Stockholms universitet

Negativ binomialfördelning och kvasi-poisson som alternativ till Poisson-fördelning vid modellering av skade-frekvens. En fallstudie.

Andrea Klemming

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:2
Matematisk statistik
Juni 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2015:2**
http://www.math.su.se/matstat

Negativ binomialfördelning och kvasipoisson som alternativ till Poissonfördelning vid modellering av skadefrekvens. En fallstudie.

Andrea Klemming*

Juni 2015

Sammanfattning

I denna uppsats har vi undersökt en viktig komponent vid bestämmande av riskpremien på sakförsäkringar, *skadefrekvensen*. Praxis inom försäkringsbranschen är att använda poissonfördelningen för att modellera skadefrekvensen med hjälp av generaliserade linjära modeller. En invändning mot poissonfördelningen är att dess varians och väntevärde ska vara lika, vilket ofta inte stämmer i verkliga räknedata. I premiesättning delar man upp försäkringskontrakten i klasser, och kvarvarande variation inom klasserna leder till att det är vanligt att variansen är större än väntevärdet, vilket kallas överspridning. Två olika metoder att arbeta med överspridning är att övergå till att anta negativ binomialfördelning eller såkallad kvasipoisson. Vi har, i en fallstudie med motorcykelförsäkringar, undersökt dessa två antaganden som alternativ till poissonfördelningen. Datamaterialet är tillhandahållet av det tidigare försäkringsbolaget *Wasa*. I vår analys har vi fått bekräftat med hjälp av överspridningstest att det är statistiskt säkerställt att det finns överspridning, och därmed är det motiverat att undersöka alternativ till poissonfördelningen. Kvasipoisson gav högre standardfel till samtliga parametrar än poissonfördelningen, vilket dock inte gav någon skillnad i klassuppdelningen; då kvasipoisson ger samma parameterskattningar som poissonfördelningen blev den överflödig i vidare analys. Genom ett likelihoodkvottest har vi med Monte Carloapproximation kunnat förkasta poissonmodellen till förmån för negativ binomialmodellen. Vi kan konstatera att vid lägre skadefrekvenser fick vi ingen större skillnad mellan skattningar för poissonmodellen och negativ binomialmodellen, medan man vid högre skadefrekvens fick högre skattningar för negativa binomialfördelningen. Det blir en svår avvägning mellan fördelningarna. Då vi använder poissonfördelningen får vi en modell oberoende av skala (modellen påverkas ej av om man mäter i procent eller promille), och ett aggregerat datamaterial ger en tillräcklig statistika. Det är två önskvärda egenskaper som negativa binomialfördelningen inte besitter. Vi behöver då avgöra om den mer korrekta modellen är värd att offra de två önskvärda egenskaperna för, eller om de borde prioriteras. Ur ett rent statistiskt perspektiv skulle jag välja negativa binomialfördelningen, men med mer branschinformation är det sannolikt att poissonfördelningen vore att föredra.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: andrea.klemming@gmail.com. Handledare: Martin Sköld och Tom Britton.

Abstract

In this thesis we have explored an important component when computing the pure premium in non-life insurance pricing, *the claim frequency*. Today, within the insurance businesses, this is modeled with generalized linear models and the distribution is assumed to be Poisson. However, in the Poisson distribution the expected value and the variance are the same, which is often incorrect in real count data. When deciding the pure premium the insurance takers are divided into groups, and variance within these groups is one reason to why it is common that the variance is larger than the mean in insurance data. This is called overdispersion. We have looked at two alternatives, which deals with overdispersion, and these alternatives are the negative binomial distribution and quasipoisson. We compared these assumptions through a case study with insurance data over motorcycles, received from the former insurance company, *Wasa*.

In our analysis we can confirm overdispersion statistically and thereby motivate trying alternatives to the Poisson distribution. Quasipoisson estimated larger standard errors to all parameters than the Poisson distribution. However, this made no difference in dividing insurance takers into groups, and since quasipoisson gives the same parameter estimates as the Poisson distribution we decided it to be redundant in further analysis.

Through a likelihoodratio test and Monte Carlo approximation of the p-value we could reject the null hypothesis of the Poisson distribution in favor of the alternative, the negative binomial distribution. In the lower estimates of the claim frequency there were no big differences between the two distributions. However, when getting to higher claim frequencies the negative binomial distribution gave higher estimates than the Poisson distribution. Even though statistic tests favor the negative binomial distribution it is a difficult choice. The Poisson distribution has other benefits, such as being scale invariant (it does not matter if we measure in percent or per mille), and aggregated data is a sufficient statistic. Both these things are properties not held by the negative binomial distribution. It is a difficult assessment if the improvement in modeling the data shown by the negative binomial distribution is worth the sacrifice of losing the two desired properties in the Poisson distribution. However, from a purely statistical point of view the negative binomial model is recommended.

Förord

Detta är ett examensarbete på 15 högskolepoäng som leder till en kandidatexamen i matematisk statistik vid Stockholms universitet. Arbetets gång underlättades mycket tack vare mina fantastiska studentkolleger, vilka bidrog till många givande diskussioner och stort stöd. Jag vill även rikta ett stort tack till mina handledare, Martin Sköld och Tom Britton. Utöver stödet i att hitta ett ämne som intresserar mig fanns de även där för råd under hela arbetets gång.

Innehållsförteckning

1 Inledning

Syftet med försäkringar är att skapa ett fungerande system för att dela på risk. Detta fungerar eftersom den beräknade förlusten för ett stort försäkringsbolag är, enligt stora talens lag, mycket mer förutsägbar än risken för en enskild person (Johansson & Ohlsson, 2010). Olika former av försäkringar har funnits länge. Det fanns, enligt grekiske filosofen Aristoteles, ett försäkringsbolag redan runt år 300 före Kristus. Detta försäkringsbolag behandlade slavar, men hur är okänt. Sveriges tidiga försäkringsformer nämns senare i historien. Omkring år 1200 nämns *brandstoden* i Skåne (då tillhörande Danmark). Denna spreds senare över till Sverige. Brandstoden var ett slags försäkring mot brand och innebar att alla hemmansägare inom ett visst område skulle ge ersättning om en av dem råkade ut för skador på grund av brand - man delade på risken (Försäkringens historia).

Idag har systemet utvecklats mycket och försäkringar är ett självklart inslag i våra liv. På marknaden finns många olika försäkringar för många olika saker, och flertalet konkurrerande försäkringsbolag. Konkurrensen på marknaden leder till att det är viktigt för försäkringsföretagen att kunden får betala en premie som är proportionell i förhållande till den ekonomiska risk hen representerar. Felaktiga premier leder automatiskt till att kunder med för höga premier övergår till andra företag och kunder med för låga premier söker sig till företaget (Johansson & Ohlsson, 2010).

