldre män, yngre kvinnor och äldre kvinnor. Alla dessa variabler låter vi vara faktorer. Man kan dela upp beståndet i fler celler – men vi väljer här få nivåer per faktor eftersom vi kommer att göra cellindelningen på ett mer elegant sätt i ett senare avsnitt. Nivåuppdelningen för de olika faktorerna sammanfattar vi i tabell 2.

3.5 Nyckeltal

I ett tidigare avsnitt beskrev vi att försäkringstagare (el. observation) i med samma värden på variablerna i designmatrisen tillhör samma cell. Låt index i beteckna aggregerad data över olika försäkringstagare som tillhör samma cell. Antalet skador i en cell i är en stokastisk variabel som vi betecknar med N_i och skadebelopp, även det en stokastisk variabel, för enskilda försäkringstagare betecknar vi med X_{ik}. Hur länge försäkringarna, som tillhör samma cell, tillsammans har gällt kallar vi för cellens sammanlagda duration och mäter denna i försäkringsår. Vi betecknar den med w_i. Vi kommer att modellera tre storheter som vi kallar *nyckeltal* för cell i: skadefrekvens, medelskada och riskpremie. Skadefrekvens S_i definieras som antalet skador per försäkringsår, medelskada M_i som skadekostnad per skada och riskpremie R_i som skadekostnad per

Tabell 2: Faktorer och nivåer.

Faktorindex j	Faktor	Nivå	Nivåbeskrivning
1	Kön-ålder	1	Kvinnor yngre än 30 år
		2	Kvinnor äldre än 30 år
		3	Män yngre än 30 år
		4	Män äldre än 30 år
2	Fordonsålder	1	0 – 2 år
		2	3 – 5 år
		3	6 – 8 år
		4	9 år eller äldre
3	Fordonsvikt	1	0 – 1000 kg
		2	1000 - 1500 kg
		3	1500 – 2000 kg
		4	2000 kg eller mer
4	Körsträcka	1	0 – 1000 mil/år
		2	1000 – 1500 mil/år
		3	1500 – 2000 mil/år
		4	2000 – 2500 mil/år
		5	2500 mil/år eller mer
5	Fordonets	1	0 – 10 kg/hk
	vikt/effekt	2	10 – 20 kg/hk
		3	20 – 30 kg/hk
		4	30 kg/hk eller mer

försäkringsår. Mellan nyckeltalen gäller sambandet

$$R_i = S_i \cdot M_i$$

När vi inte menar ett speciellt nyckeltal använder vi beteckningen Y_i. Vi sammanfattar detta i tabell 3.

Vi vill skatta förväntad skadekostnad för försäkringstagare i varje cell. Denna skattning delar man vanligtvis upp i att först skatta $E[S_i]$ – hur ofta vi förväntar att en skada uppstår – och sedan $E[M_i|N_i = n_i]$ – hur mycket vi förväntar att skadan kostar. Genom att sedan multiplicera de två får vi fram hur mycket vi förväntar oss att försäkringstagarna i cellen kommer att kosta per försäkringsår. Vi har följande sats från [1].

Sats 3.1. *Mellan $E[R]$, $E[S]$ och $E[M]$ gäller sambandet*

$$E[R] = E[S] \cdot E[M]$$

Tabell 3: Nyckeltal.

Nyckeltal	Y_i
Skadefrekvens	$S_i = \frac{\text{Antal skador}}{\text{Duration}} = \frac{N_i}{w_i}$
Medelskada	$M_i = \frac{\text{Skadekostnad}}{\text{Antal skador}} = \frac{\sum_{k=1}^{N_i} X_{ik}}{N_i}$
Riskpremie	$R_i = \frac{\text{Skadekostnad}}{\text{Duration}} = \frac{\sum_{k=1}^{N_i} X_{ik}}{w_i}$

Denna sats använder vi senare när vi har skattat $E[S]$ och $E[M]$ för att få fram $E[R]$.

3.6 Multiplikativ struktur för väntevärden

Väntevärdena för nyckeltal kommer fortsättningsvis att antas ha en multiplikativ struktur:

$$E[Y_i] = \mu_i = \mu_0\gamma_{i1} \cdot \ldots \cdot \gamma_{ik} \cdot \ldots \cdot \gamma_{iK}$$

där γ_{ik} är *relationstal* för variabel k och rad i i data på listform. För varje variabel väljer vi en basnivå och alla basnivåer utgör tillsammans en bascell. Det är lämpligt att välja basnivåer med hög exponering eftersom vi då får en stabil skattning av bascellens väntevärde. För basnivåerna låter vi $\gamma_{ik} = 1$. Detta leder senare till att modellen får en unik lösning när vi skattar parametrarna. Man kan tolka μ_0 som bascellens väntevärde av nyckeltalet och relationstalen som de procentuella avvikelserna från det.

Väljer vi länkfunktion $g(x) = \log(x)$ får vi $g^{-1}(x) = e^x$ och därmed

$$\begin{aligned}
\mu_i &= e^{\boldsymbol{X}_i\boldsymbol{\beta}} \\
&= e^{\sum_{j=1}^{m}\beta_j x_{ij}} \\
&= e^{\beta_1 x_{i1}} \cdot \ldots \cdot e^{\beta_m x_{im}} \\
&= \mu_0\gamma_{i1} \cdot \ldots \cdot \gamma_{ik} \cdot \ldots \cdot \gamma_{iK}
\end{aligned}$$

vilket är sambandet mellan relationstalen, parametrarna och väntevärdet för observation i.

3.7 Fördelningar för nyckeltalen

Vi gör även i detta skede ett par modellantaganden: Vi antar att antalet skador och skadekostnader är oberoende mellan olika försäkringstagare. Vi antar även att de är oberoende i tid, dvs. olika skador och skadekostnader för samma försäkringstagre är oberoende. Slutligen antar vi att om två försäkringstagare tillhör samma cell och har samma exponering, så har de samma fördelning för antalet skador och skadekostnader.

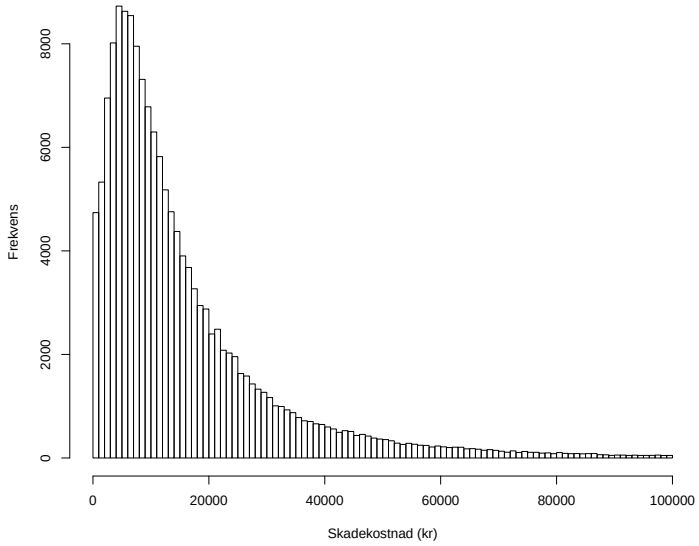
Låt $N(t)$ beteckna antalet skador för ett försäkringskontrakt under intervallet $[0, t]$ och $N(0) = 0$. Den stokastiska processen $\{N(t), t \geq 0\}$ kallas för skadeprocessen och med de två sista antagandena ovan samt ett antagande om att skador inte anhopar, kan man enligt [1] visa att skadeprocessen är en Poissonprocess. Vi antar därför att antalet skador för ett försäkringskontrakt under ett tidsintervall är Poissonfördelat med parameter $w_i\mu_i$ där w_i är tidsintervallets längd. På grund av det första modellantagandet ovan får vi en sammansatt Poissonprocess om vi låter två försäkringskontrakt tillhöra samma cell och får alltså även på aggregerad nivå att antalet skador är Poissonfördelat. Man kan visa att skadefrekvensen i en cell är relativt Poissonfördelad om man antar att antalet skador i en cell är Poissonfördelat[4] med parameter och väntevärde $w_i\mu_i$.

