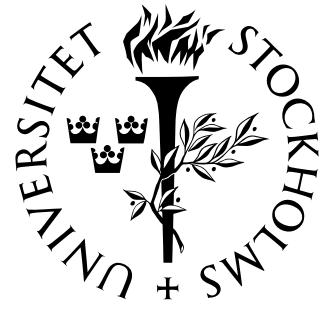

Matematisk statistik
Stockholms universitet

Mått på kvarstående systematisk variation mellan individer efter indelning i premieklasser inom sakförsäkring

Kajsa Järnmalm

Examensarbete 2006:15

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

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

Matematisk statistik
Stockholms universitet
Examensarbete **2006:15**,
http://www.math.su.se/matstat

Mått på kvarstående systematisk variation mellan individer efter indelning i premieklasser inom sakförsäkring

Kajsa Järnmalm*

november 2006

Sammanfattning

Inom skadeförsäkring används i allmänhet generaliserade linjära modeller (GLM) för att beräkna risker och korrekt premienivå. För att kunna uppskatta den risk det innebär att försäkra en specifik individ innehåller GLM vissa informativa variabler, såsom kön, ålder och geografisk zon, vars värden hjälper oss bedöma individernas beteende och vidare deras premier. Syftet med det här examensarbetet är att undersöka den kvarstående systematiska variationen mellan individer efter att de delats in i premieklasser som bestäms beroende av värdena på dessa variabler. För att förtydliga presenteras en metod med vilken vi kan bestämma tariffens förklaringsgrad, eller hur mycket av individernas beteende vi kan förklara med hjälp av den tariff vi använder oss av idag. Redovisade resultat är baserade på beräkningar gjorda på data ur Ifs kunddatabas för privata trafikförsäkringar i Sverige.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91 Stockholm. E-post: kajsa@jarnmalm.se Handledare: Bengt Eriksson (If Skadeförsäkring) och Ola Hössjer.

Measures of the remaining systematic variance between individuals when divided into individual premium groups in non-life insurance

Kajsa Järnmalm*

June of 2006

Abstract

In non-life insurance, generalized linear models (GLMs) are commonly used to estimate risk and calculate a correct tariff. To determine the risk of insuring a specific individual the GLM contains certain variables such as sex, age and geographic area, the values of which help to calculate a certain individuals behaviour and through that their unitary premium. The purpose of this thesis is to investigate the still remaining variation between individuals after they have been divided into individual premium groups, depending on the value of these variables. To further amplify, it will present a method with which we can determine the degree of explanation of the tariff or how much of the individual's behaviour we can account for by the use of our model. My findings are a result of calculations made with the use of information in If's Data Warehouse for private liability car insurance in Sweden.

* Postal address: Dept. of Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden.
E-mail: kajsa@jarnmalm.se. Supervisors: Bengt Eriksson (If P&C Insurance) and Ola Hössjer.

Förord

Det här är ett examensarbete i matematisk statistik med inriktningen försäkringsmatematik, skriven på matematiska institutionen vid Stockholms universitet. Arbetet skrevs på uppdrag av If Skadeförsäkring under affärsområdet Privat. Jag vill innerligt tacka min handledare på If, Bengt Eriksson, som tillsammans med Vilhelm Luttemo (Ledare, motorförsäkringar Sverige) tagit initiativ till det här arbetet och som bidragit med sin kunskap och inspiration genom hela arbetsprocessen. Jag vill också tacka Ola Hössjer, professor på matematiska institutionen vid Stockholms universitet, som med enormt stort engagemang tagit sig an uppdraget som min handledare för det här examensarbetet. Till sist vill jag även passa på att tacka och skicka en varm hälsning till alla mina arbetskamrater och underbara människor på C3 på If i Bergshamra, som varit enormt välkomnande och som är en fantastisk samling människor allihop.

Innehåll

Kapitel 1

Introduktion

Tanken med en försäkring är att många delar på risken för något som sällan inträffar. Försäkringsbolagets kunder betalar en premie mot att försäkringsbolaget ersätter skador som inträffar under den avtalade försäkringsperioden. Trafikförsäkringen är en enligt trafikskadelagen (1975:1410) obligatorisk försäkring för alla motordrivna fordon som brukas i trafik i Sverige. Det kan gälla bil, mc, moped och även en del motorredskap. Trafikförsäkringen ersätter personskada som drabbar förare och passagerare i egna fordonet samt person- och sakskada som drabbar en eventuell motpart. Samtliga beräkningar utförda i det här examensarbetet är gjorda på If:s trafikförsäkringar för personbil i Sverige med hjälp av programvarupaketet SAS 9.01.

1.1 Bakgrund

Årspremien som kunden betalar för sin försäkring bör lämpligen motsvara väntevärdet av det belopp försäkringsgivaren kommer att betala ut till följd av försäkringstagarens skador under året (=risken). Lägger man sig på en illa beräknad prisnivå, som är för låg eller för hög, riskerar man antingen att inte täcka upp sina utgifter eller att skrämma bort prismedvetna kunder. För att kunna beräkna en korrekt premie fordras en uppskattning av den risk kunden innebär för företaget eller hur många skador han eller hon kan antas orsaka under försäkringsperioden. Denna risk bedöms utifrån vissa beskrivande variabler, kallade premieargument, som man anser påverkar sannolikheten för skada. Premieargumenten delas i sin tur in i *klasser* beroende av premieargumentens observerade värde. En utförlig redogörelse för hur premieargumenten och deras indelning ser ut lämnas till kapitel två.

På If använder man, som brukligt är inom skadeförsäkring, en generaliserad linjär modell (GLM) innehållande dessa omsorgsfullt utvalda förklaringsvariabler, eller premieargument, för att beräkna den så kallade *tariffen* som styr kundernas respektive premienivå. Den fråga som ligger till grund för det här examensarbetet är hur god tariffens kvalité egentligen är.

1.2 Syfte

Syftet med det här examensarbetet är att bedöma hur god tariffens kvalité är och att definiera ett mått och en metod med vilken vi kan beräkna denna. Ett lämpligt mått torde vara att beräkna den återstående variationen mellan individer efter att de delats in i sina respektive premiegrupper. Detta på grund av att om vi med hjälp av våra premieargument hade förklarat individernas fulla beteende, så hade premien varit perfekt beräknad och denna *excessvarians* skulle summera till noll. Ett annat sätt att tänka oss kvalitén på tariffen och som torde vara lättare för gemene man att förstå är att tänka i *förklaringsgrad*, ett tydligare mått på hur mycket vår modell förklarar av skillnaden mellan individer. En tänkvärd anmärkning är att om premien beräknats perfekt vore försäkringsbolagens bonussystem överflödigt. Det finns sedan tidigare rapporter som berör det här ämnet, men ingen som enligt min vetskap verkligen fokuserar på just de här frågorna eller ännu hellre på ett pedagogiskt och lättbegripligt sätt lyckas förklara hur dessa beräkningar kan utföras och varför de förtjänar vår uppmärksamhet. Syftet med detta examensarbete är följaktligen att fylla detta tomrum.

1.3 Målsättning

Målet med det här examensarbetet är att definiera ett lämpligt mått på den kvarstående variansen mellan individer då de delats upp i premieklasser och att beskriva en metod med vilken man kan beräkna tariffens förklaringsgrad. Jag ämnar också diskutera relevanta resultat i samband med dessa beräkningar.