För att uppnå målet med premier som motsvarar risken använder försäkringsbolagen till stor del egna historiska data på försäkringskontrakt, ibland med kompletteringar från externa källor, och bearbetar informationen med statistiska verktyg. Sedan *generaliserade linjära modeller* introducerades som verktyg vid premieanalys på 90-talet används dessa som standard i många länder, och det är även dessa vi kommer att använda. Vi kommer att undersöka en komponent i premiesättningen kallad *skadefrekvens*, där man undersöker hur ofta skador sker i olika premieklasser. Praxis idag är att modellera detta med hjälp av poissonfördelningen (Johansson & Ohlsson, 2010). Precis som modeller i allmänhet är poissonantagandet dock en förenkling av verkligheten, och det finns andra fördelningar som skulle kunna åtgärda vissa av problemen poissonantagandet stöter på. Vi kommer med hjälp av olika statistiska verktyg undersöka två av dessa, negativa binomialfördelningen och kvasipoisson. Detta kommer att göras genom en fallstudie med hjälp av programmet R (R Core Team, 2015) och ett datamaterial över försäkringskontrakt för motorcyklar, tillhandahållet av tidigare försäkringsbolaget Wasa.

Vi börjar med att i Avsnitt 2 gå igenom bakgrundsteori för sakförsäkringar. Där kommer vi även gå vidare med mer specifik statistisk teori relevant för vår modellering av skadefrekvensen och jämförande av modeller. I Avsnitt 3 fortsätter vi genom att introducera vårt datamaterial och tillämpa teorin för att jämföra våra modeller. I slutet av Avsnitt 3 gör vi en kort sammanfattning av de viktigaste resultaten för att sedan diskutera vad de innebär och vilka slutsatser som kan dras i Avsnitt 4, diskussionen.

2 Teori

I teoridelen börjar vi med att allmänt gå in på försäkringar för att sedan fokusera på det som behövs för vår modellering. Vi kommer, innan teori beskrivs, att hänvisa till dess källa då det är större stycken från samma källa. När det är korta avsnitt från en annan källa kommer den att hänvisas till i texten, oftast i slutet av den delen hänvisningen gäller. Allt är skrivet med de olika källorna som nämns som grund, men är till viss del anpassad för att bättre passa till den modellering vi kommer göra senare.

2.1 Försäkringar

Tills vidare kommer informationen från Johansson & Ohlsson (2010) då inte annat anges.

Premiesättningen på försäkringar baseras på den förväntade kostnaden. I premiesättningen brukar man skilja mellan hur man sätter den totala premienivån, vilket bland annat baseras på förväntade administrativa kostnader, och hur man ska fördela premierna mellan försäkringstagare. Intresset i denna uppsats ligger i hur premier fördelas i de olika premieklasserna, vilket innebär att det relevanta är hur totala premienivån fördelas mellan försäkringstagare.

Vid fördelningen tar man hänsyn till att olika försäkringstagare bidrar med olika stor risk för skador, och när en skada uppstår bidrar de till olika förväntade kostnader. Det är inom detta område statistiska metoder blir aktuella. För att differentiera mellan olika försäkringstagare använder man sig av olika faktorer, vilka fungerar som variabler vid modellerande. Faktorerna kan kategoriseras enligt följande:

- Försäkringstagarens egenskaper.
- Det försäkrade objektets egenskaper.
- Egenskaper hos den geografiska zonen där försäkringstagaren bor.

Om man har tillräckligt med data skulle man kunna sätta premier efter de observerade premierna. Detta skulle dock ofta leda till väldigt varierande premier, både tidsmässigt och mellan klasser, då det inom vissa klasskombinationer ibland helt saknas skador. För att hålla premierna stabilare över tid och mer jämnt fördelade mellan premieklasser använder man statistiska metoder för att beräkna ett förväntat värde för *riskpremien*. Den definieras som skadekostnaden dividerat med durationen (längden på ett försäkringskontrakt), vilket innebär att riskpremien är medelkostnaden per premieår. Riskpremien kan delas upp i två faktorer, *skadefrekvens* och *medelskada*. Det innebär antalet skadeanmälningar dividerat med försäkringens duration, respektive den totala skadekostnaden dividerat med antalet skador. Vi får

$$\text{Riskpremie} = \text{Skadefrekvens} \cdot \text{Medelskada} = \frac{\text{Antal skador}}{\text{Duration}} \cdot \frac{\text{Skadekostnad}}{\text{Antal skador}}.$$

Denna uppdelning är vanlig vid modellering av riskpremien. Man gör då två statistiska analyser, en för skadefrekvensen och en för medelskadan. Riskpremien,

skadefrekvensen och medelskadan kallas för nyckeltal. Nyckeltalen är en kvot, där täljaren är en respons och nämnaren responsens exponering. Antal skador behandlas därför som en stokastisk variabel vid modellering av skadefrekvensen, där durationen är antalet skadors exponering. Vid modellering av medelskadan är antal skador istället skadekostnadens exponering och behandlas därmed ej som slumpmässig. Vi kommer fokusera på den del av premiesättningen som ges av skadefrekvensen.

2.2 Skadefrekvens

Vi kommer hädanefter arbeta med skadefrekvensen som responsvariabel. Modellbyggandet bygger då på tre grundläggande antaganden.

Antagande 1 Alla försäkringskontrakt är oberoende för våra responsvariabler.
Antagande 2 Våra responsvariabler är oberoende av tiden.
Antagande 3 Två försäkringskontrakt i samma tariffcell (försäkringskontrakten tillhör samma klass för varje faktor) med samma duration är likafördelade.

Man kan argumentera för att dessa antaganden inte alltid är uppfyllda, men även argument till varför vi fortfarande kan anta dessa inom vissa försäkringar finns. För mer om detta se sidorna 7-8 i Johansson & Ohlsson (2010).

De olika faktorerna vi har tillgång till kommer vi gruppera för att skapa premieklasser. Genom antagande 3 och genom att justera för försäkringskontraktens duration (förklaras närmre i Avsnitt 2.6.3 Offsetvariabel) får försäkringskontrakt tillhörande samma tariffcell samma intensitet. De tillsätts därmed samma riskpremie.

Vi vill avgöra hur skadefrekvensen beror av ett antal kovariater. Vi har räknedata som responsvariabel och man brukar inom premiesättning använda sig av multiplikativa modeller (förklaras närmre i Avsnitt 2.6.1 Multiplikativa modeller). För att modellera detta övergår man därför från vanlig multipel linjär regression, där responsvariabeln antas normalfördelad och modellen additiv, till att använda sig av generaliserade linjära modeller, GLM:er. Linjär regression är ett specialfall av GLM:er.