I figur 1 visar vi medelskadan över försäkringstagare med minst en skada. Det *ser* ut som att gammafördelningen kan vara ett lämpligt alternativ. Gammafördelningen har även vissa lämpliga egenskaper; den är positiv och snedvriden till höger och har en standardavvikelse proportionell mot dess väntevärde. Det är rimligt att anta att små och stora skadekostnader varierar proportionellt mot storleken på beloppen.

3.8 GLM i R

Vi beskriver kortfattat hur förfarandet går till. Innan vi får fram skattningar måste datamaterialet förberedas för analys. Det innebär att vi definierar faktorer, nivåer och nyckeltal och sätter basnivåer för faktorerna. Därefter aggregerar man eventuellt observationer med samma kombination av faktornivåer för att få färre rader och snabbare skattningar av parametrar. Man förlorar dock viss

[4]Vanligtvis är variansen för antalet skador i en cell större än dess väntevärde vilket inte är förenligt med antagandet om att antalet skador är Poissonfördelat. Detta beror på att försäkringstagarna i samma cell inte är helt homogena. Det är möjligt att byta ut Poissonfördelningen mot över-spridd Poisson eller negativ binomial för att få bättre anpassning.



Figur 1: Histogram över medelskadan per försäkringstagare för försäkringstagare med minst en skada. Skadekostnader över 100 000 kronor finns ej med.

information om man aggregerar data. Se [1] för en diskussion kring detta. Vi ger ett urval av funktioner för dessa ändamål i appendix.

Efter att man har förberett data i R är det enkelt att få fram skattningar i en GLM. Det gör vi med

```
glm(formula, family = ..., weights = ..., ...)
```

där `formula` skrivs på formen

```
respons ~ variabel1 + ... + variabelN + offset(...),
```

`family` är vilken familj fördelningen tillhör, `weights` är eventuell vikt för varje observation och `offset(...)` är en konstant (vektor) enligt tidigare avsnitt. När vi är intresserade av skattningar för skadefrekvensen sätter vi helt enkelt

`formula = skadefrekvens ~ ägarålder + ... + fordonsviktpereffekt,`

`family = poisson(link = "log")` och `weights = duration`. Man kan, i specialfallet med Poissonfördelningen, även använda sig av

`formula = antal skador ~ ägarålder + ... + offset(log(duration)),`

`family = poisson(link = "log")` men utan någon vikt.

För medelskadan gäller å andra sidan

```
formula = medelskada ~ ägarålder + ... + fordonsviktpereffekt,
```

`family = Gammma(link = "log")` och `weights = antalskador`.

Parameterskattningar får vi enkelt fram med `coef(glm(...))`. Eftersom vi använder en log-länk får vi relationstalen om vi exponerar parameterskattningarna med `exp(coef(glm(...)))`.

3.9 Skattade relationstal

Sats 3.1 och antagandet om en multiplikativ struktur hos väntevärdena för nyckeltalen ger samband mellan relationstalen och bascellernas väntevärden enligt $\gamma_k^R = \gamma_k^S \cdot \gamma_k^M$ och $\mu_0^R = \mu_0^S \cdot \mu_0^M$. Detta, tillsammans med en generaliserad modell ger relationstalen i tabell 4. Vi ger ett exempel på hur mycket en man som fyllt 31 år får betala enligt denna tariff.

Exempel 3.2. En man som fyllt 31 år och som försäkrar ett nytt fordon, vilket körs strax under 1500 mil per år, med vikt kring 1400 kg och med 15 kg per hk får riskpremien

$$765.1931 \cdot 1.0000 \cdot 1.0984 \cdot 1.0000 \cdot 1.0342 \cdot 1.1337 \approx 985 \text{ kronor}$$

Relationstalen verkar vara rimliga, exempelvis utgör yngre försäkringstagare större risk än äldre och fordon som körs längre sträckor per år löper större risk att skadas oftare vilket medför att relationstalen ökar med körsträckan. Tittar vi på relationstalen för medelskadan ser vi att det inte är någon signifikant skillnad mellan olika körsträckenivåer förutom att de som kör allra mest utmärker sig något. De relationstal som inte är signifikant skilda från basnivåns relationstal sätter vi till 1. Detta är en av fördelarna av att analysera skadefrekvens och medelskada var för sig istället för riskpremien direkt: vi får ut mer information från analysen när vi ser om det är skadefrekvensen eller medelskadan som påverkar riskpremien mest för de olika nivåerna.

Behållningen med modellen är att den är enkel: det är enkelt att skatta konfidensintervall, lägga till eller minska antalet variabler eller helt byta fördelning för skadefrekvens och medelskada. Tariffen skapar dock stora glapp mellan olika celler. En manlig kund som fyller 30 år men samtidigt inte gör några andra ändringar minskar sin premiekostnad med över 60 procent. Man kan tänka sig

att introducera fler nivåer på varje faktor men då riskerar man att få celler med för lite data vilket ger en skakig tariff. Den tredje tariffen (Avsnitt 4) inkluderar splines för de kontinuerliga variablerna för att få en tariff med färre hopp i prissättningen mellan celler.

Tabell 4: Basnivåer för varje faktor är skrivna med fetstil. Relationstal med kursiv text är icke-signifkant skilda (5 %) från basnivån och vi sätter dem därför till 1.

Tariff				
		Relationstal γ_j		
Faktor	Nivåbeskrivning	Skadefrekvens	Medelskada	Riskpremie
Kön-ålder	Kvinnor yngre än 30 år	1.4647	1.1066	1.6209
	Kvinnor äldre än eller 30 år	1.0640	0.9824	1.0452
	Män yngre än 30 år	1.7360	1.2335	2.1425
	Män äldre än eller 30 år	1.0000	1.0000	1.0000
Fordonsålder	0 – 2 år	1.0910	*1.0000*	1.0984
	3 − 5 år	1.0000	1.0000	1.0000
	6 – 8 år	0.9114	1.0417	0.9493
	9 år eller äldre	0.6247	0.9139	0.5714
Fordonsvikt	0 – 1000 kg	0.7233	0.8741	0.6329
	1000 - 1500 kg	1.0000	1.0000	1.0000
	1500 – 2000 kg	1.2098	1.2049	1.4588
	2000 kg eller mer	1.3194	1.5884	2.1050
Körsträcka	**0 − 1000 mil/år**	1.0000	1.0000	1.0000
	1000 – 1500 mil/år	1.0349	*1.0000*	1.0342
	1500 – 2000 mil/år	1.1469	*1.0000*	1.1466
	2000 – 2500 mil/år	1.2468	*1.0000*	1.2470
	2500 mil/år eller mer	1.4784	1.0730	1.5863
Fordonets	0 – 10 kg/hk	1.7027	1.5888	2.7075
vikt/effekt	10 – 20 kg/hk	1.1333	*1.0000*	1.1337
	20 − 30 kg/hk	1.0000	1.0000	1.0000
	30 kg/hk eller mer	0.5012	0.5243	0.2739
μ_0		0.0493	15521	765

4 Generaliserad additiv modell

4.1 Allmänt

Vissa av variablerna i föregående avsnitt är svåra och tidskrävande att dela upp i disjunkta mängder som motsvarar någon nivå. Nivåuppdelning medför dessutom problemet att välja antal nivåer; väljer man många riskerar man för lite data i varje cell vilket medför osäkra skattningar och väljer man för få får man stora hopp i tariffen. I tariffen från föregående avsnitt som baseras på en generaliserad linjär modell får vi exempelvis ett stort hopp i riskpremie mellan yngre och äldre personer. Lösningen i detta avsnitt är att införa kontinuerliga kurvanpassningar till de kontinuerliga variablernas väntevärden. Detta är en generalisering av den den generaliserade linjära modellen och kallas för generaliserad additiv modell.