Kapitel 2

Datamaterialet

Datamaterialet som analyserats i det här examensarbetet kommer från If:s kunddatabas över privata motorförsäkringar i Sverige, specifikt trafikförsäkringar för personbil. Ur detta register har vi valt att enbart betrakta försäkringar som utan avbrott är giltiga under en treårsperiod, inom perioden 1 januari 2002 t.o.m. 31 december 2005, vilket resulterat i n=439.283 observationer. Här har de observationer där en eller flera variabler innehållt saknade värden, så kallade *missing values*, eliminerats för att förenkla våra beräkningar. Skadorna som analyseras är trafikskador med ersättning. Skador där ”vår” förare inte är vållande räknas därigenom inte.

Med en observation menas en enskild försäkring och det totala antalet skador över treårsperioden för denna enskilda försäkring. I registret finns för varje försäkring uppgifter, inte endast om antalet skador, utan även ytterligare betydande information som beskriver kunden samt det försäkrade objektet. Här finns, förutom den självklara informationen om försäkringstyp, teckningsdatum, försäkrings- och kundnummer, exempelvis information om bilens ålder, körsträcka, kundens kön och geografiskt område. Detta är en del av det data man använder för att bygga tariffen och sedermera beräkna kundernas premier med.

2.1 Premieargumenten

Ett premieargument är en variabel som beskriver den försäkrade kunden eller det försäkrade objektet och som på så sätt hjälper försäkringsbolaget att bedöma risken för skada under försäkringsperioden. Premieargumenten delas i sin tur in i *klasser* beroende av premieargumentens observerade värde. Indelningen för vårt datamaterial redovisas i *tabell 2.1* på nästa sida. Premieargumenten svarar mot de variabler som innefattas i vår GLM och som sedermera hjälper oss att beräkna tariffen och premienivån.

Premieargument	Klass	Variabel	Klassbeskrivning
Kundår	1	0- 2	Vid den tidpunkt då vi gjort vårt
	2	3- 5	urval har kunden haft försäkring hos
	3	6-10	bolaget i så många år som intervallet
	4	11-	anger.
Geografisk zon	0		Vardera zon motsvarar ett
	1		geografiskt område i Sverige.
	2		Landet är indelat i 19 zoner efter
	3		postnummerkoder. Varje zon
	4		påverkar risken och därmed också
	5		premien positivt eller negativt.
	6		Som exempel är risken för skada
	7		högre i Stockholmsområdet än på
	8		landsbygden.
	9		
	10		
	11		
	12		
	13		
	14		
	15		
	16		
	17		
	18		
Bilålder	1	0- 2	Åldern på det försäkrade objektet.
	2	3- 5	
	3	6- 8	
	4	9-12	
	5	13-16	
	6	17-	
Premieklass	0	000	Premieklassen bestäms av bilmärke
	1	001	och -modell.
	2	002	
	3	003	
	4	004	
	5	005	
	6	006	
	7	007	
	8	008	
	9	009	*forts. på nästa sida*

forts. från föregående sida

Körsträcka	1 2 3 4 5		Ju högre variabelvärde, desto längre körsträcka har kunden angett för det försäkrade objektet.
Kön	K M	Kvinna	Kundens kön.
Ålder	1 2 3 4 5 6 7 8 9 10 11 12 13 14	0-19 24-24 25-26 27-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-	Kundens ålder.

Tabell 2.1 Premieargument för trafikförsäkring på personbil

2.2 Tariffen

Tariffen ligger som bas då man fastställer premien för försäkringen. Med nämnda premieargument som utgångspunkt beräknas en tariff fram som är direkt kopplad till den risk som försäkringen avser. Tariffen baseras på en generaliserad linjär modell, en så kallad GLM, och kan därför räknas fram med hjälp av standardproceduren ”*Proc Genmod*” i SAS. (Vår GLM beskrivs närmare i kapitel 3). Premien bestäms sedan utifrån en grundpremie som influeras av tariffen och premieklassernas respektive risknivå, samt eventuella rabatter eller bonusklasser.

Alla försäkringar vars observerade värden hamnar i samma intervall för varje premieargument, och därmed tillhör samma klass i samtliga premieargument, sägs tillhöra samma *tariffcell*. De försäkringar som tillhör samma tariffcell betraktas som likvärdiga ur prissättningssynpunkt. Man delar alltså in alla försäkringsavtal som har

lika stor risk i samma grupp, med målet att alla försäkringstagare betalar en premie som motsvarar den egna förväntade skadekostnaden. Poängen med en sådan indelning är också att skillnaden *mellan* olika tariffceller skall vara stor i jämförelse med skillnaden mellan individer inom *samma* tariffcell.

Kapitel 3

Generaliserade linjära modeller

Generaliserade linjära modeller (GLM) är en utvidgning av den traditionella linjära modellen, där den beroende variabeln kan antas följa andra mönster än normalfördelningens. GLM kan därför användas vid ett större urval av dataanalys-problem och är en rik klass av statistiska modeller inom vilken man kan finna specialfall som har tillämpningar inom olika delar av försäkringsmatematiken. Vanligt är diskreta fördelningar såsom binomial- eller poissonfördelning. Generellt gäller att alla fördelningar som tillhör familjen Exponentiella dispersionsmodeller (EDM) kan beskrivas genom GLM. Samtliga fördelningar med vilka man kan skriva frekvensfunktionen på formen

$$f_{Y_{il}}\left(y, \theta_i, \phi\right) = \exp\left\{\frac{y\,\theta_i - b\left(\theta_i\right)}{\phi / w_i} + c\left(y, \phi, w_i\right)\right\},$$

där $w \geq 0,\ \phi > 0,\ f_{Y_{il}}(0) = 0$

Vidare gäller

$\mu_i = E(Y_{il}) = b'(\theta_i)$, där Y_{lj} betecknar antal skador för individ l i tariffcell i och

$\mathrm{Var}(Y_{il}) = b''(\theta_i)\frac{\phi}{w}$, där $b''(\theta_i) = v(\mu_i)$ kallas variansfunktionen.

Vi låter n_i beteckna antalet individer i tariffcell i och Y_{il} antalet skador för individ l i tariffcell i, $i=1,2,\ldots,I$ $l=1,2,\ldots,n_i$. Det totala antalet individer ges alltså av $n=n_1+n_2+\ldots+n_I$, där I anger antal tariffceller.

I GLM tillåter man att väntevärdets linjära struktur skapas genom en länkfunktion, $g(\mu_i)=\eta_i=\mathbf{x}_i'\boldsymbol{\beta}$, där $\mathbf{x}_i=(x_{i1},x_{i2},\ldots,x_{ip})$ är den så kallade *designvektorn* av dummyvariabler som är identisk för alla individer i tariffcell *i,* och $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)$ är vektorn av regressionsparametrar. Länkfunktionens utseende bestäms beroende av vilken modell man önskar använda för att beskriva data. För den klassiska linjära modellen har vi exempelvis en så kallad *identitetslänk*, $g(\mu_i)=\eta_i=\mu_i$. Vissa länkfunktioner visar sig vara ”naturliga” för vissa fördelningar. Dessa länkfunktioner benämns *kanoniska* länkar och erhålls då $g(\cdot)=b'^{-1}(\cdot)$, vilket medför att $\eta_i=\theta_i$.