2.3 Generaliserade linjära modeller

Informationen övergår nu till att vara från Agresti (2013) då inte annat anges. GLM:er inkluderar linjära modeller för transponerade medelvärden där responsvariabeln tillhör exponentialfamiljen. GLM:er är uppbyggda av tre komponenter: en slumpmässig, en systematisk och en länkfunktion.

2.3.1 Slumpmässig komponent

Den slumpmässiga komponenten i GLM:er består av en responsvariabel med oberoende observationer och som tillhör den naturliga exponentialfamiljen. Det innebär att dess täthetsfunktion kan skrivas på formen

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp(y_iQ(\theta_i)),$$

där $Q(\theta_i)$ är den naturliga parametern och y_i står för responsvariabeln, i vårt fall skadefrekvensen. En mer allmän form av exponentialfamilj är en exponentiell dispersionsfamilj. Till den hör fördelningar med täthetsfunktion som kan skrivas på formen

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right), \tag{1}$$

där θ_i är den naturliga parametern. Det tillåts även en spridningsparameter, ϕ. När spridningsparametern betraktas som en konstant kan den exponentiella dispersionsfamiljen förenklas till den naturliga exponentialfamiljen.

2.3.2 Systematisk komponent

Den systematiska komponenten illustrerar en linjär koppling mellan den slumpmässiga komponenten och kovariaterna. Om N är antalet observationer, $p+1$ antalet parametrar (p är antalet förklarande variabler) i modellen och vi har

$$\eta_i = \sum_{j=0}^{p} \beta_j x_{ij}, \quad i = 1, ..., N,$$

så är η_i den systematiska komponenten. I matrisform får vi

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

där:

$\boldsymbol{\eta}$ är en N-dimensionell kolonnvektor,
$\boldsymbol{\beta}$ står för modellparametrarna och är en $p+1$-dimensionell kolonnvektor, och
$\boldsymbol{X}$ är en $N \times (p+1)$-matris.

När modellen har ett intercept består första kolonnen i $\boldsymbol{X}$-matrisen av N stycken ettor.

2.3.3 Länkfunktion

Den sista komponenten i GLM:er, länkfunktionen, fungerar som en länk mellan slumpmässiga och systematiska komponenten. Den består av en funktion som omvandlar väntevärdet för den slumpmässiga komponenten till den systematiska komponenten. Länkfunktionen är monoton och differentierbar. Den ger sambandet mellan väntevärdet och förklaringsvariablerna genom

$$g(\mu_i) = \sum_{j} \beta_j x_{ij}, \quad i = 1, ..., N. \tag{2}$$

På matrisform får vi

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

där $\boldsymbol{\mu}$ är en N-dimensionell kolonnvektor. Den länkfunktion som transformerar väntevärdet μ_i till den naturliga parametern $Q(\theta_i)$ kallas för kanonisk länkfunktion. Vid linjär regression används länkfunktionen $g(\mu_i) = \mu_i$, vilken även kallas identitetslänken (Johansson & Ohlsson, 2010).

Istället för att skriva vår modell i termer av β, som vi beskrivit GLM:er i, kommer vi till större delen att övergå till beteckningar med λ. Detta gör vi för att vi har variabler uppdelade i klasser och vi får en skattning för varje klass, vilket innebär att exempelvis för 5 olika åldersgrupper får vi 5 olika β-skattningar fast den underliggande faktorn är densamma. Vi kommer fortsättningsvis istället skriva λ_i^X, där X är ett index som står för vilken faktor det är, och i ett index för vilken klass i den faktorn. Om vi exempelvis har en faktor med tre olika klasser, varav en basklass (som då blir interceptet), får vi då vi skriver i termer av β att

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

Där antas värdet β_0 i basklassen och för varje extra klass i faktorn lägger man till ytterligare ett β och ett X_j. Våra X blir indikatorvariabler. Med vår nya beteckning får vi

$$\eta_i = \lambda + \lambda_i^X,$$

och vid tillägg av klasser är det helt enkelt ett värde till i kan anta.

2.4 Maximumlikelihoodskattningar

Lutningskoefficienterna i en GLM skattas genom maximumlikelihoodmetoden. Man får maximumlikelihoodskattningar genom att maximera likelihoodfunktionen eller loglikelihoodfunktionen (en monoton transformation) med avseende på β. Om vi har en stokastisk variabel Y med fördelningen $F(y;\beta)$ definieras likelihoodfunktionen som

$$L(\beta) = \begin{cases} \prod_{i=1}^n p(y_i;\beta), & \text{om } Y \text{ är diskret} \\ \prod_{i=1}^n f(y_i;\beta), & \text{om } Y \text{ är kontinuerlig,} \end{cases}$$

där y_1, y_2,..., y_n är ett slumpässigt stickprov från Y (Alm & Britton, 2008).

Enligt vår definition av GLM:er tillhör alla fördelningar man kan modellera med denna metod exponentialfamiljen. Man kan därmed härleda maximumlikelihoodskattningarna för en allmän GLM genom att börja i definitionen för en exponentiell dispersionsfamilj, se ekvation (1) för definitionen. Efter en del arbete fås likelihoodekvationerna

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \ldots. \tag{3}$$

Vi får β från sambandet i ekvation (2), vilken kan skrivas om till $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$ [Agresti, 2013].

Likelihoodekvationerna är ofta inte möjliga att lösa explicit och man kan då istället övergå till lösning med hjälp av iterationer. Vi kommer använda glm funktionen i programvaran R, vilket använder sig av iterationer genom metoden *Fisher scoring* för en implicit lösning. När man använder sig av Fisher scoring får man även ut kovariansmatrisen som en biprodukt (Venables & Ripley, 2002).

2.5 Önskvärda egenskaper

När man arbetar med skadefrekvensen som responsvariabel har man vissa egenskaper man vill att modellen ska ha. Bland annat vill man att det aggregerade datamaterialet ska vara en tillräcklig statistika, och att modellen ska vara oberoende av skala. Vi beskriver detta närmre.

2.5.1 Aggregerad data tillräcklig statistika

Vid riskpremieanalys har man i många fall tillgång till väldigt stora mängder data. Ett alternativ för att förenkla hantering av data är att aggregera den, vilket kan ge bland annat kortare beräkningstider. Samtidigt är det viktigt att man inte förlorar information vid aggregeringen. Detta kan kontrolleras med hjälp av tillräckliga statistikor. Informationen angående tillräckliga statistikor och faktoriseringskriteriet kommer från Olive (2014).

Om vi har ett slumpmässigt stickprov från en stokastisk variabel med okända parametrar $\boldsymbol{\beta}$, så är en tillräcklig statistika för $\boldsymbol{\beta}$ en statistika som innehåller all möjlig information om de okända parametrarna. För oss behöver en tillräcklig statistika innehålla all möjlig information angående väntevärdet för skadefrekvensen. En tillräcklig statistika skulle kunna bestå av hela datamaterialet, men det finns även statistikor som använder mindre delar av materialet och fortfarande är tillräckliga. Om vi aggregerar datamaterialet får vi en statistika som använder mindre delar av materialet och därmed gör modellerandet mindre beräkningstungt och om statistikan är tillräcklig förlorar vi ingen information.