En generaliserad additiv modell [1, 2] har en väntevärdesstruktur enligt

$$g(\mu_i) = \boldsymbol{X_i}\boldsymbol{\beta} + f_1(x_{i1}) + \ldots + f_j(x_{ij}) \ldots + f_J(x_{iJ}) \tag{3}$$

där g är länkfunktion, i som vanligt är observationsnummer och rad i data på listform, $\boldsymbol{X_i}$ är en radvektor ur en designmatris $\boldsymbol{X}$ och $\boldsymbol{\beta}$ är en parametervektor. Funktionerna f_j är godtyckliga och i detta avsnitt vill vi använda oss av splines för att anpassa en kurva för de kontinuerliga variablerna ägarens ålder, fordonets ålder, fordonets vikt och fordonets vikt per effekt. De kategoriska variablerna kön och körsträcka är fortfarande faktorer och ingår precis som tidigare i $\boldsymbol{X}$. Tanken är att detta ska ge oss en jämnare tariff utan stora glapp mellan närliggande celler då även relationstalen blir kontinuerliga för de kontinuerliga variablerna.

4.2 Splines

Splines och B-splines som vi nu ämnar definiera kan definieras på olika sätt [1, 2]. Följande definitioner är från [1]. Vi betecknar *knopar* – tal i stigande ordning – med $u_1, \ldots, u_r$.

Definition 4.1. *Spline.* En funktion på intervallet $[u_1, u_r]$ kallas för en spline av grad j om den är $j-1$ gånger kontinuerligt deriverbar och på varje intervall $[u_k, u_{k+1}]$ är ett polynom av grad j.

Definitionen säger att en spline är en kurva ihopsatt av polynom definierade på intervall. I ändpunkterna på dessa intervall kräver vi att kurvan är $j-1$

gånger kontinuerligt deriverbar. För ett polynom av grad tre behöver alltså både derivatan och andraderivatan vara lika i knoparna.

Exempel 4.1. Kubiska splines. Om vi definierar $s(x)$ i intervallet $[u_1, u_r]$ som $s(x) = p_k(x)$ för $x \in [u_k, u_{k+1}]$ och $k = 1, \ldots, r-1$, där $p_k(x) = c_{k0} + c_{k1}x + c_{k2}x^2 + c_{k3}x^3$ och c_{kl} valda så att $p'_{k-1}(x) = p'_k(x)$ och $p''_{k-1}(x) = p''_k(x)$ för $k = 2, \ldots, r-1$ får vi en spline $s(x)$ av grad 3.

Vi kommer dock ej att använda oss av definitionen av splines direkt. Satsen nedan säger att varje spline $s(\cdot)$ kan skrivas som en linjärkombination av s.k. B-splines. B-splines har trevliga numeriska egenskaper och är väl beprövade inom numerisk analys.

Definition 4.2. B-splines. För $k = 1, \ldots, r-2$ sätt

$$B_k^0(x) = \begin{cases} 1 & : x \in [u_k, u_{k+1}) \\ 0 & : x \notin [u_k, u_{k+1}) \end{cases}$$

och för det sista intervallet

$$B_{r-1}^0(x) = \begin{cases} 1 & : x \in [u_{r-1}, u_r] \\ 0 & : x \notin [u_{r-1}, u_r] \end{cases}$$

För $m \geq 0$ definierar vi B-splines rekursivt genom

$$B_k^{m+1}(x) = \begin{cases} \frac{x-u_{k-m-1}}{u_k-u_{k-m-1}} B_{k-1}^m(x) + \frac{u_{k+1}-x}{u_{k+1}-u_{k-m}} B_k^m(x) & : k = 1, \ldots, r+m \\ 0 & : k \leq 0 \vee k \geq r+m \end{cases}$$

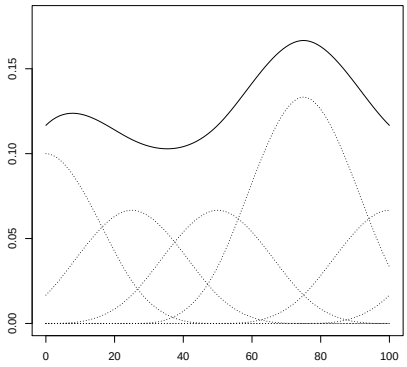
Vi låter $u_k = u_1$ för $k \leq 0$ och $u_k = u_r$ för $k \geq r+1$. Här betecknar $m+1$ splinens grad.

B-splines består alltså av två trappfunktioner[5] – varav den ena är definierad på halvöppna intervall och den andra på ett slutet intervall (det sista) – och ett rekursivt samband mellan B-splines av grad m och $m+1$. I appendix beräknar vi $B_k^3(x)$ för att ge läsaren en känsla för dess utseende. Nästa sats säger att dessa B-splines agerar byggstenar åt en spline $s(\cdot)$.

Sats 4.1. *För en given mängd av r knopar kan en spline s av grad m skrivas som*

$$s(x) = \sum_{k=1}^{r+m-1} c_k B_k^m(x)$$

[5]Man kan även tänka indikatorfunktioner.



Figur 2: De streckade kurvorna är B-splines av grad 3 multiplicerade med sina koefficienter. Summan av dessa bildar splinen, representerad av den heldragna kurvan.

med unika konstanter $c_1, \ldots, c_{r+m-1}$.

Bevis. Återfinns i [1]. I figur 2 visar vi en grafisk representation av sats 3.

Använder vi oss av satsen ovan på kubiska splines får vi

$$s(x) = \sum_{k=1}^{r+2} c_k B_k^3(x),$$

med unika konstanter $c_1, \ldots, c_{r+2}$. Eftersom vi uteslutande använder oss av kubiska splines (B-splines av grad 3) skriver vi hädanefter inte ut index m men samtidigt börjar vi använda oss av j som tidigare för att indikera variabelindex.

Vi låter $c_k = \beta_{jk}$ och på så vis binder vi ihop den allmänna formeln (3) för en GAM med splines eftersom vi nu har fått parametrar β_{jk} som inte ingår i $\boldsymbol{\beta}$ att skatta:

$$\begin{aligned} g(\mu_i) &= \boldsymbol{X_i\beta} \\ &+ \sum_{k=1}^{r_1+2} \beta_{1k} B_{1k}(x_{i1}) + \ldots + \sum_{k=1}^{r_j+2} \beta_{jk} B_{jk}(x_{ij}) + \ldots + \sum_{k=1}^{r_J+2} \beta_{Jk} B_{Jk}(x_{iJ}) \end{aligned}$$

4.3 Ändligt antal värden

Även om en variabel är kontinuerlig observerar vi blott ett ändligt antal värden av den eftersom antalet observationer är ändligt. Av beräkningsmässiga skäl

kommer vi att approximera dessa till närmevärden och anpassar därefter en kurva till närmevärdena. Vi låter dessa närmevärden betecknas av den monotont växande följden $(z_{jk})_{k=1}^{m}$. Av naturliga skäl låter vi $u_{j1} \leq z_{j1}$ och $z_{jm} \leq u_{jr}$ där u_{jk} betecknar en knop enligt tidigare. Det faller sig då naturligt att beteckna nyckeltal och vikter för närmevärdena med $y_{z_{jk}}$ respektive $w_{z_{jk}}$ eller om det framgår av sammanhanget vilken variabel som man menar, bara y_{z_k} respektive w_{z_k}. Vi passar även på att inför beteckningen $z_{k'}$ för det närmevärde som vi väljer till basvärde och med vilket uttrycket $\frac{s(z_k)}{s(z_{k'})}$ skalar om splinen till en kurva för relationstal istället för en kurva för väntevärden.