3.1 Modellen

För att SAS skall kunna skatta våra parametrar måste vi kommunicera önskad respons, samt vilka övriga variabler som är aktuella för modellering till programmet. I vårt fall ser modellen ut som

$$Antal\,skador = kundår\,zon\,bilålder\,premieklass\,körsträcka\,kön \times ålder$$

Detta innebär att *antal skador* är responsvariabel och att denna beror på samtliga variabler i högerledet. Multiplikationstecknet mellan variablerna kön och ålder beror på ett samspel dessa variabler emellan. Ty tariffen varierar inte enskilt efter om individen är man eller kvinna. Man menar istället att risken beror på kundens kön i kombination med dess ålder. Som exempel tillhör män i åldrarna 20-24 en högre riskklass än kvinnor i samma ålder, men de innebär också en högre risk än exempelvis män i åldrarna 35-39. Man kan alltså inte påstå att män som individuell grupp innebär en högre risk än kvinnor. Risken för skada beror på kön, men denna fluktuerar mellan åldersgrupperna för både kvinnor och män.

Låt γ_0 beteckna baspremien och $\gamma_{j1},\ldots,\gamma_{jk_j}$ de olika nivåerna för premieargument j, (j=1,…,6). Här anger k_j antalet nivåer för premieargument j, d.v.s. $k_1=4$, $k_2=19$, $k_3=6$, $k_4=10$ och $k_5=5$. Lägg märke till att premieargument sex svarar mot samspelsvariabeln $kön \times ålder$. Antalet nivåer för premieargument sex är således $k_6=2\times 14$. Låt i_j beteckna index för den nivå på premieargument j som svarar mot tariffcell i. Medan den klassiska linjära modellen är additiv, $\left(\mu_i=\gamma_0+\gamma_{1i_1}+\gamma_{2i_2}+\ldots+\gamma_{6i_6}\right)$, är vår modell multiplikativ, d.v.s, $\mu_i=\gamma_0\,\gamma_{1i_1}\,\gamma_{2i_2}\ldots\gamma_{6i_6}$. Anledningen till att vi vill använda oss av en multiplikativ modell är att vi är intresserade av att se den *relativa* förändringen och inte den *absoluta* förändringen hos väntevärdet.

Responsvariabel för vår modell är som nämnt *antal skador*, vilka följer en Poissonfördelning, som har en log-länk som kanonisk länk, $g(\mu_i)=\eta_i=\log(\mu_i)$. Genom

logaritmering får vi $\log(\mu_i)=\log(\gamma_0)+\log(\gamma_{1i_1})+\log(\gamma_{2i_2})+\ldots+\log(\gamma_{6i_6})$ och samma linjära struktur som i additiva modellen. Den här typen av modell kallas en loglinjär modell. Modellen parameteriseras så att man väljer en referenscell för varje premieargument. Dessa noteras $\gamma_{11}=\gamma_{21}=\gamma_{31}=\ldots=\gamma_{61}=1$ och blir noll vid logaritmering. Övriga parametrar mäter nu hur mycket risken avviker från referenscellen. Parametrarna γ_{1i} visar relationstalen för premieargument nummer ett, γ_{2i} för premieargument nummer två, o.s.v. Om vi introducerar beteckningarna $\beta_1=\log(\gamma_0)$, $\beta_2=\log(\gamma_{12})$, $\beta_3=\log(\gamma_{13})$, $\beta_4=\log(\gamma_{14})$, $\beta_5=\log(\gamma_{22})$, o.s.v., där vi utelämnat de parametrar som blir noll efter logaritmering, och inför

dummyvariablerna $x_{ij}=\begin{cases} 1 & \text{om } \beta_j \text{ ingår i } \mu_i \\ 0 & \text{annars} \end{cases}$,

kan vi skriva

$$E\left(Y_{il}\right)=\mu_i=exp\left(\eta_i\right)=exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) \quad , \quad l=1, 2, \ldots, n_i$$

som på ett naturligt sätt beskriver hur vi vill att vår GLM ska återge hur responsen Y_{il} påverkas av de förklarande variablerna, x_{ij}. Här betecknar p antalet parametrar i modellen, $p=1+\sum_{j=1}^{6}\left(k_j-1\right)=67$,

och I står för antalet tariffceller, $I=\prod_{j=1}^{6} k_j=638.400$.

Vi kan också skriva

$$\log\left(\mu_i\right)=\sum_{j=1}^{p} x_{ij}\beta_j \qquad i=1, 2, \ldots, I$$

för att få samma struktur som den linjära modellen, med skillnaden att μ är logaritmerad i vänsterledet.

3.2 Proc Genmod

Ett plus med att använda generaliserade linjära modeller är att man med fördel kan använda sig av standardproceduren Proc Genmod i SAS för parameterskattning med mera. Proceduren anpassar en generaliserad linjär modell till data genom maximum likelihood skattning av parametervektorn $\boldsymbol{\beta}$ och skattar modellens parametrar numeriskt iterativt. Det kan nämnas att ML-skattningarna för denna modell ger samma resultat som tillämpning av Jungs marginalsummemetod.
För att uppskatta vikten av huvudeffekterna i modellen kan man använda sig av de så kallade *type 1* och *type 3* optionerna i Proc Genmod för att låta genomföra statistiska test för signifikanserna av dessa variabler. Resultaten av dessa analyser redovisas i SAS som tabeller i outputfönstret:

Source	Deviance	DF	Chi-square	Prob>ChiSq
Intercept	163605			
kundår	163429	3	169,15	<0.0001
zon	162592	18	802,12	<0.0001
bilålder	162281	5	298,74	<0.0001
premieklass	161944	9	323,14	<0.0001
körsträcka	161764	4	173,03	<0.0001
kön x ålder	161190	27	550,46	<0.0001

Tabell 3.1: Resultat type1 analys

Source	DF	Chi-square	Prob>ChiSq
kundår	3	148,85	<0.0001
zon	18	863,84	<0.0001
bilålder	5	190,78	<0.0001
premieklass	9	294,18	<0.0001
körsträcka	4	255,99	<0.0001
kön x ålder	27	550,46	<0.0001

Tabell 3.2: Resultat type3 analys

I *Tabell 3.1* ovan representerar varje värde i ”Deviance” kolumnen deviansen för modellen innehållande variabeln på samma rad och samtliga variabler på raderna ovan denna. Den devians som anges på raden för kundår i tabellen är exempelvis deviansen för modellen inrymmande kundår och intercept. Ju fler termer som inkluderas i modellen, desto mindre blir deviansen. Kolumnen ”DF” redovisar antal frihetsgrader. Ett type 1 test innebär att man testar signifikansen för modellen så att man lägger till en variabel i taget. Man börjar således med att testa signifikansen för ett likelihoodkvottest som för varje variabel testar om den aktuella raden ska läggas till de ovanstående variablerna eller inte. De resulterande p-värdena (i sista kolumnen ”Prob>ChiSq”) på varje rad visar på signifikansen för en modell innehållande variabeln på den aktuella rad, givet variablerna på alla föregående rader. Vi kan tydligt se av tabellen för type 1 testet ovan att samtliga variabler i modellen är starkt signifikanta med ett p-värde på <0.0001. Type 3 analysen ger samma resultat med samtliga variabler starkt signifikanta enligt *Tabell 3.2* ovan. Här testas signifikansen av varje enskild variabel utesluten ur modellen en i taget, givet att samtliga övriga variabler är med i modellen. På så sätt prövas variablernas individuella bidrag till modellen. Att en variabel är starkt signifikant innebär att variabeln har stor betydelse då man vill bestämma skadefrekvensen hos de försäkrade.