Om $(Y_1, ..., Y_n)$ har en simultan täthetsfunktion som beror på en parametervektor $\boldsymbol{\beta} \in \boldsymbol{B}$, där $\boldsymbol{B}$ är parameterrummet så är en statistika $T(Y_1, ..., Y_n)$ tillräcklig med avseende på $\boldsymbol{\beta}$ om fördelningen för $(Y_1, ..., Y_n)$ betingat på $\boldsymbol{T} = \boldsymbol{t}$ är oberoende av $\boldsymbol{\beta}$ för alla värden på t. Man kan hitta tillräckliga statistikor genom att använda sig av *faktoriseringskriteriet*.

Faktoriseringskriteriet
Om vi har en stokastisk variabel med likelihoodfunktionen $f(y_i; \boldsymbol{\beta})$ där $\boldsymbol{\beta} \in B$ med statistikan $T(\boldsymbol{Y})$ så är $T(\boldsymbol{Y})$ en tillräcklig statistika för $\boldsymbol{\beta}$ omm, för alla y och alla β i parameterrummet B,

$$f(\boldsymbol{y}; \boldsymbol{\beta}) = g(T(\boldsymbol{y}); \boldsymbol{\beta})h(\boldsymbol{y}).$$

2.5.2 Skalinvariant modell

Informationen är än en gång från Johansson & Ohlsson (2010) då inte annat anges. När man arbetar med skadefrekvensen vill man arbeta med en modell som är skalinvariant, alltså oberoende av skala. Detta för att man inte vill att inferensen ska bero på om man mäter skadefrekvensen i procent eller promille. Detta innebär rent tekniskt att om vi har en positiv konstant c och en stokastisk variabel Y från en specifik fördelningsfamilj, så är Y oberoende av skala om cY tillhör samma fördelningsfamilj som Y. Detta skulle exempelvis vara om Y är poissonfördelad och även cY är det, även om cY har en annan intensitet. Man

kan visa att av alla fördelningar som tillhör exponentiella dispersionsmodeller i GLM:er så är det enbart de med variansfunktionen $v(\mu) = \mu^p$ som är oberoende av skala (μ är väntevärdet). Dessa kallas Tweedie modeller.

2.6 Förberedande modellbyggande

Oavsett fördelningsantagande så finns det några komponenter alla våra modeller har gemensamt. Vi börjar därför gå igenom den information modellerna har gemensamt för att sedan bli mer fördelningsspecifika.

2.6.1 Multiplikativa modeller

Tidigare nämndes att man vid modellering av premier ofta föredrar multiplikativa modeller. Detta beror på att när man använder multiplikativa modeller så förändras premien med en faktor om man byter klass i en variabel och därmed anpassas förändringen till hur hög premien är. Detta bedöms mer realistiskt än i en additiv modell där premien förändras med samma term oavsett hur hög premien är i övrigt. Om vi har två parametrar γ_{ij}, där i står för vilken faktor det är och j för vilken klass det är i faktorn, kan en multiplikativ modell skrivas som

$$\mu_{kl} = \gamma_0 \gamma_{1k} \gamma_{2l}.$$

Vi får att k är ett index för vilken klass inom faktor 1 det är och l motsvarande för faktor 2. Eftersom vi arbetar med faktorer uppdelade i klasser behöver vi ha en basklass för varje faktor. Alla skattningar för övriga klasser sker då i förhållande till basklassen. När alla faktorer befinner sig i sin basklass antas värdet γ_0.

2.6.2 Loglänk

Vid modellering av skadefrekvensen kommer vi att använda oss av den naturliga logaritmen som länkfunktion, oavsett fördelningsantagande. Vi får då modellen

$$\log(\mu_i) = \sum_j \beta_j x_{ij}.$$

Vi kan se att detta ger en multiplikativ modell då vi för medelvärdet får sambandet

$$\mu(x_i) = \exp\left(\sum_j \beta_j x_{ij}\right).$$

Om vi skriver detta med λ-beteckningarna kan vi förtydliga sambandet genom att anta att vi har två kovariater, vilket med samma beteckningar som i Avsnitt 2.6.1 Multiplikativa modeller ger

$$\mu_{kl} = e^{\lambda_0} e^{\lambda_k^{x_1}} e^{\lambda_l^{x_2}} = \gamma_0 \gamma_{1k} \gamma_{2l}.$$

Varje klassförändring i x_j ger då den multiplikativa förändringen $e^{\lambda_i^{x_j}}$ av väntevärdet för skadefrekvensen, där i står för vilken klass det är.

2.6.3 Offsetvariabel

Ibland har man viss information om väntevärdet innan man modellerat data. Ett exempel på detta skulle kunna vara att de olika observationerna sker under specifika tidsperioder och man förväntar sig ett proportionellt samband. Man kan lösa detta genom användandet av en såkallad offsetvariabel. Vi kommer att använda durationen som offsetvariabel i vår modellerering och då man använder sig av naturliga logaritmen som länkfunktion blir koefficienten framför offsetvariabeln 1. Detta ger oss den systematiska komponenten

$$\eta_i = \log(\text{duration}_i) + \sum_j \beta_j x_{ij},$$

vilket är ekvivalent med uttrycket

$$\log\left(\frac{\mu_i}{\text{duration}_i}\right) = \sum_j \beta_j x_{ij}.$$

2.6.4 Konfidensintervall

När man använder sig av ml-skattningar, vilket vi gör, kan man basera konfidensintervall för β-skattningarna på att de är asymptotiskt normalfördelade och väntevärdesriktiga. Skattningarnas kovariansmatris överensstämmer med fisherinformationen. För varje parameter blir ett approximativt 95% konfidensintervall $\hat{\beta}_j \pm 1.96\sqrt{c_{jj}}$ där c är kovariansmatrisen.

I vår modell med naturliga logaritmen som länkfunktion kan det vara mer relevant att beräkna konfidensintervall för $e^{\beta_j} = \gamma_j$ än för β_j. Detta gör vi genom att först beräkna konfidensintervallet för β_j, vilket vi kallar (a, b). Vi använder sedan detta för att beräkna konfidensintervallet (e^a, e^b) för $e^{\beta_j} = \gamma_j$. Konfidensintervallet är symmetriskt för β_j, men ej för e^{β_j}.

2.7 Fördelningsantagande

Vi börjar med att motivera varför man inom försäkringsbranschen väljer att modellera antal skador som poissonfördelade (skadefrekvensen som relativ poisson). Vi utvecklar sedan från grundantagandet och motiverar alternativen kvasipoisson och negativ binomial.