Exempel 4.2. Ett observerat värde av vikten för ett fordon i datamaterialet är 1132 kg. Vi avrundar detta till säg närmsta hela 50 kg av skälen ovan. Det nya värdet blir då 1150 kg.

4.4 Skattning av parametrar

Vi antar nu att väntevärdet μ bara beror på en kontinuerlig variabel, dvs. $g(\mu(x)) = \sum_{k=1}^{r+2} \beta_k B_k(x)$ med någon länkfunktion. Vi är i behov av ett mått för att välja $\boldsymbol{\beta}$ för en splinen på något sätt. Det görs i [1, 2] genom att införa en diskrepans med en extra "straffterm":

$$\Delta(s(x;\boldsymbol{\beta})) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + \lambda \int_a^b (s''(x))^2 dx$$

Diskrepansen för den relativa Poissonfördelningen är

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i w_i (y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \hat{\mu}_i - y_i)$$

och för gammafördelningen

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i w_i (\log\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i}{\hat{\mu}_i} - 1)$$

Båda kan ses som viktade mått på hur långt observationerna y_i ligger från skattningarna $\hat{\mu}_i$. Om man exempelvis låter $\hat{\mu}_i = y_i$ (den mättade modellen) så blir båda diskrepanserna 0. Den andra termen, $\lambda \int_a^b (s''(x))^2 dx$, är ett mått på splinens totala acceleration i kvadrat över intervallet $[a, b]$. λ är en parameter som agerar vikt; ett stort värde på λ ger integralen – och därmed kurvaturen hos grafen – större betydelse än diskrepansen medan ett litet värde fungerar omvänt. Värdet för λ bestämmer man antingen subjektivt eller med någon metod. Vi använder oss av ett förfaringssätt som kallas för approximativ korsvalidering som vi beskriver senare.

Allmänt har vi att

$$\begin{aligned} s(x) &= \sum_{k=1}^{r+2} \beta_k B_k(x) \\ &= \beta_1 B_1(x) + \ldots + \beta_{r+2} B_{r+2}(x) \end{aligned}$$

och därmed att

$$\begin{aligned} (s''(x))^2 &= (\beta_1 B_1''(x) + \ldots + \beta_{r+2} B_{r+2}''(x))^2 \\ &= \sum_{k=1}^{r+2} \sum_{l=1}^{r+2} \beta_k B_k''(x) \beta_l B_l''(x) \end{aligned}$$

Vi sätter

$$\Omega_{kl} = \int_a^b B_k''(x) B_l''(x) dx$$

och får

$$\Delta(s(x;\boldsymbol{\beta})) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + \lambda \sum_{k=1}^{r+2} \sum_{l=1}^{r+2} \beta_k \beta_l \Omega_{kl}$$

Beroende på vilken fördelning vi nu antar för $\boldsymbol{y}$ och vilken länkfunktion vi väljer får vi olika $\Delta(s(x;\boldsymbol{\beta}))$ att minimera. Ett enklare fall än Poisson- och gamma-fallen är att anta att $\boldsymbol{y}$ är normalfördelad och välja identitetslänk. Vi visar nu härledningen för det fallet, för att visa principen. Vi deriverar $\Delta(s(x;\boldsymbol{\beta}))$ med avseende på β_l för att få de partiella derivatorna.

$$\frac{\partial \Delta}{\partial \beta_l} = -2 \sum_i w_i \left(y_i - \sum_{j=1}^{m+2} \beta_j B_j(x_i) \right) B_l(x_i) + 2\lambda \sum_{j=1}^{m+2} \beta_j \Omega_{jl}$$

Antag att vi har skapat närmevärden $(z_k)_{k=1}^m$ för variabeln. Då får vi

$$-2 \sum_i w_i \left(y_i - \sum_{j=1}^{m+2} \beta_j B_j(x_i) \right) B_l(x_i) =$$

$$-2 \sum_{k=1}^{m} w_{z_k} \left(y_{z_k} - \sum_{j=1}^{m+2} \beta_j B_j(z_k) \right) B_l(z_k)$$

Låter vi de partiella derivatorna vara lika med noll får vi ekvationerna

$$\sum_{k=1}^{m} \sum_{j=1}^{m+2} w_{z_k} \beta_j B_j(z_k) + \lambda \sum_{j=1}^{m+2} \beta_j \Omega_{jl} = \sum_{k=1}^{m} w_{z_k} y_{z_k} B_l(z_k)$$

för $l = 1, \ldots, m+2$. Detta löser man vanligtvis numeriskt.
I modellen för nyckeltalens väntevärden använder vi en log-länk och då får vi

inte ett linjärt ekvationssystem. I [1] visar man att man kan iterera fram en lösning för de fallen. Skrivet på matrisform blir ekvationsystemet.

$$\boldsymbol{\beta}^{(n+1)} = (\boldsymbol{B^t W^{(n)} B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B^t W^{(n)} y^{(n)}}$$

där

$$\boldsymbol{B} = \begin{bmatrix} B_1(z_1) & \dots & B_{r+2}(z_1) \\ \vdots & \ddots & \vdots \\ B_1(z_m) & \dots & B_{r+2}(z_m) \end{bmatrix},$$

$\boldsymbol{W^{(n)}}$ är en diagonalmatris med omskalade vikter enligt nedan, $\boldsymbol{y^n}$ är en vektor med omskalade observationer och $\boldsymbol{\Omega}$ är en kvadratisk matris med Ω_{kl} som element. Då Y_i antas vara poissonfördelad är vikterna $w_{z_k}\exp(s^{(n)}(z_k))$ och observationerna $\frac{y_{z_k}}{\exp(s^{(n)}(z_k))} - 1 + s^{(n)}(z_k)$. Antas däremot Y_i gammafördelad är vikterna $\frac{w_{z_k}y_{z_k}}{\exp(s^{(n)}(z_k))}$ och observationerna $1 - \frac{\exp(s^{(n)}(z_k))}{y_{z_k}} + s^{(n)}(z_k)$.

4.5 Approximerad korsvalidering

Vi beskriver nu metoden med approximerad korsvalidering att välja λ i parameterskattningar för en spline. Låt y_{z_k} beteckna nyckeltal för observationer med närmevärde z_k. Antag att vi tar bort z_k och y_{z_k} ur datamaterialet och därefter anpassar en spline $s_k^\lambda(x)$ till resterande datapunkter. Med ett bra värde på λ bör väntevärdet $\mu(x) = s_k^\lambda(x)$ kunna ge en rimlig prediktion av det borttagna nyckeltalet y_{z_k}. Detta bör gälla för alla k och ett mått för detta kan vara diskrepansen

$$C(\lambda) = D(\boldsymbol{y}_{z_k}, \boldsymbol{s}_k^\lambda)$$

Idén är att välja det λ för vilket $C(\lambda)$ minimeras för $k = 1 \dots m$. Man kan visa [1] att med

$$\boldsymbol{A}^{(n)} = \boldsymbol{B}(\boldsymbol{B^t W^{(n)} B} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B^t W^{(n)}}$$

och omskalade vikter och observationer enligt ovan (för de olika fördelningarna) i iteration n är uttrycket

$$C^{(n)}(\lambda) = \sum_{k=1}^{m} w_{z_k}\left(\frac{y_{z_k} - s^{(n)}(z_k)}{1 - a_{kk}^{(n)}}\right)^2$$

en approximation till korsvalideringen. Här är $s^{(n)}$ splinen för hela datamaterialet där man på något sätt funnit ett minimerande λ. Detta används till nästa iteration och söks sedan på nytt.