Kapitel 4

Mått på kvarstående variation

4.1 Total variation i portföljen

Vi har, under den treårsperiod vi valt att iaktta, n=439.283 observationer. En observation motsvarar en enskild individs individuella försäkring under perioden och lämnar uppgift om antalet skador i det aktuella tidsintervallet. För att förenkla vår notation låter vi fortsättningsvis Y_i beteckna antal skador för individ i, i=1,2,…,n. *Tabell 4.1* nedan ger en överblick över individernas skadefrekvens under perioden. Vi ser att de flesta individer inte har några skador alls och att vi har färre observationer ju högre skadantal.

skadeantal	n = antal individer/observationer	totalt antal skador
0	411 495	0
1	26 264	26 264
2	1 436	2 872
3	83	249
4	5	20
	∑ **439 283**	**29 405**

Tabell 4.1: Skadefrekvens

För en modell utan indelning i tariffceller antas alla individer ha samma förväntat antal skador, $\mu=E(Y_i)$, och varians, $\sigma^2=Var(Y_i)$.

Väntevärdet skattas genom

$$\hat{\mu} = \sum_{i=1}^{n} \frac{Y_i}{n} = \frac{totalt\ antal\ skador}{n} = \frac{29.405}{439.283} = 0{,}0669386250$$

och variansen genom

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{\left(Y_i - \hat{\mu}_i\right)^2}{n} = \frac{n \times z - totalt\ antal\ skador^2}{n^2} \quad , där\ z = \sum_{i=0}^{4} skadeantal_i^2 \times n_i$$

$$= \frac{439.283 \times 32.835 - 29.405^2}{439.283^2} = 0{,}0702660238$$

Om skadorna hade genererats av en homogen Poissonprocess vore väntevärde och varians lika. Att $\hat{\sigma}^2 > \hat{\mu}$ innebär att vi har en excessvarians, v=Var(Y_i)–E(Y_i), som skattas genom

$$\hat{v} = \hat{\sigma}^2 - \hat{\mu} = 0{,}0703 – 0{,}0669 = 0{,}0033274$$

Excessvariansen kommer utav ”blandningen”, d.v.s. individernas skilda skadefrekvenser, och visar variansen i blandningsfördelningen. Poissonmodellen bygger på antagandet att skadorna för varje individ inträffar oberoende av varandra i tiden enligt en Poissonprocess, där varje individ har sin egen intensitet. I Poissonmodellen utan överspridning antas premieargumenten förklara de individuella skadeintensiteterna helt. I annat fall tillåts stokastiska skadeintensiteter, där slumpvariationen härrör från individuella egenskaper som vi inte registrerar, men antar påverkar skadeintensiteten.

Vi antar att $Y_i \sim Po(\Lambda_i)$, där Λ_i anger förväntat antal skador för individ i. Antal skador får en *blandad Poissonfördelning* om vi antar att Λ_i är en stokastisk variabel med $\mu = E(\Lambda_i)$. Detta ger

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{E}\left(Y(Y-1)\right) + \mathrm{E}(Y) - \mathrm{E}^2(Y) \\ &= \mathrm{E}\left(\Lambda^2\right) + \mathrm{E}(Y) - \mathrm{E}^2(\Lambda) \\ &= \mathrm{E}(Y) + \mathrm{Var}(\Lambda), \end{aligned}$$

där vi i andra ledet utnyttjar ekvation (4.5) i ref [3] för faktoriella moment hos en Poissonfördelning. Excessvariansen är alltså identisk med variansen i blandningsfördelningen, v=Var(Λ_i).

Att vi får en excessvarians är inte ett överraskande resultat, då vi är fullt medvetna om att portföljen inte är homogen. Det är naturligtvis tänkbart att det finns skillnader mellan exempelvis individerna i olika åldrar och från olika geografiska områden. Det bör samtidigt förtydligas att excessvarians kan orsakas även av andra företeelser än okända individuella egenskaper. Om vi exempelvis har ett modelleringsfel som står till följd av att premieargumentens effekt inte är riktigt multiplikativ, så kan detta fel ge upphov till excessvarians.

Vi antar att Λ_i har samma fördelning, $\Gamma(\alpha,\beta)$, med täthetsfunktion

$$u(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \quad \lambda > 0$$

Gammafördelningen motiveras av att en negativ binomialfördelning passar bra till vårt data (se resonemang kring *Tabell 4.2* resp. *4.3* nedan). Antal skador får nu en blandad Poissonfördelning, Y_i=Po(Λ_i), och med hjälp av gammafördelningen kan vi formulera sannolikheterna för att en slumpmässigt vald individ uppvisar k skador,

$$p_k(\alpha,\beta) = \int_0^{\infty} P\left(Y_i = k \,\middle|\, \Lambda = \lambda_i\right) u\left(\lambda_i\right) d\lambda$$

$$= \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\, n!} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{n}$$

Detta är en negativ binomialfördelning. För att närmare undersöka huruvida ovannämnda är en bra modell för datamängden kan vi utföra ett enkelt χ^2-test. Detta gör vi genom att först skatta observationstal i Poisson respektive negativ binomial, (se *Tabell 4.2* nedan) för att jämföra med det faktiska antalet observationer från Tabell 4.1 ovan. Observationstal för Poisson skattas genom

$$\hat{p}_k \times n = p_k\left(\hat{\mu}\right) \times n = \left(\frac{\hat{\mu}^k}{k!} e^{-\mu}\right) \times n \quad , k = 0,1, \ldots,4 \quad ,$$

där p_k (μ) anger sannolikhetsfunktionen $p(Y_i=k)$ då Y_i är Poissonfördelad med väntevärde μ.

Motsvarande för negativ binomial skattas med hjälp av sannolikhetsfunktionen ovan,

$$\hat{p}_k \times n = p_k\left(\hat{\alpha},\hat{\beta}\right)\times n = \frac{\Gamma\left(\hat{\alpha}+k\right)}{\Gamma\left(\hat{\alpha}\right)k!}\left(\frac{\hat{\beta}}{\hat{\beta}+1}\right)^{\alpha}\left(\frac{1}{\hat{\beta}+1}\right)^{k}\times n \qquad , k = 0,1, \ldots,4 \quad ,$$

där $\hat{\alpha}$ = 1,3466 och $\hat{\beta}$ = 20,12 är skattningar av α och β, erhållna genom momentmetoden (ref [3]).

I *Tabell 4.3* anges våra beräknade χ^2-värden. Dessa beräknas i enlighet med den principiella formeln

$$\chi^2 = \sum_{k=1}^{4} \frac{\left(n_k - \hat{p}_k \times n\right)^2}{\hat{p}_k \times n},$$

där n_k står för observerat värde och $\hat{p}_k \times n$ för förväntat värde, i vårt fall skattade observationstal med Poisson respektive negativ binomial.