2.7.1 Poisson

Motiveringen för antagandet att antal skador är poissonfördelat börjar med fördelningen i varje enskild tariffcell.

Vi antar att antalet skador för varje försäkringskontrakt $N(t)$ är en stokastisk process för tiden (0,t], där N(0)=0. Under modellantaganden liknande cellhomogenitet och att responsvariablerna är oberoende av tiden, och villkoret att skadeanmälningarna ej ankommer i grupper är processen $\{N(t); t \geq 0\}$ en poissonprocess. Detta motiverar antagandet att antalet skador för varje försäkringskontrakt är poissonfördelat. Enligt tidigare antagande om oberoende

mellan försäkringskontrakt får vi även att den aggregerade datan med alla försäkringskontrakt i en cell poissonfördelad.

Poissonfördelningen tillhör exponentialfamiljen och kan skrivas

$$p(y;\theta) = \frac{1}{y!}e^{-\theta}e^{y\log\theta},$$

på exponentialfamiljform. Vi ser att den naturliga parametern i poissonfördelningen är $Q(\mu) = \log\mu$, vilket innebär att den kanoniska länkfunktionen är $g(\mu) = \log(\mu)$. Vi kommer alltså att använda den kanoniska länkfunktionen när vi modellerar skadefrekvensen med poissonfördelningen. Vi kan även se att $T(\boldsymbol{y}) = \sum_i y_i$, vilket innebär att aggregerad data är en tillräcklig statistika för poissonfördelningen enligt faktoriseringskriteriet. Om vi har en poissonfördelad stokastisk variabel Y med parameter λ får vi $E[Y] = Var(Y) = \lambda$ (Alm & Britton, 2008). Vi kan därmed konstatera att poissonfördelning är oberoende av skala enligt definitionen för Tweedie models (se Avsnitt 2.5.2 Skalinvariant modell). Ett problem med att väntevärdet ska vara lika med variansen i poissonfördelningen är att variansen ofta är större än medelvärdet i verkliga räknedata. Detta kallas överspridning. En konsekvens av överspridning är felaktiga standardardfel (Agresti, 2013).

Antagandet att skadefrekvensen är poissonfördelad är inte självklart. Även om man kan anta att ett försäkringskontrakt har en poissonfördelad skadefrekvens, blir antagandet mindre tillförlitligt när man lägger ihop flera försäkringskontrakt till en klass (vilket behövs för att kunna skatta en intensitet). När man grupperar föräkringskontrakten till klasser finns även variation i skadefrekvensen mellan kontrakten i samma tariffcell. Denna variation är en bidragande orsak till möjlig överspridning när man modellerar skadefrekvensen med poissonfördelningen.

Det finns olika sätt att testa för överspridning. Vi kommer att arbeta med ett test där man först utför en regression för att sedan testa för överspridning. Vi kommer ge en översikt över metoden och hänvisar den intresserade läsaren till vår källa Cameron & Trivedi (1990) för en mer detaljerad beskrivning.

I detta överspridningstest behöver vi inte specificera vilken fördelning vi testar mot poissonfördelningen, utan vi ställer istället upp en mer allmän variansfunktion. Vi har en responsvariabel Y_i och förklarande variabler $\boldsymbol{X}_i$. Observationerna är oberoende. Medelvärdet för responsvariabeln är

$$E[Y_i] = \mu_i = \mu(X_i, \beta)$$

och vi ställer upp variansfunktionen

$$Var[Y_i] = \mu_i + \alpha \cdot g(\mu_i),$$

där g är en specificerad funktion som avbildar från den positiva reella tallinjen till den positiva reella tallinjen. Om variabeln är poissonfördelad så är $\alpha = 0$, vilket är vår nollhypotes. När vi vill veta om vi har överspridning testar vi mothypotesen $\alpha > 0$ ($\alpha < 0$ innebär underspridning). Vi får

H_0: $\text{Var}(Y_i)=\mu_i$
H_1: $\text{Var}(Y_i)=\mu_i + \alpha \cdot g(\mu_i)$,

där $\alpha > 0$.

Enligt definition är $Var(Y_i) = E[(Y_i - E[Y_i])^2]$, vilket under H_1 ger

$$Var(Y_i) = E[(Y_i - \mu_i)^2] = \mu_i + \alpha \cdot g(\mu_i) = E[Y_i] + \alpha \cdot g(\mu_i).$$

Vi vill ha ett uttryck för $\alpha \cdot g(\mu_i)$ och skriver därför om uttrycket till

$$E[(Y_i - \mu_i)^2 - Y_i] = \alpha \cdot g(\mu_i).$$

Med hjälp av detta uttryck kan man utföra en regression. Denna kan man använda för att skatta α och testa $\alpha = 0$ med en t-statistika, vilken är asymptotiskt normalfördelad under nollhypotesen.

Det finns ett flertal olika metoder för att hantera överspridning i räknedata, varav vi kommer arbeta med två. I ena metoden kommer vi övergå till såkallad kvasipoisson och i andra fallet anta negativ binomialfördelning. I överspridningstestet kommer vi utföra testet med $g(\mu_i) = \mu_i$, vilket gör att mothypotesen motsvarar variansfunktionen för negativa binomialfördelningen.

2.7.2 Kvasilikelihood

Information angående kvasilikelihood och negativ binomialfördelning kommer från Agresti (2013). I tidigare modeller används maximumlikelihoodmetoden för att skatta parametrarna. Skattningarna baseras då på den fördelning man antar för responsvariabeln. Detta kräver att man för responsvariabeln antar en fördelning man kanske inte är säker på och att fördelningen tillhör exponentialfamiljen. När man beräknar ml-skattningarna i en GLM använder man sig dock inte av hela fördelningen, utan enbart av förhållandet mellan väntevärde och varians (Venables & Ripley, 2002). Med detta som bakgrund väljer man i kvasilikelihood inte en specifik fördelning för Y_i, utan bestämmer istället förhållandet mellan väntevärde och varians. Man kan övergå till kvasilikelihoodskattningar och därmed undvika behovet av fördelningsantagande och tillåta en större andel fördelningar till GLM. Man får ekvationer för parametrarna som i likelihoodekvationerna (3) med förändringen $Var(Y_i) = v(\mu_i)$, där $v(\mu_i)$ är sambandet mellan varians och väntevärde. Man får då skattningsekvationerna

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{v(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, ...$$

En variant av kvasilikelihoodskattningar är kvasipoisson. Kvasipoisson kan användas för att hantera överspridning i data. Man antar då att sambandet mellan väntevärde och varians är $v(\mu_i) = \phi\mu_i$, där ϕ (vid överspridning) är någon konstant större än 1. Man tillåter alltså en spridningsparameter. I kvasipoisson får vi samma parameterskattningar som vanlig Poisson, men standardavvikelserna förändras med faktorn

$$\sqrt{\frac{\sum_i (y_i - \hat{\mu}_i)^2 / v^*(\hat{\mu}_i)}{N - p}},$$

där $N - p$ är antalet observationer subtraherat med antalet parametrar.