4.6 GAM i R

Till R finns ett tilläggsbibliotek kallat `fda`[6]. Genom

```
basis <- create.bspline.basis(..., breaks = ..., norder = 4)
```

skapar vi först ett objekt som innehåller B-splinesfunktioner. Ett av argumenten, `breaks`, är för att specifiera knoparna $u_1, \ldots, u_r$ i definitionen för B-splines. `norder = 4` talar om att vi tänker använda kubiska splines. Med `plot(basis)` kan man se basfunktionerna.

För att explicit räkna fram $\boldsymbol{B}$ använder vi `eval.basis` med objektet som vi nyss skapade:

```
B <- eval.basis(evalarg = ..., basisobj = basis, ...)
```

där `evalarg` är i vilka punkter vi evaluerar B-splines. Man kan även få fram funktionsvärden för B-splines direkt med funktionen `bsplineS`. Vi vill även få fram matrisen $\boldsymbol{\Omega}$. Det gör vi enkelt med

```
Omega <- bsplinepen(basisobj = basis, ...)
```

Därefter skapar vi en funktion

```
1  betaSolver <- function(...) {
2    basis     <- create.bspline.basis(..., breaks = ..., norder = 4)
3    B         <- eval.basis(evalarg = ..., basisobj = basis)
4    Omega     <- bsplinepen(basisobj = basis)
5    Beta      <- innerBetaSolver(..., B, Omega, ...)
6    return(Beta)
7  }
```

där `innerBetaSolver` är en annan funktion som löser ekvationsystem i förra avsnittet. För korsvalideringen använder vi 1000 förbestämda värden på λ per iteration och variabel inom ett visst intervall som vi testar oss fram till. Vi observerar att λ kan vara stort ($\approx 10^{12}$) och litet (≈ 1) beroende på vilken variabel man arbetar med.

Vill man göra det lätt för sig kan man använda ett annat tilläggspaket för R skapat av Wood [2], kallat `mgcv`, för GAM. Det inkluderar en funktion `bam(...)` speciellt framtagen för generaliserade additiva modeller på stora datamängder. Med `gam.plot(...)` kan man därefter rita varje spline.

[6] Functional Data Analysis. Se [4].

Tabell 5: Faktorvariabler och kontinuerliga variabler i vår GAM.

Faktorindex	Faktorvariabel	Nivå	Nivåbeskrivning
1	Kön	1	Kvinna
		2	Man
2	Körsträcka	1	0 – 1000 mil/år
		2	1000 – 1500 mil/år
		3	1500 – 2000 mil/år
		4	2000 – 2500 mil/år
		5	2500 mil/år eller mer
Variabelindex	Kontinuerlig variabel	Närmevärden (z_{jk})	Knopar (u_{jk})
3	Ägarålder	Hela år	$18, 19, 20, \ldots, 99$
4	Fordonsålder	Hela år	$0, 1, 2, \ldots, 35$
5	Fordonsvikt	Närmsta 50 kg	$600, 650, 700, \ldots, 3200$
6	Fordonets vikt/effekt	Hela kg/hk	$4, 5, 6, \ldots, 42$

4.7 Skattade relationstal

Vår väntevärdesstruktur ser nu ut på följande sätt.

$$\begin{aligned}\log(\mu_i) &= \boldsymbol{X_i\beta} \\ &+ \sum_{k=1}^{r_3+2} \beta_{3k} B_{3k}(x_{i3}) + \sum_{k=1}^{r_4+2} \beta_{4k} B_{4k}(x_{i4}) \\ &+ \sum_{k=1}^{r_5+2} \beta_{5k} B_{5k}(x_{i5}) + \sum_{k=1}^{r_6+2} \beta_{6k} B_{6k}(x_{i6})\end{aligned}$$

Parametrar för kön och körsträckecell ingår i $\boldsymbol{\beta}$ medan parametrar för fordonsägarens ålder, fordonets ålder, fordonets vikt och fordonets vikt/effekt ingår i respektive summa av B-splines. Närmevärden $\{z_j\}$ och knopar $\{u_j\}$ för varje variabel finns i tabell 5. Skattningsförfarandet går till enligt följande.

1. Initiala parametrar för en variabel i taget – som ensam förklarande variabel – skattar vi antingen med ML-ekvationer (faktorer) eller diskrepans med straffterm (kontinuerliga variabler). Vi gör detta för alla variabler utom en.

2. Nyckeltal och vikter skalar vi om enligt ett förfarande som finns beskrivet i [1] när man har flera faktorer och kontinuerliga variabler i modellen, innan vi gör skattningar för variabeln vars parametrar inte skattades i (1) . Idén

är att överföra skattningar på ett fall med enbart en förklarande variabel. Det är samma sak som att betrakta parametrar för övriga variabler som en offset.

3. Nyckeltal och vikter skalas om på nytt för nästa variabel (som ensam förklarande variabel och på samma sätt som i punkt 2). Relationstal för den variabeln skattas nu om.

4. (3.) upprepas för alla variabler tills dess att skattningarna konvergerat enligt något kriterium, exempelvis att skillnaden mellan relationstalen i en iteration och iterationen efter är mindre än något ϵ.

Härledningar för hur β_{jk} skattas med flera kontinuerliga variabler och faktorvariabler för skadefrekvens och medelskada finns alltså beskrivet i [1].

En generaliserad additiv modell med både faktorvariabler och kontinuerliga variabler ger skattningar av relationstalen i tabell 6 och för var och en av de kontinuerliga variablerna, en spline i figurer 3 – 10 nedan. Vi noterar att de relationstal som ligger en bit från splinen uteslutande är relationstal med låg vikt men med förhållandevis många skador eller hög skadekostnad. Det är främst omkring minimum och maximum för variablerna som kurvorna blir osäkra p.g.a. vi har få observationer där. Vi ser att vi fångar intressanta fenomen som man missar i tidigare tariffer. Bland annat att medelskadan för fordonets ålder till en början ökar stadigt (upp till 10 år) för att sedan minska ner till ett minimum och sedan höjs ånyo. Rent intuitivt borde det istället vara en minskande kurva; kanske är det så att försäkringsbolaget är för frikostiga med skadekrav för bilar som inte är helt nya. I övrigt verkar kurvorna rimliga. Notera att bascellens väntevärde i tabell 6 är beroende av val av basnivåer för både faktorvariabler och kontinuerliga variabler och det är därför de skiljer sig avsevärt från förra avsnittet. Där gav vi även ett exempel på vad en försäkringstagare med vissa egenskaper får för riskpremie. Vi tittar nu på en försäkringstagare med samma egenskaper. Siffrorna i exemplet kommer från tabell 6 och varje spline (som vi inte redovisar explicit utom i grafisk form).