Observerad **exponering = ant. individer/observ.**	**Poisson($\hat{\mu}$=0.0669)**	**NegBin(α=1.3466,β=20.12)**
411 495	410 841	411 503
26 264	27 501	26 241
1 436	920	1 458
83	21	77
5	0	4
439 283	**439 283**	**439 283**

Tabell 4.2 Skattade observationstal

Poisson($\hat{\mu}$=0.0669)	**NegBin(α=1.3466,β=20.12)**
1,04	0
55,65	0,02
288,77	0,33
189,97	0,46
63,08	0,27
598,52	**1,09**

Tabell 4.3 χ^2-värden

Det beräknade värdena för Poisson respektive negativ binomial i tabellen jämförs med p-värdet på 99,5%-ig nivå med tre respektive två frihetsgrader och fördelningen förkastas om vårt beräknade χ^2-värde överstiger motsvarande p-värde enligt $598{,}52 > \chi^2_{0.005}(3) = 12{,}8$ (förkastas) och $1{,}09 < \chi^2_{0.005}(2) = 10{,}6$ (förkastas ej). P-värdet hämtar man för respektive antal frihetsgrader ur en χ^2-tabell och antalet frihetsgrader för χ^2-fördelningen beräknas här som $antal\,celler - 1 - antal\,skattade\,parametrar$. Resultatet antyder att negativ binomial ger en god anpassning till data.
Med hjälp av värdena i *Tabell 4.1* kan vi beräkna variationskoefficienten i blandningsfördelningen genom formeln

$$CV = \frac{\sqrt{Var(\Lambda)}}{E(\Lambda)}$$

som skattas som

$$\hat{CV} = \frac{\sqrt{\hat{v}}}{\hat{\mu}} = \frac{\sqrt{0{,}0033}}{0{,}0669} = 86{,}17\% \ .$$

Denna skattning gäller oberoende av blandningsfördelning.

Eftersom vi anpassat en $\Gamma(1.3466,20.12)$ som blandningsfördelning för Λ och *CV* i gammafördelningen är lika med $\frac{1}{\sqrt{\hat{\alpha}}}$, kan *CV* även skattas som $\frac{1}{\sqrt{1{,}3466}} = 86{,}17\%$.

Variationskoefficienten (*CV, eng. coefficient of variation*) är ett spridningsmått som mäter hur stor standardavvikelsen är i förhållande till det aritmetiska medelvärdet och är alltså direkt relaterad till excessvariansen. Genom detta mått får vi en uppfattning om hur mycket Λ_i avviker från sitt väntevärdet, (0,0669). I ref [6] och ref [7] redovisas, på motsvarande sätt som i vår *Tabell 4.*1, ett antal tabeller avseende trafikskador från olika länder. Tabellerna finns återgivna i Appendix A med variationskoefficienten beräknad som ovan. Både högre och lägre värden för $\hat{CV}$ observeras i dessa tabeller. Riskspridningen mellan individer tycks därför inte vara avvikande i Sverige jämfört med andra länder.

4.2 Kvarstående variation i premiegrupper

När vi använder SAS-proceduren Proc Genmod för att skatta parametrarna i vår modell kan vi också utnyttja optionen ”*pred*” för att beräkna en premie som motsvarar varje enskild kunds förväntade skadeantal. Genom att sedan avrunda dessa värden till två decimaler får vi ett antal premiegrupper vars exponering redovisas nedan i *Tabell 4.4*.

Premiegrupp	Antal individer	Antal skador
0,00	10	0
0,01	1513	13
0,02	625	13
0,03	7464	226
0,04	35478	1448
0,05	78467	3934
0,06	100746	6092
0,07	84760	5945
0,08	57053	4517
0,09	35487	3236
0,10	20145	1963
0,11	9945	1058
0,12	4344	516
0,13	1818	221
0,14	837	122
0,15	391	60
0,16	126	23
0,17	43	10
0,18	15	5
0,19	8	2
0,20	4	1
0,21	3	0
0,22	1	0
Σ	**439283,00**	**29405,00**

Tabell 4.4 Exponering per premiegrupp

I samband med detta kan vi i *Diagram 4.1* nedan redogöra för en jämförelse mellan den skattade blandningsfördelningen, $\Gamma\left(\hat{\alpha},\hat{\beta}\right)$, (gråtonade staplar) och den beräknade GLM-tariffens fördelning (svarta staplar).

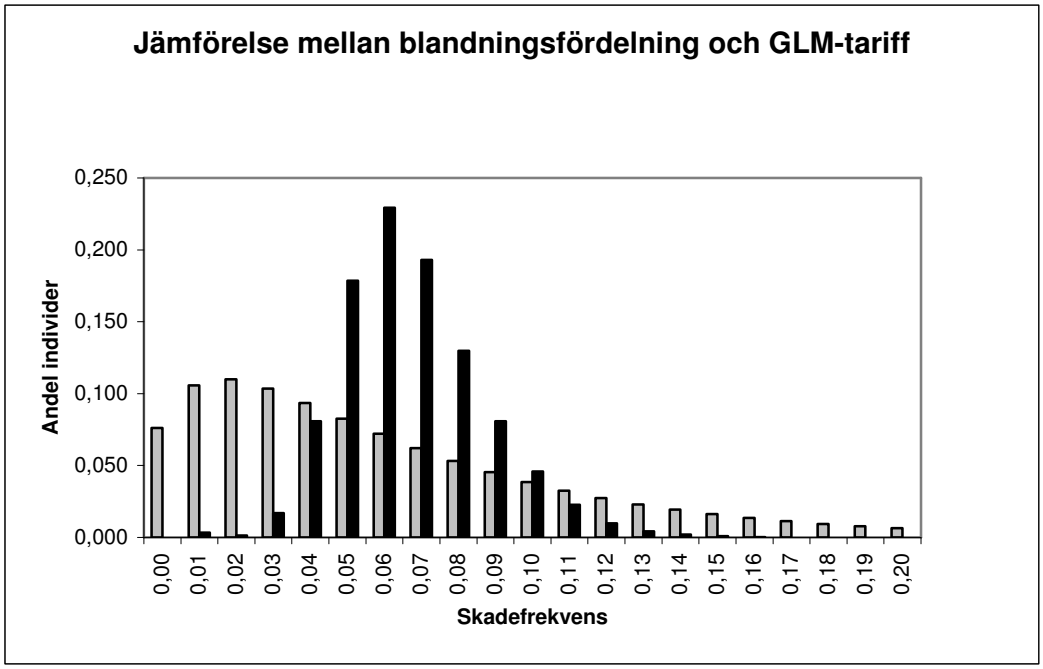

Diagram 4.1

Vi ser att GLM-premien ger en mindre spridning än den vi skattat för blandningsfördelningen. Variansen är 0,00038272 och motsvarande variationskoefficient är 28,87%. Varje stapel motsvarar andelen individer för en enskild premiegrupp.

Om vi nu återgår till *Tabell 4.4* ser vi exempelvis att de 100.746 individerna i den största gruppen (premiegrupp 0.06) i tabellen ovan har en skadefrekvens enligt GLM som för samtliga håller sig inom intervallet [0.055,0.065). Vi kan nu använda samma metod som tidigare för att kontrollera homogeniteten inom denna grupp.

skadeantal	n = antal individer/observationer	totalt antal skador
0	94964	0
1	5485	5 485
2	284	568
3	13	39
	100 746	**6 092**

Tabell 4.5 Tabellvärden för "premiegrupp" 0.06

Resultatet är

$$\hat{\mu} = \frac{totalt\ antal\ skador}{n} = \frac{6.092}{100.746} = 0,0605$$

$$\hat{\sigma^2} = \frac{n \times z - totalt\ antal\ skador^2}{n^2} \quad , \ där\ z = \sum_{i=0}^{4} skadeantal_i^2 \times n_i$$

$$= \frac{100.746 \times 6.738 - 6.092^2}{100.746^2} = 0,0632 \quad , \text{ vilket i sin tur ger en excessvarians}$$

$$\hat{v} = \hat{\sigma^2} - \hat{\mu} = 0,0632 - 0,0605 = 0,0027 \quad , \text{ och variationskoefficienten}$$

$$\hat{CV} = \frac{\sqrt{\hat{v}}}{\hat{\mu}} = \frac{\sqrt{0,0027}}{0,0605} = 85,89\%$$

Vi ser här att variationskoefficienten i blandningsfördelningen inom denna premiegrupp ligger väldigt nära den totala variationskoefficienten för hela portföljen (86,17%). Den relativa spridningen inom den enskilda gruppen är alltså nästan lika stor som spridningen i hela portföljen!
För varje grupp kan vi göra samma beräkning som vi gjort för gruppen med premienivå 0,06 ovan. Resultatet blir som följer av *Tabell 4.6* på nästkommande sida, där de premiegrupper som saknar tillräckligt underlag för beräkning av variationskoefficienten har tagits bort.