2.7.3 Negativ binomial

Ett annat sätt att hantera överspridning i poissonfördelningen är att övergå till en *konjugerad blandningsmodell* (conjugate mixture model), vilket innebär en modell där responsvariabeln har en specifik fördelning betingat på en parameter. Parametern har även den en specifik fördelning. I vårt fall är responsvariabeln poissonfördelad betingat på intensiteten som är en stokastisk variabel. Ett lämpligt alternativ är att se intensiteten som en gammafördelad variabel. Detta eftersom en poissonfördelad stokastisk variabel med gammafördelad intensitet är negativt binomialfördelad. I vårt fall innebär det att försäkringskontrakten inom en klass tillåts variera med en parameter $\Lambda \sim \text{Gamma}(\alpha, \beta)$. Vi använder en annan parametrisering än Agresti (2013), och i dess parametrisering skulle det bli $\text{Gamma}(\alpha, \mu)$, där $\mu = \alpha\beta$. Vi får $y|\Lambda \sim \text{Poisson}(\Lambda)$, $\text{E}[\Lambda] = \alpha\beta$ och $\text{Var}(\Lambda) = \alpha\beta^2$. Detta ger

$$E[Y] = E[E[Y|\Lambda]] = E[\Lambda] = \alpha\beta$$

och

$$Var(Y) = E[Var(Y|\Lambda)] + Var(E[Y|\Lambda]) = E[\Lambda] + Var(\Lambda) = \alpha\beta + \alpha\beta^2.$$

En naturlig omskrivning ges av $\mu = \alpha\beta$ och $r = 1/\alpha$, vilket ger $E[Y] = \mu$ och $Var(Y) = \mu + r\mu^2$. Vi får då spridningsparametern r. Om spridningsparametern går mot 0 så går fördelningen mot poissonfördelningen.

Negativa binomialfördelningen tillhör den exponentiella dispersionsfamiljen och vi kan skriva $Y \sim NB(\mu, r)$ som

$$p(y; \mu, r) = \exp\left(y \log\left(\frac{r\mu}{1 + r\mu}\right) + \frac{1}{r}\log\left(\frac{1}{1 + r\mu}\right) + \log\left(\frac{\Gamma(r^{-1} + y)}{\Gamma(r^{-1})y!}\right)\right).$$

Vi kan se att $\theta = \log\left(\frac{r\mu}{1+r\mu}\right)$. Vi kan då konstatera att om r känt vore, enligt faktoriseringskriteriet, $T(\boldsymbol{y}) = \sum_i y_i$ en tillräcklig statistika. I vårt fall skattas r, och är alltså inte känd, vilket innebär det aggregerade datamaterialet inte är en tillräcklig statistika på grund av faktorn $\exp\left(\log\left(\frac{\Gamma(r^{-1}+y)}{\Gamma(r^{-1})y!}\right)\right)$.

I negativa binomialfördelningen finns alltså ytterligare en parameter vi behöver skatta, spridningsparametern. När vi gör GLM:er med negativa binomialfördelningen använder vi funktionen glm.nb i R och vi presenterar därför kort hur den skattas där, vilket beskrivs närmre i Venables & Ripley (2002). Vi kommer då använda en annan parametrisering än tidigare. Vi antar att intensiteten Λ i den poissonfördelade responsvariabeln är gamma(θ)/θ-fördelad. Vi får då att Λ har väntevärde 1 och varians $1/\theta$. Detta ger

$$Y|\Lambda \sim \text{Poisson}(\mu\Lambda), \quad \theta\Lambda \sim \text{gamma}(\theta)$$

och

$$E(Y) = \mu, \quad var(Y) = \mu + \mu^2/\theta.$$

Vi får sedan spridningsparametern genom en maximumlikelihoodskattning för θ, vilken hålls konstant för hela den anpassade modellen. Värt att ha i åtanke är att vi får skattningen för $\theta = 1/r$ i R, då man ofta annars arbetar med skattningen för r, vilken vi tidigare även refererat till som spridningsparametern. Vi kan konstatera, genom variansfunktionen, att negativa binomialfördelningen inte är oberoende av skala enligt definitionen för Tweedie modeller (se Avsnitt 2.5.2 Skalinvariant modell).

2.8 Verktyg för modelljämförelse

Det finns olika sätt att testa kvaliteten på modeller. Vi kommer bland annat att arbeta med de två måtten AIC och devians, vars information kan återfinnas i Agresti (2013). Vi kommer även arbeta med transformen av sannolikhetsintegral (Czado *et al.*, 2009) och Monte Carloapproximationer av p-värde (Davison & Hinkley, 2013).

2.8.1 Akaikes informationskriterium

När man ska avgöra vilken modell man vill använda finns det flera saker man bör ha i åtanke. Man vill att modellen ska förklara mycket av variationen i data. En mer komplex modell, med fler förklarande variabler är på vissa sätt närmre verkligheten än en enklare modell. Det är dock en avvägning, då en enklare modell ofta är att föredra av andra skäl. Ett mått på modellens anpassning där man tagit hänsyn till detta är Akaikes informationskriterium, AIC. Den bedömer modellen baserat på ett speciellt väntevärde för hur nära de skattade värdena ligger de verkliga. Man vill minimera måttet

$$\text{AIC} = -2(\text{maximerad loglikelihood} - \text{antal parametrar i modellen}).$$

Måttet innebär inte att man måste välja modellen med lägst AIC. Det är en riktlinje för vilka modeller som kan vara aktuella. För att måttet ska vara relevant måste man ha samma responsvariabel.

2.8.2 Devians

Ett annat mått man kan använda för modellanpassning är deviansen. Vi kommer fokusera på residualdeviansen, vilken jämför modellen man testar mot den mättade modellen. En mättad modell är en modell med alla huvudeffekter och alla möjliga samspel. Deviansen ser då ut som likelihoodkvotstatistikan där nollhypotesen är att den modell vi testar håller mot alternativhypotesen att den inte gör det och vi behåller den mättade modellen. Deviansen blir då

$$D = -2(\text{loglikelihood under } \mathrm{H}_0 - \text{loglikelihood för den mättade modellen}).$$

När D är deviansen får vi att den normerade deviansen är D/ϕ, och det måttet är alltså korrigerat för spridningsparametern.