Exempel 4.3. En man som fyllt 31 år och som försäkrar ett nytt fordon, vilket körs strax under 1500 mil per år, med vikt kring 1400 kg och med 15 kg per hk får skattad skadefrekvens

$$0.0968 \cdot 1.1085 \cdot 1.0011 \cdot 0.7380 \cdot 0.9481 \cdot 1.0064 \cdot 1.0000 \approx 0.0756$$

och medelskada

$$19327 \cdot 1.0656 \cdot 1.0029 \cdot 0.7847 \cdot 1.0177 \cdot 0.9752 \cdot 1.0000 \approx 16085.46$$

vilket ger riskpremie

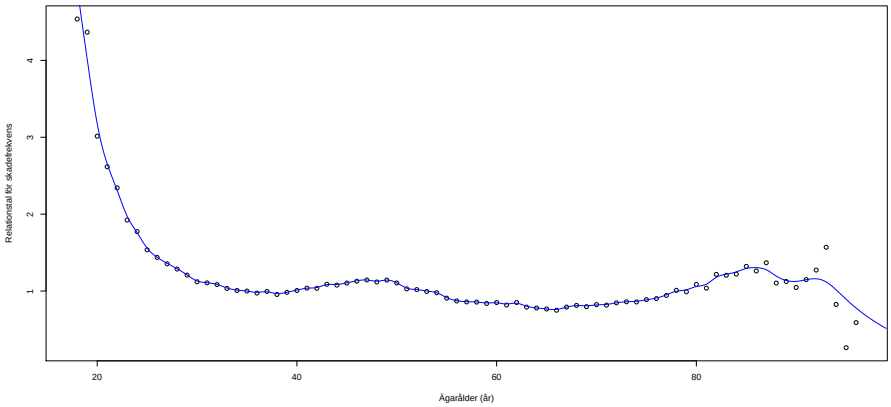
$$16085 \cdot 0.0756 \approx 1217 \text{ kronor}$$

Det betyder att denna kund förmodligen var subventionerad av andra kunder i den tidigare tariffen (tabell 4).

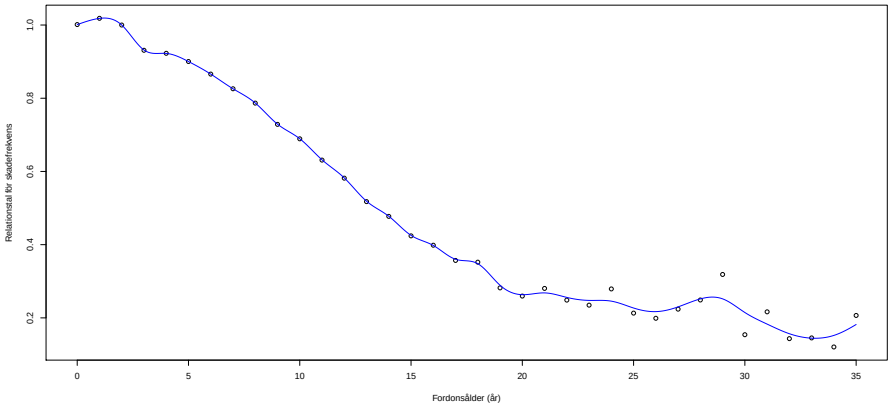
I samband med förra tariffen gav vi även ett exempel på att en manlig kund som precis fyllt 30 år minskar sin riskpremie med över 60 procent. I den här tariffen minskar han bara sina kostnader med strax över åtta procent om man räknar från dess att han precis fyllt 29 år. Här har vi dessutom alternativet att kurvintegrera splinen och låta försäkringstagare förändra sina kostnader kontinuerligt.

Tabell 6: Basnivåer för varje faktor är skrivna med fetstil.

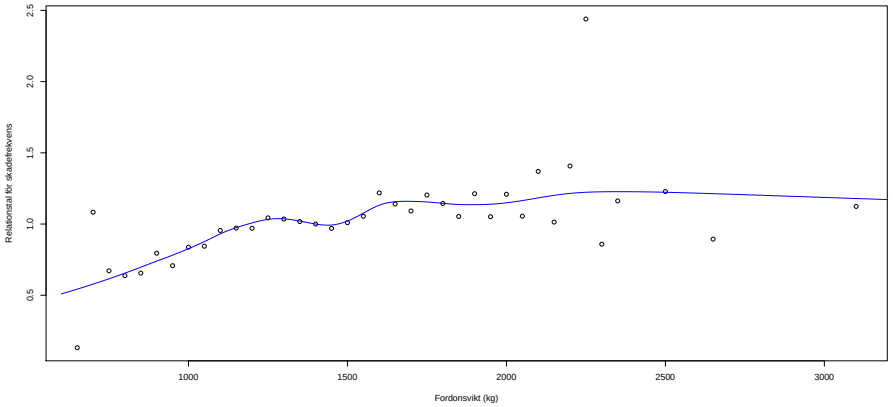
		Tariff		
		Relationstal		
Faktor	Nivåbeskrivning	Skadefrekvens	Medelskada	Riskpremie
Kön	**Kvinnor**	1.0000	1.0000	1.0000
	Män	0.9481	1.0177	0.9649
Körsträcka	**0 – 1000 mil/år**	1.0000	1.0000	1.0000
	1000 – 1500 mil/år	1.0064	0.9752	0.9814
	1500 – 2000 mil/år	1.1173	0.9622	1.0751
	2000 – 2500 mil/år	1.2150	0.9805	1.1913
	2500 mil/år eller mer	1.4406	1.0473	1.5087
μ_0		0.0968	19327	1929



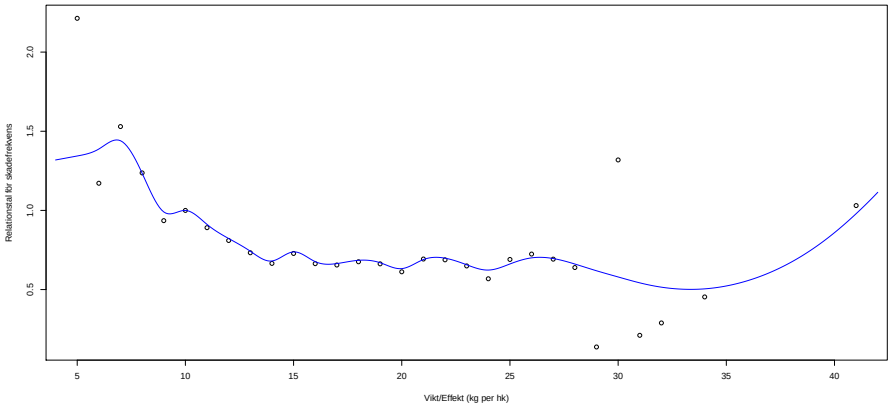
Figur 3: Kontinuerliga relationstal för skadefrekvensen för ägarens ålder.



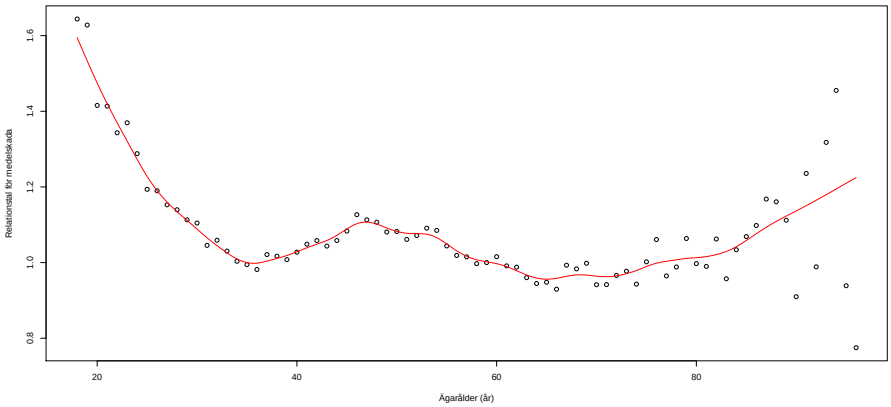
Figur 4: Kontinuerliga relationstal för skadefrekvensen för fordonets ålder.