Premiegrupp	Antal individer	Antal skador	$\hat{CV}$
…	…	…	…
…	…	…	…
0,03	7464	226	158,34%
0,04	35478	1448	109,57%
0,05	78467	3934	80,80%
0,06	100746	6092	86,81%
0,07	84760	5945	54,54%
0,08	57053	4517	79,91%
0,09	35487	3236	66,61%
0,10	20145	1963	63,33%
0,11	9945	1058	73,89%
0,12	4344	516	60,87%
0,13	1818	221	51,53%
0,14	837	122	35,31%
0,15	391	60	117,87%
…	…	…	…
…	…	…	…

Tabell 4.6 Skattat $\hat{CV}$ per "GLM-premie"

Den, med hänsyn till antal individer, sammanvägda variationskoefficienten för samtliga premiegrupper är 78,12%. Variationskoefficienten utan indelning i premiegrupper skattades tidigare till 86,17% och har alltså reducerats efter indelningen.

4.3 Mått på kvarstående variation utan indelning i premiegrupper

Vi kan även kvantifiera den överspridning som kvarstår efter att hänsyn tagits till premieargumenten direkt baserat på varje individs GLM-premie. Vi går alltså inte omvägen via indelning i premiegrupper.

I avsnitt 4.1 betraktade vi hela datamaterialet som homogent och använde oss av excessvariansen som överspridningsmått och antog stokastiska skadeintensiteter Λ_i med samma väntevärde, $E(\Lambda_i)=\mu$. Vid Poissonregression å andra sidan förutsätts att varje individ har en fix (icke-stokastisk) skadeintensitet, $\mu_i = \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$, som

varierar som funktion av regressionsparametrarna β_j och designvariablerna, $x_{i1},x_{i2},\ldots,x_{ip}$, för respektive individ *i*. För att ge utrymme åt den individuella variation som inte förklaras av premieargumenten låter vi nu återigen skadeintensiteten Λ_i vara stokastisk, men med ett väntevärde, $E(\Lambda_i)=\mu_i$, som varierar mellan individer på samma sätt som i GLM-modellen (Poissonregressionen). Låt vidare $\hat{\mu}_i$ vara den skattade skadeintensiteten för individ *i* enligt GLM-modellen. Definiera så Pearsons χ^2-statistika som

$$\chi^2 = \sum_i \frac{\left(Y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}$$

Genom att dividera denna med antalet frihetsgrader får vi ett uttryck för *överspridningen,* ϕ, som i vårt fall skattas till

$$\hat{\phi} = \frac{1}{DF}\sum_i \frac{\left(Y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i} = \frac{Pearson\ chi\ square}{DF} = \frac{457891{,}21}{439216} = 1{,}0425 > 1$$, där

$DF = n - p$ anger antalet frihetsgrader, ref [4].

ϕ kvantifierar i vilken utsträckning vi har högre variation i data än vad som förväntas vid en Poissonregression där hela den individuella variationen förklaras av kovariaterna. Det ϕ vi får från GLM är baserat på den individdatastruktur vi har valt. Om man först aggregerar data till tariffcellnivå får man ett lägre $\hat{\phi}$.

Man använder i allmänhet ϕ för att beskriva hur mycket man behöver bredda sitt konfidensintervall, där båda sidor av intervallet multipliceras med ϕ. Att $\hat{\phi}>1$ innebär just att vi har en överspridning (för $\phi<1$ har vi istället en *underspridning*). Överspridningen skattas även automatiskt i proceduren Proc Genmod och redovisas då i tabellen ”*Goodness of fit*” i SAS outputfönster.

Notera att om hela datamaterialet istället antas vara homogent så beräknas ϕ för hela portföljen till:

$$\hat{\phi} = \frac{\hat{\sigma}^2}{\hat{\mu}} = \frac{0{,}0702660238}{0{,}0669386250} = 1{,}0497 > 1.$$

Alternativt uttryckt,

$$\hat{\phi} = \frac{1}{DF}\sum_i \frac{\left(Y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i} = 1{,}0497$$, där $\hat{\mu}_i = \hat{\mu}$ är lika med ”*enhetspremien*”, (0,0669).

För att kontrollera om vi har en konstant överspridningsfaktor, ϕ, kan vi behöva utföra en del beräkningar. Vi utgår från att överspridningen har formen $v_k = c\mu_k^a$, där v_k=Var(Λ_i) för alla individer i premiegrupp k=*1,2,…,23* och c och a är okända konstanter. Om vi ersätter v_k och μ_k med skattningar och logaritmerar får vi ett linjärt samband

$$\log\left(\hat{v}_k\right) = \log(c) + a \times \log\left(\hat{\mu}_k\right)$$, med intercept *log(c)* och lutningen *a*.

Om en viktad minsta kvadratskattning anpassas till punkterna ($\log(\hat{\mu}_k)$,$\log(\hat{v}_k)$) med vikter proportionella mot antalet individer i respektive premiegrupp k erhålls $\hat{c} = 0{,}0404$ och $\hat{a} = 0{,}9028$. Eftersom skattningen av a ligger nära 1 kan vi alltså anta sambandet $v_k = c\mu_k$. Eller uttryckt för individer (i) istället för premiegrupper (k): Var(Λ_i) $= c\mu_i$. Nedan plottas excessvarianserna per premiegrupp med minsta kvadratskattningen inlagd som rät linje.

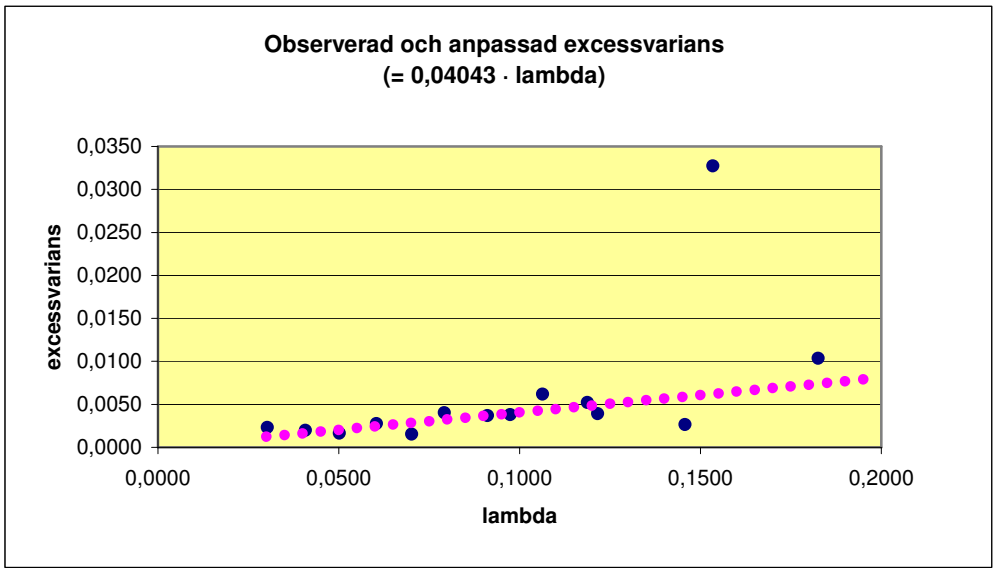

Diagram 4.2

Med hjälp av $\hat{\phi}$ kan alltså kvarstående varians skattas till

$$\left(\hat{\phi}-1\right)\times\hat{\mu}=\left(1{,}04252-1\right)\times 0{,}06694=0{,}0028462 \qquad (1)$$

Beräkning (1) ovan belyser det faktum att det $\hat{\phi}$ vi får ut genom Pearsons χ^2-statistika är direkt kopplat till den kvarstående variansen endast genom multiplikation med medelfrekvensen. På samma sätt är förklaringsgraden kopplad till $\hat{\phi}$. För att mera exakt kunna separera varianskomponenterna använder vi dock i T*abell 4.7* nedan en alternativ skattning av inomcellsvariansen.