2.8.3 Transformen av sannolikhetsintegral

Vi kommer använda oss av transformen av sannolikhetsintegral (probability integral transform), PIT, för jämförelse av fördelningar. PIT använder sig av faktumet att de värden sannolikhetsfunktionen för en antagen fördelning blir då vi observerar ett specifikt värde, y ($P(Y \leq y)$), är likformigt fördelade på intervallet (0,1). Detta om fördelningen vi antagit är kontinuerlig och observationerna kommer från den antagna fördelningen. Skadefrekvensen är dock diskret (med en offsetvariabel), vilket innebär att PIT behöver anpassas före användning. Det finns förslag både på ett randomiserat och icke randomiserat alternativ av PIT, varav vi kommer att arbeta med det icke randomiserade. Vi väljer att arbeta med det icke-randomiserade alternativet för att vi då undviker att införa extra slump i metoden. Man ställer då upp fördelningsfunktionen $F(u)$ betingat på observerade y och får

$$F(u|y) = \begin{cases} 0, & \text{då } u \leq P_{y-1} \\ (u - P_{y-1})/(P_y - P_{y-1}), & \text{då } P_{y-1} \leq u \leq P_y \\ 1, & \text{då } u \geq P_y. \end{cases}$$

Vi får alltså fördelningsfunktionen för en likformig stokastisk variabel på intervallet (0,1). För att bedöma resultatet kan man aggregera över n prediktioner och jämföra medelvärdet för PIT

$$\bar{F}\,(u) = \frac{1}{n}\sum_{i=1}^{n} F^{(i)}(u|y^{(i)}).$$

Vi kommer att kontrollera för likformighet genom att plotta ett icke-randomiserat PIT-histogram och kontrollera detta. Vi plottar PIT-histogrammet genom att räkna ut

$$f_j = \bar{F}\left(\frac{j}{J}\right) - \bar{F}\left(\frac{j-1}{J}\right),$$

där J är antalet staplar vi valt. Vi plottar histogrammet för jämt fördelade $j = 1, ..., J$ med höjden f_j för stapel j och kontrollerar det färdiga histogrammet för likformighet. Olika avvikelser från likformighet tyder på olika problem med modellvalet. Exempelvis tyder ett upp och nervänt U-format histogram på överspridning och ett U-format på underspridning.

2.8.4 Monte Carlo

Monte Carlo är en metod man kan använda för att räkna ut approximativa p-värden genom bootstrap-metoder när exakta p-värden kan vara väldigt svåra, eller omöjliga, att beräkna.

Vi har en nollhypotes H_0 och en alternativhypotes H_1. Vi vill att H_1 ska beskriva de avvikelser från H_0 som är viktigast, eller troligast, att se. Vi skapar en teststatistika T, vanligt är att man använder sig av likelihoodkvotstatistikan,

$T = -2$(maximerad loglikelihood under H_0−maximerad loglikelihood under H_1).

Vi kommer använda oss av ett grundläggande Monte Carlotest där vi jämför vårt observerade värde för vår statistika, t, med S oberoende värden på T. De S oberoende värdena på T får vi genom att generera S stycken nya slumpade stickprov där fördelningsvalet måste stämma under nollhypotesen och de nya stickproven måste vara baserade på originella datamaterialet. För oss kommer de vara baserade på det originella datamaterialet via en anpassad modell, men det kan också vara direkt. Vi får då replikerade dataset, vilka vi använder för att räkna ut de simulerade värdena på teststatistikan. Vi betecknar dessa $t_1^*, ..., t_S^*$ och vet att under H_0 är $t, t_1^*, ..., t_S^*$ lika troliga värden på T. Vi definierar k som antalet simulerade t^*-värden som är större än det observerade värdet t och får för ett kontinuerligt T Monte Carlo p-värdet

$$p = P(T \geq t|H_0) \doteq p_{mc} = \frac{k+1}{S+1}.$$

3 Modellering

I modelleringsdelen kommer vi nu att börja med att beskriva vårt datamaterial och bearbeta det i förberedelse för vidare analys. Vi kommer sedan tillämpa den teori vi precis lärt oss för att testa våra modeller och antaganden. Vi kommer att använda oss av programmet R. Modelleringsavsnittet avslutas med en kort sammanfattning på de viktigaste resultaten i modelleringen.

3.1 Datamaterial

Vi har tillgång till ett datamaterial med partiell kaskoförsäkring för motorcyklar, tillhandahållit från det tidigare försäkringsbolaget, *Wasa*. Partiell kaskoförsäkring innebär att försäkringen bland annat täcker stöld, men även vissa andra orsaker av skada på fordonet, såsom brand (Johansson & Ohlsson, 2010). Datamaterialet gäller alla försäkringar och skadeanmälningar för motorcyklar under åren 1994-1998. Datamaterialet innehåller faktorerna ägarålder, kön, geografisk zon, MC-klass, fordonets ålder, försäkringens duration i år, antal skadeanmälningar, skadeanmälningarnas kostnad och bonusklasser baserade på antalet skadeanmälningar. Man startar i bonusklass 1 och ökar en bonusklass för varje år utan skadeanmälan. Efter första fordringen minskar bonusklassen med 2 och man kan inte återkomma till bonusklass 7 förrän efter 6 åtföljande år utan skadeanmälningar.

Innan vi skapar en modell för skadefrekvensen behöver vi bearbeta våra faktorer. Först utesluter vi variabeln kön i all modellering med bakgrund av lagändring år 2012 (Kommissionens riktlinjer för könsneutrala premier). Då förbjöds könsdifferentierade premier även om skillnaden baseras på statistik. Vi bortser även från alla försäkringskontrakt med duration 0, då det ej är möjligt att skadeanmälan gjorts på dem och de därmed är irrelevanta i vidare analys.

Även alla försäkringskontrakt tillhörande en ägare yngre än 16 år har bortsetts från, då de ej får köra motorcykel. Vi behöver inte ta ställning till hantering av skadeanmälningar som inte lett till någon utbetalning då inga finns i vårt datamaterial. När denna inledande bedömning av data gjorts har vi ett material med 8 variabler och 62436 observationer, där varje försäkringskontrakt är en observation.

3.2 Premieklasser

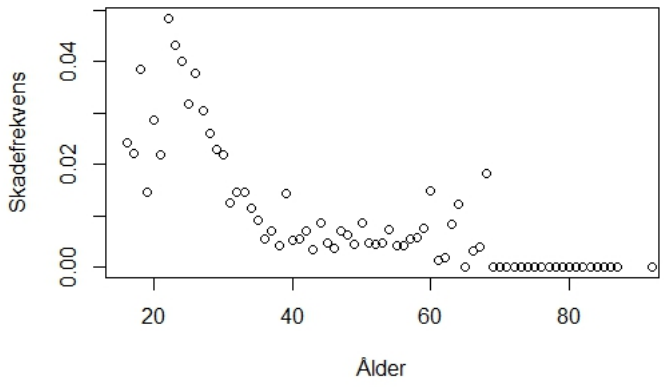
Nu går vi vidare genom att dela upp de kontinuerliga kovariaterna till kategorier för att bilda premieklasser. Vi utgår från den ursprungliga uppdelningen på försäkringsbolaget Wasa. Denna återfinns i Tabell 2.8 (Johansson & Ohlsson, 2010) och vi återger den här som Tabell 1.