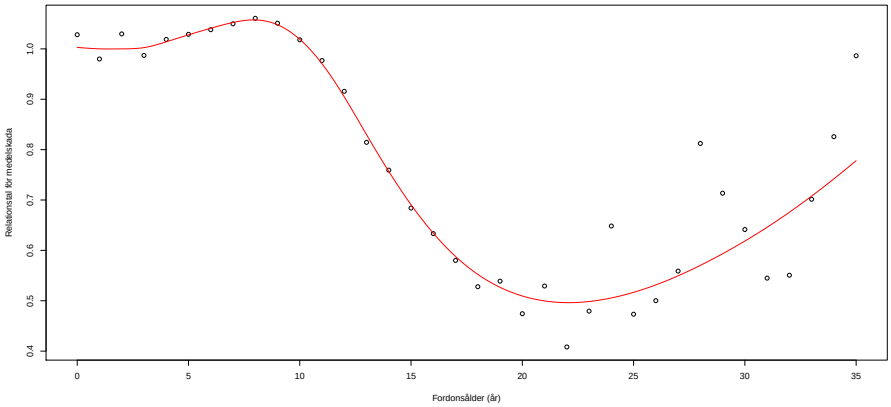
Figur 5: Kontinuerliga relationstal för skadefrekvensen för fordonets vikt.



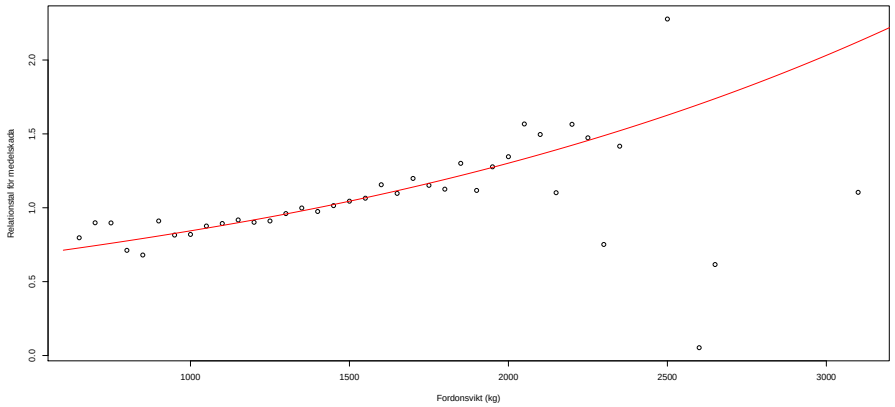
Figur 6: Kontinuerliga relationstal för skadefrekvensen för fordonets vikt per effekt.



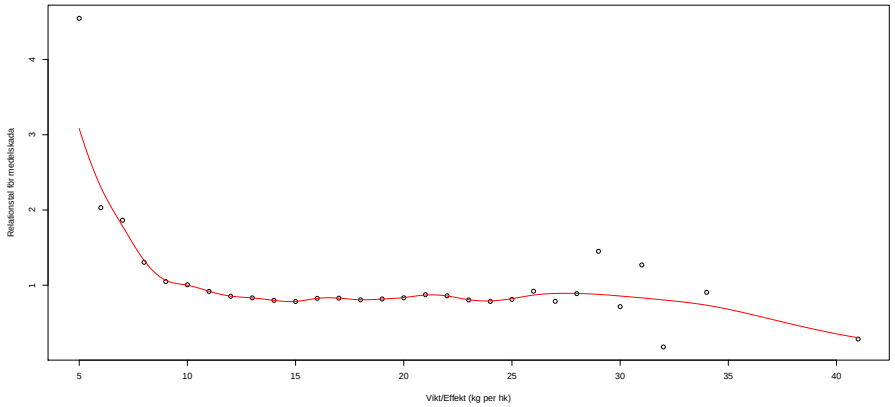
Figur 7: Kontinuerliga relationstal för medelskadan för ägarens ålder.



Figur 8: Kontinuerliga relationstal för medelskadan för fordonets ålder.



Figur 9: Kontinuerliga relationstal för medelskadan för fordonets vikt.



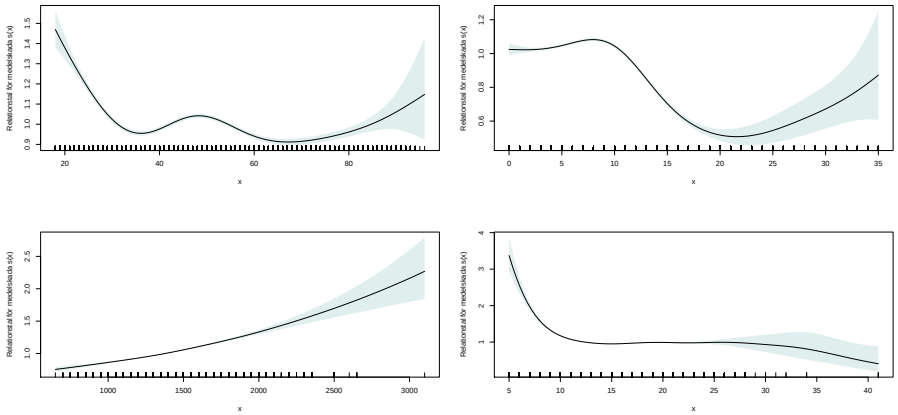
Figur 10: Kontinuerliga relationstal för medelskadan för fordonets vikt per effekt.

5 Diskussion

I detta arbete utgår vi från två relaterade modeller och implementerar m.h.a. programmeringsspråket R två stora program för att beräkna tariffer. Även om R inte har en alldeles självklar syntax så vill författaren ändå framföra åsikten att R har en mer intuitiv och modern syntax än SAS. Inkluderar man även det faktum att både R och en mängd olika grafiska gränssnitt (exempelvis RStudio) är gratis så är R definitivt något man kan ersätta SAS med när det kommer till statistisk analys och implementering av egna modeller. Det finns även tilläggsbibiliotek till R för att speciellt kunna arbeta med stora datamängder för GLM och GAM (och givetvis även den vanliga linjära modellen) men även SQL varför det lämpar sig för försäkringsdata. Med vårt egenhändigt programmerade GAM-program tar det mindre än tio minuter att få fram splines och relationstal för alla variabler vilket kan anses acceptabelt. Å andra sidan tar det mindre än tio sekunder med tilläggsbiblioteket `mgcv`.

Den generaliserade additiva modellen med splines använder vi när vi vill använda oss av kontinuerliga variabler och inte enbart faktorvariabler. Anledningarna var dels att man kan misstänka att vissa av dem har en underliggande kontinuerlig kurva för nyckeltalen, dels att få en mer konkurrenskraftig tariff. Figurerna i föregående avsnitt ger oss även ett par outliers som kan vara intressanta att undersöka vidare, bland annat relationstalen för vissa fordonsvikter och fordonseffekter. Dessa skulle vi möjligtvis ha missat i tariffen baserad på en vanlig GLM. En del naturliga frågeställningar dyker även upp. Hur mycket bör man lita på kurvorna i början och slutet av variablernas definitionsmängder? Här hade det varit bra med konfidensområden (något som är enkelt att få med Woods [2] tilläggsbibiliotek). Kan man kurvintegrera B-splines på ett enkelt sätt för att få en riskpremie som ändras på ett ”naturligt” sätt för en försäkringstagare under giltighetstiden?

Det finns hur mycket som helst att undersöka och inkludera i dessa modeller. Vi har inte tagit med konfidensintervall för relationstalen eller testat om varje faktor ska ingå eller ej i modellerna. Vi har ej heller gjort någon modelldiagnostik. Man kan även tänka sig att inkludera samspel mellan faktorerna och kredibilitet i en GAM. Från början inkluderade vi även en tredje modell med kredibilitetsskattningar för olika bilmärken. På grund av tidsbrist hann vi dock ej inkludera denna i texten. Däremot har vi jämfört (figur 11) parameterskattningar och grafer med de man får m.h.a. tilläggspaketet i R för GAM skapat av Simon N. Wood [2] – och de stämmer bra överens!