	Varians	**Förklaringsgrad**	$\hat{\text{CV}}$
Inom-cell:	0,0029572	88,8%	0,8124
Mellan-cell:	0,0003735	11,2%	0,2887
Total:	0,0033306	100,0%	0,8621

Tabell 4.7

I *Tabell 4.7* har vi beräknat inom- och mellancellersvarianserna (*ic* respektive *mc*) enligt

$$\hat{\sigma}_{ic}^2=\sum_i \frac{\left(Y_i-\hat{\mu}_i\right)^2-\hat{\mu}_i}{n}=\sum_i \frac{\left(Y_i-\hat{\mu}_i\right)^2}{n}-\hat{\mu}$$

$$\hat{\sigma}_{mc}^2=\sum_i \frac{\left(\hat{\mu}_i-\hat{\mu}\right)^2}{n}$$

där vi i sista ledet utnyttjade att likelihoodekvationerna för GLM-skattningen medför att $\sum_i\left(Y_i-\hat{\mu}_i\right)=0$. Väsentligen har vi delat upp den skattade excessvariansen $\hat{v}$ från avsnitt 4.1 i två termer, där $\hat{\sigma}_{mc}^2$ svarar mot den del som förklaras av premieargumenten och $\hat{\sigma}_{ic}^2$ mot kvarstående excessvarians. Notera dock att $\hat{\sigma}_{ic}^2+\hat{\sigma}_{mc}^2=0{,}0033306$ skiljer sig något från värdet 0,0033274 på $\hat{v}$ i avsnitt 4.1. Skillnaden har att göra med att den GLM-skattade totalvariationen inte exakt kan delas upp i kvadratsummor (i motsats till linjära regressionsmodeller). Vidare kan vi

konstatera att den skattade inomcellsvariansen skiljer sig något från (1), främst beroende på att individer viktats olika beroende på premienivå.

Variationskoefficienterna i *Tabell 4.7* har beräknats genom

$$\hat{CV}_{ic} = \frac{\hat{\sigma}_{ic}}{\hat{\mu}}$$

$$\hat{CV}_{mc} = \frac{\hat{\sigma}_{mc}}{\hat{\mu}}$$

där variationskoefficienten för inom-cell, 81,24%, kan uttryckas som ”*premiefelet*” för en slumpmässigt vald individ i beståndet.

Förklaringsgraden för inom- respektive mellan-cell är variansen för respektive grupp dividerat på den totala variansen. Vi ser att vi endast förklarar 11,2% av den totala systematiska variationen mellan individer med hjälp av premieargumenten!

	Varians	**Förklaringsgrad**
Inom-cell:	0,0029572	4,21%
Mellan-cell:	0,0003735	0,53%
Brus:	0,0669393	95,26%
Total:	0,0702700	100,00%

Tabell 4.8

Tabell 4.8 ovan åskådliggörs att hela 95,26% av den totala variationen förklaras av ren slump (*brusvarians*) medan resterande 4,21+0,53=4,74% av variationen är individuell, och i sin tur kan delas upp i förklarad och icke-förklarad individuell variation enligt *Tabell 4.7*.

Notera att redovisad förklaringsgrad gäller för den tariff vi har beräknat med vår modell. Denna kan sakna något premieargument som kan tänkas vara aktuellt att ha med i en modell för att beräkna tariffen, exempelvis ett argument som besvarar om någon förare under 25 år kör bilen. Skattningar av ϕ och förklaringsgrad bör dock inte förändras betydligt tack vare upplysning om detta, även om man kan hoppas på en lite högre förklaringsgrad. Vi har heller inte inkluderat något bonussystem i analyserna eftersom ett sådant just eftersträvar att korrigera premien med hänsyn till individernas skadehistorik. Det skulle därför inte gå att kvantifiera kvarstående individuell variation, då just denna individuella variation i sådana fall tillåts påverka tariffindelningen.

Ett högre värde på ϕ innebär en lägre förklaringsgrad. Om nu ϕ med en förhållandevis enkel beräkning kan besvara frågan om tariffens kvalitet kan man undra varför det inte får ta större plats i litteraturen. Med insikt om den låga förklaringsgraden uppstår genast nya frågor. En viktig fråga i sammanhanget är hur försäkringsbolagets förklaringsgrad står sig gentemot konkurrerande bolag. Om man endast lyckas förklara 11,2% av individens beteende med tariffen, kan den ändå betraktas som effektiv om övriga försäkringsbolag endast lyckas förklara 10%? Oavsett svaret på den frågan så står klart att detta område lämnar rum för vidare efterforskning.

Kapitel 5

Diskussion

Med tanke på den betydande mängd information värdet på ϕ kan erbjuda är det underligt att dess värde inte har fått stå mer i centrum. Detta i synnerhet då detta värde är så lättillgängligt. Vid GLM-analys erbjuds man en skattning, $\hat{\phi}$, av ϕ, beräknad som funktion av Pearsons χ^2-statistika. Detta värde hamnar lätt i skuggan av övrig analys. Det finns litteratur sedan långt tidigare som berör området ”*excess varians*”, men ingen enligt min vetskap som riktigt vill fokusera på våra beräkningar och lyckas belysa resultaten på ett lättbegripligt sätt. Venezian (ref [6]) behandlar frågan, även om han fokuserar på betydelsen av körsträckans variation som förklaring av överspridning.

Brockman (ref [1]) är kanske den som tydligast behandlar vår problemställning. I sin artikel ”*Statistical motor rating: Making effective use of your data*” fäster han uppmärksamhet på den kvarstående variationen inom och mellan cell och artikeln innefattar en del funderingar och beräkningar som är relevanta för denna rapport. Han framför även hur ”*adverse selection*”, eller asymmetrisk information, kan leda till en obalanserad prissättning, vilken gör att försäkringsbolaget riskerar att förlora de kunder som drabbats av en orättvis tariff och bli kvar med kunder som kostar mer än tariffen avslöjar. Han påpekar också att en korrekt tariff är ett viktigt stöd även för underwriters, då pålitliga relationstal är en stor hjälp vid deras prissättning. Han menar att man inte ska underminera en underwriters magkänsla, men att det är en sak att *tro* att du vet en sak och en annan att *veta* att du vet.

Individer med en förhållandevis hög skadefrekvens (*de flesta kunder har inga skador alls under treårsperioden)* har en starkt negativ effekt på hela portföljen och ger upphov till en högre medelfrekvens, men hur bedömer man en enskild individs benägenhet att drabbas av en olycka? De premieargument som står till grund för tariffen har alla visat sig vara signifikanta vad gäller att bedöma risken för skada hos

en försäkring, men det är också visat att tariffen förklarar en relativt liten del av individens beteende. Mellan de individer som enligt GLM beräknas få samma skadefrekvens återstår fortfarande stora avvikelser som vi inte kan förklara. Att vi inte lyckats förklara den större delen av individens beteende är inte förvånande då man håller i åtanke att den individuella variationen är lika med all information som vi inte har registrerat. En fråga vad gäller tariffens kvalitet är om vi bör söka nya premiestyrande variabler och vilka dessa i sådana fall bör vara. Självklart finns det en mängd frågor som, om besvarade, hade varit relevanta för beräkning av tariffen, men som man av uppenbara skäl inte kan ställa till kunden.

Vi vill försöka förstå varför vissa individer har högre skadefrekvens än andra, men det finns omständigheter som kan ge missvisande utslag. En sådan situation uppstår exempelvis om en individ har en dubbelt så lång *faktisk* körsträcka som en annan individ som lämnat samma uppgifter. Då kommer den förra individen att ha dubbelt så många skador och enligt data tolkas som dubbelt så riskfylld som kund, även om båda kunder i själva verket har lika stor relativ skadefrekvens. Våra uppgifter om kundens körsträcka är helt och hållet baserat på vad kunden själv uppger i försäkringen och är inte grundat på faktiska siffror. En särskilt olycksdrabbad väg är en annan sak som kan bidra till fler skador, men som egentligen inte betyder att den olycksdrabbade är en sämre förare. Risken är lika stor för alla som kör på den. Ovannämnda exempel tydliggör att en högre skadefrekvens inte nödvändigtvis behöver bero på att en individ är mera ansvarslös eller vårdslös än andra individer.

Frank A. Haight menar i sin artikel ”*Accident Proneness: The history of an idea*” (ref [2]) att vissa kunder är mer benägna än andra att orsaka en skada på grund av att vissa individer faktiskt är ”klantigare” än andra. Han menar dock att dessa individer kompenserar för denna egenskap på ett sätt som gör dem svåra att identifiera, vilket i sin tur medför att det blir besvärligt att finna värdet på λ_i. Han nämner även i sin artikel en lustig teori som framfördes av Franz Alexander (en känd psykoanalytiker under mitten på 1900-talet), som menade att alla olyckor är medvetna val av individen, vilket enligt Alexander skulle innebära att vårt ”gäckande” λ_i är ett mått på en ”dödslängtan” mer än ett mått på oaktsamhet hos individen!

Haight tar i sin artikel upp att om vi kunde identifiera de individer som är orsaken till majoriteten av alla skador kunde man förhindra ett stort antal skador genom att dra in dessa individers rätt att köra bil. (*För vårt data kan beräknas att de 20% kunder som har högst skadefrekvens står för hela 47% av skadorna!*) Detta är förstås inte en fullständig lösning om man inte blint räknar med att dessa individer laglydigt accepterar beslutet och i framtiden verkligen avstår från att köra bil.

Huvudorsaken till att vi har svårt att skatta alla individuella λ_i är att så få olyckor sker under en given period på några år. Vår låga skadefrekvens kan förefalla förvånande om man jämför med värden som bygger på utländska data, men om man exempelvis lyssnar till kommentarer av Jean Lemaire (ref [5]) är detta ett normalt

värde för ett land i Norden. Enligt hans ”*Bonus-Malus in Automobile Insurance*” varierar skadefrekvensen kraftigt mellan olika länder, vilket känns helt logiskt om man jämför trafiksituationen exempelvis i Norden med den i länder kring Medelhavet.

Appendix A

Tabeller

I detta appendix finner vi tabellerna till vilka hänvisas i kapitel 4.1.

”Table 1” ur ref [6]

Antal skador	Exponering = antal individer/observationer	Antal skador totalt
0	41434	0
1	10162	10 162
2	2034	4 068
3	418	1 254
4	88	352
5	20	100
6	5	30
7	4	28
	54 165	**15 994**

Tabell A.1

Väntevärde:	*0,2953*
Varians:	*0,3622*
Excess:	*0,0670*
Variationskoefficienten i blandningsfördelningen:	*0,8764*

Tabell ”Belgium 1975-1976“ ur ref [7]

Antal skador	Exponering = antal individer/observationer	Antal skador totalt
0	96978	0
1	9240	9 240
2	704	1 408
3	43	129
4	9	36
5	0	0
6	0	0
7	0	0
	106 974	**10 813**

Tabell A.2

Väntevärde:	*0,1011*
Varians:	*0,1074*
Excess:	*0,0064*
Variationskoefficienten i blandningsfördelningen:	*0,7894*

Tabell "Great Britain 1968" ur ref [7]

Antal skador	Exponering = antal individer	antal skador totalt
0	370412	0
1	46545	46 545
2	3935	7 870
3	317	951
4	28	112
5	3	15
6	0	0
7	0	0
	421 240	**55 493**

Tabell A.3

Väntevärde:	*0,1317*
Varians:	*0,1385*
Excess:	*0,0068*
Variationskoefficienten i blandningsfördelningen:	*0,6252*

Tabell "Switzerland 1961" ur ref [7]

Antal skador	Exponering = antal individer	antal skador totalt
0	103704	0
1	14075	14 075
2	1766	3 532
3	255	765
4	45	180
5	6	30
6	2	12
7	0	0
	119 853	**18 594**

Tabell A.4

Väntevärde:	*0,1551*
Varians:	*0,1793*
Excess:	*0,0242*
Variationskoefficienten i blandningsfördelningen:	*1,0022*

Tabell "Germany 1960" ur ref [7]

Antal skador	Exponering = antal individer	antal skador totalt
0	20592	0
1	2651	2 651
2	297	594
3	41	123
4	7	28
5	0	0
6	1	6
7	0	0
	23 589	**3 402**

Tabell A.5

Väntevärde:	*0,1442*
Varians:	*0,1639*
Excess:	*0,0196*
Variationskoefficienten i blandningsfördelningen:	*0,9718*

Tabell "Zaire 1960" ur ref [7]

Antal skador	Exponering = antal individer	antal skador totalt
0	3719	0
1	232	232
2	38	76
3	7	21
4	3	12
5	1	5
6	0	0
7	0	0
	4 000	**346**

Tabell A.6

Väntevärde:	*0,0865*
Varians:	*0,1225*
Excess:	*0,0360*
Variationskoefficienten i blandningsfördelningen:	*2,1940*

Litteraturförteckning

[1] M. J. Brockman, T. S. Wright, *Statistical motor rating: Making effective use of your data*, J.I.A. 119, III 457-543.

[2] Frank A. Haight, *Accident proneness: The history of an idea*. Institute of Transportation Studies, University of California, Irvine, U.S.A., augusti 2001.

[3] Björn Johansson, *Matematiska modeller inom sakförsäkring*. Kompendium 2001

[4] Björn Johansson, Esbjörn Ohlsson, *Prissättning inom sakförsäkring med generaliserade linjära modeller*. Kompendium 2001.

[5] Jean Lemaire, *Bonus-Malus in Automobile Insurance*. Kluwer Academic Publishers, 1995.

[6] Emilio C. Venezian, *The distribution of automobile accidents – Are relativities stable over time?*

[7] Astin Bulletin, *no 2*. 1997.

[8] SAS Institute Inc., *SAS/STAT 9.1 User's guide, volume 3.* SAS Publishing 2004.