Figur 11: Kontinuerliga relationstal med konfidensintervall för medelskadan skapade med tilläggsbiblioteket mgcv för GAM av Wood [2].

6 Appendix

6.1 Datamaterial

Tabell 7: Utdrag ur datamaterialet.

i	Kön	Ägarålder	Fordonets ålder	Tillverkare
1	K	18	1	74
2	K	18	3	1
3	K	18	5	21
4	K	18	5	65
5	K	18	7	2
⋮	⋮	⋮	⋮	⋮
i	Körsträckecell	Fordonsvikt	Fordonets vikt per effekt	
1	1	1104	16.5	
2	1	1602	14.2	
3	1	1533	20.0	
4	3	1115	10.7	
5	1	1060	19.3	
⋮	⋮	⋮	⋮	
i	Duration	Antal skador	Kostnad	
1	0.25000	0	0	
2	0.45000	0	0	
3	0.03611	0	0	
4	0.69166	1	24872	
5	0.21667	0	0	
⋮	⋮	⋮	⋮	

I tabellen ovan ger vi de fem första raderna ur datamaterialet. Man kan låta en variabel vara en faktor med olika nivåer (beroende på val av modell). I sådant fall transformeras värdena för variabeln om till nivåbeteckningen i tabellen och faktorn får ett antal dummy-variabler i designmatrisen $\boldsymbol{X}$.

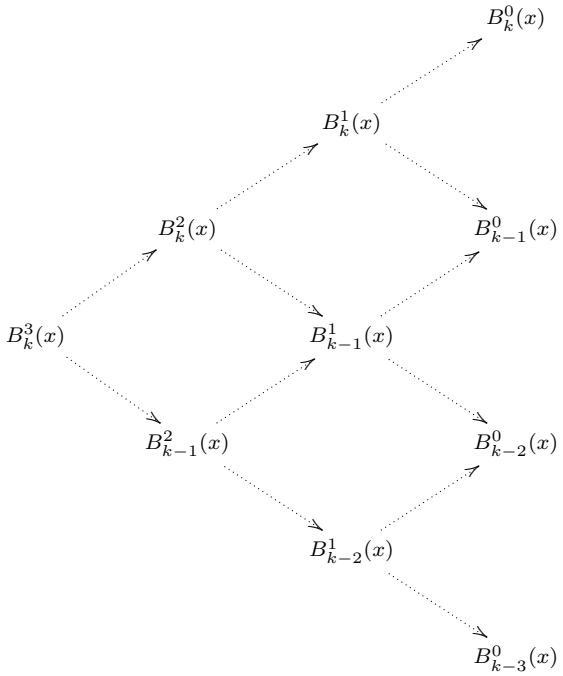
Ett enkelt sätt att snabba upp beräkningar är att lägga rader med skadekostnader i en separat tabell och skatta parametrar för medelskadan från den och att ha separata program för de två nyckeltalen.

6.2 Diverse funktioner i R

Vi ger här ett urval av användbara funktioner i R i alfabetisk ordning som vi använder i programmeringen för att beräkna tarifferna.

Funktion	Funktionsbeskrivning
`aggregate(...)`	Aggregerar data enligt någon funktion. Exempelvis summation.
`as.factor(...)`	Ändrar typ till faktor för ett objekt.
`cbind(...)`	Sätter ihop två objekt på bredden. Exempelvis två kolonner i en matris eller data frame.
`count(...)`	Räknar unika värden av en variabel och skapar en data frame med antalet förekomster av dessa värden. Tillhör paketet `plyr`.
`cut(...)`	Delar upp en numerisk variabel i intervall och gör en faktor med nivåer som motsvarar var och ett av intervallen.
`ddply(...)`	Applicerar en funktion för varje delmängd av en data frame och kombinerar resultatet i en data frame. För exempelvis beräkning av funktionsvärden när data är på listform. Tillhör paketet `plyr`.
`fitted.values(...)`	Ger predikterade värden.
`glm(...)`	Ger ett objekt av typen GLM. Bland attributen för objektet finns parameterskattningar, predikterade värden och konfidensintervall.
`ifelse(...)`	Istället för att behöva skriva en `if`- och `else`-sats.
`merge(...)`	Slår ihop två data frames.
`relevel(...)`	Sätter basnivå.
`subset(...)`	Ger en delmängd av ett objekt. Exempelvis en data frame.
`with(...)`	Arbetar i en miljö, exempelvis en data frame och modifierar denna
`within(...)`	eventuellt.

6.3 B-splines av grad 3



Vi visar resultatet men utelämnar stegen i uträkningen av B-splines av grad 3. Enligt Cox-de Boors formel ger varje B-spline upphov till två nya B-splines men av lägre grad som i diagrammet ovan. Tredjegradspolynomet blir

$$
\begin{aligned}
B_k^3(x) &= \frac{x-u_{k-3}}{u_k-u_{k-3}}\frac{x-u_{k-3}}{u_{k-1}-u_{k-3}}\frac{x-u_{k-3}}{u_{k-2}-u_{k-3}}B_{k-3}^0(x)\\
&+ \frac{x-u_{k-3}}{u_k-u_{k-3}}\frac{x-u_{k-3}}{u_{k-1}-u_{k-3}}\frac{u_{k-1}-x}{u_{k-1}-u_{k-2}}B_{k-2}^0(x)\\
&+ \frac{x-u_{k-3}}{u_k-u_{k-3}}\frac{u_k-x}{u_k-u_{k-2}}\frac{x-u_{k-2}}{u_{k-1}-u_{k-2}}B_{k-2}^0(x)\\
&+ \frac{u_{k-1}-x}{u_{k+1}-u_{k-2}}\frac{x-u_{k-2}}{u_k-u_{k-2}}\frac{x-u_{k-2}}{u_{k-1}-u_{k-2}}B_{k-2}^0(x)\\
&+ \frac{x-u_{k-3}}{u_k-u_{k-3}}\frac{u_k-x}{u_k-u_{k-2}}\frac{u_k-x}{u_k-u_{k-1}}B_{k-1}^0(x)\\
&+ \frac{u_{k-1}-x}{u_{k+1}-u_{k-2}}\frac{x-u_{k-2}}{u_k-u_{k-2}}\frac{u_k-x}{u_k-u_{k-1}}B_{k-1}^0(x)\\
&+ \frac{u_{k+1}-x}{u_{k+1}-u_{k-2}}\frac{u_{k+1}-x}{u_{k+1}-u_{k-1}}\frac{x-u_{k-1}}{u_k-u_{k-1}}B_{k-1}^0(x)\\
&+ \frac{u_{k+1}-x}{u_{k+1}-u_{k-2}}\frac{u_{k+1}-x}{u_{k+1}-u_{k-1}}\frac{u_{k+1-x}}{u_{k+1}-u_k}B_k^0(x)
\end{aligned}
$$

Utanför de fyra intervallen $[u_{k-3}, u_{k-2})$, $[u_{k-2}, u_{k-1})$, $[u_{k-1}, u_k)$ och $[u_k, u_{k+1})$ är $B_k^3(x)$ lika med 0. Vi ser att varje intervall har ett eget polynom pga. trapp-

funktionerna $B^0(x)$. Situationen kompliceras dock av att själva splinen $s(x)$ är en summa av B-splines (med tillhörande koefficienter).

Referenser

[1] E. Ohlsson, B. Johansson *Non-Life Insurance Pricing with Generalized Linear Models.*
Springer, 2010.

[2] Simon N. Wood, *Generalized Additive Models: An Introduction with R.*
Chapman and Hall, 2006.

[3] Michael J. Crawley, *The R Book.*
Wiley, 2007.

[4] J. O. Ramsay et al, *Functional Data Analysis.*
GPL, 2012.