Tabell 1: Faktorer och premieklasser med beskrivning

Faktor	Klass	Klassbeskrivning
Geografisk zon	1	Centrala och semicentrala delar av Sveriges tre största städer
	2	Förorter och medelstora städer
	3	Mindre städer som ej inräknas i klass 5 eller 7
	4	Små städer som ej inräknas i klass 5 eller 7
	5	Nordliga städer
	6	Nordlig landsbygd
	7	Gotland
MC-klass	1	EV-ratio -5
	2	EV-ratio 6-8
	3	EV-ratio 9-12
	4	EV-ratio 13-15
	5	EV-ratio 16-19
	6	EV-ratio 20-24
	7	EV-ratio 25-
Fordonsålder	1	0-1 år
	2	2-4 år
	3	5- år
Bonusklass	1	1-2
	2	3-4
	3	5-7

EV-ratio är ett mått på en motorcykels prestanda. Måttet baseras på motorcykelns styrka i förhållande till dess vikt, (motorns kraft i kW x 100)/(motorcykeln vikt i kg + 75) (Johansson & Ohlsson, 2010).

Den sista variabeln vi vill dela upp i klasser är ägaråldern. I teoridelen nämndes att responsens intensitet ska vara densamma i en klass. Därför utgår vi från Figur 1, en plot över skadeintensiteten mot ålder, för att dela upp i klasser med liknande intensitet.



Figur 1: Plott över skadefrekvensen mot ålder

När vi delar upp i klasser tar vi även hänsyn till att vi inte vill ha för många olika klasser. Då kan det konstateras genom Figur 1 att det blir svårt att få samma intensitet i en hel premiegrupp, och det blir mer approximativt. Uppdelningen som valdes med Figur 1 som grund kan ses i Tabell 2, vilken är en påfyllnad till Tabell 1.

Tabell 2: Fortsättning Tabell 1

Faktor	Klass	Klassbeskrivning
Ägarens ålder	1	16-24 år
	2	25-30 år
	3	31-40 år
	4	41-60 år
	5	60- år

Nu återstår valet av basklasser innan modellbyggandet kan börja. Vi vill ha den klass med högst duration som basklass. Våra basklasser presenteras i Tabell 3.

Tabell 3: Premieklassernas basklasser

Variabel		Basklass	Duration i år
Ägarålder	X_1	4	41742.3
Geografisk zon	X_2	4	32619.8
MC-klass	X_3	3	21662.3
Fordonets ålder	X_4	3	50508.3
Bonusklass	X_5	3	35726.5

Detta kommer fungera som en grund för vidare analys av skadefrekvensen. Värt att notera angående skadefrekvensen är att vi har ett datamaterial med ett fåtal observationer med en skada, ännu färre med två och ingen alls med tre skador. Exakta fördelningen kan ses i Tabell 4.

Tabell 4: Antal skador i hela datamaterialet

Antal skador	0	1	2
Antal kontrakt	61770	639	27

3.3 Skadefrekvens

Tills vidare arbetar vi med skadefrekvensen. Vilket tidigare nämnts så får vi

$$\text{Skadefrekvens} = \frac{\text{Antal skador}}{\text{Duration}}.$$

Vi kommer att modellera skadefrekvensen med hjälp av GLM:er där vi först antar poissonfördelningen och sedan jämför med negativ binomialfördelning och kvasipoisson. Vi använder naturliga logaritmen som länkfunktion och med naturliga logaritmen av durationen som offsetvariabel. Vi får modellen

$$\log\left(\frac{\mu_{ijklm}}{\text{Duration}}\right) = \lambda_0 + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^{X_3} + \lambda_l^{X_4} + \lambda_m^{X_5}.$$

Detta ger

$$\log(\mu_{ijklm}) = \lambda + \lambda_i^{X_1} + \lambda_j^{X_2} + \lambda_k^{X_3} + \lambda_l^{X_4} + \lambda_m^{X_5} + \log(\text{Duration}),$$

oavsett fördelningsantagande och vi får

$$\mu_{ijklm} = e^{\lambda} e^{\lambda_i^{X_1}} e^{\lambda_j^{X_2}} e^{\lambda_k^{X_3}} e^{\lambda_l^{X_4}} e^{\lambda_m^{X_5}} \cdot \text{Duration} = \gamma_0\gamma_{1i}\gamma_{2j}\gamma_{3k}\gamma_{4l}\gamma_{5m} \cdot \text{Duration}.$$

När vi gör denna GLM med de klasser presenterade i Tabell 1 och 2 får vi att ett flertal klasser inte är signifikant skilda från varandra. För att utveckla modellen skapar vi 95%-iga konfidensintervall för alla γ_{ij}. Vi använder oss sedan av tre villkor för att utveckla modellen.

1. Om två eller fler parametrars konfidensintervall innesluter varandras parameterskattningar läggs klasserna ihop.

   Exempel: Klass 5-7 för faktorn geografisk zon. I Tabell 5 kan vi se parameterskattningarna och deras konfidensintervall (vid poissonantagande). Samtliga parameterskattningar ligger innanför alla tre konfidensintervall och vi lägger därför ihop dessa klasser.

Tabell 5: Konfidensintervall för geografisk zon klass 5-7 vid poissonantagande

	γ_{2j}	Konfidensintervall
Geografisk zon 5	0.7942	(0.4070, 1.5500)
Geografisk zon 6	1.0858	(0.6699, 1.7599)
Geografisk zon 7	0.7020	(0.0984, 5.0101)

2. Klasser vars parameterskattningar ej är signifikanta i modellen har lagts ihop med basklassen.

   Exempel: Geografisk zon klass 5-7 var ej signifikant skilda från basklassen även efter de lagts ihop. Vi lägger därför ihop dem med basklassen för faktorn.

3. Att lägga ihop klasser har skett med en restriktion. Faktorer som ägarålder, fordonsålder och MC-klass har alla en naturlig ordning och vi lägger därför inte ihop klasser som inte angränsar.

   Exempel: I vår färdiga modell har vi en klass, MC-klass 7, som inte är signifikant skild från faktorns basklass, MC-klass 4. MC-klass 7 kan däremot inte läggas ihop med MC-klass 5 och 6, vilket är anledningen till att vi i vår färdiga modell har en klass, MC-klass 7, som ej är signifikant skild från faktorns basklass.

Parameterskattningarna för den slutliga modellen kan återfinnas i Tabell 6, där de klasser som inte omnämns är hoplagda med respektive faktors basklass. Vi slutar upp med samma uppdelning för samtliga fördelningar.

Tabell 6: Relationstal (γ_{ij}) och standardfel på linjära skalan för samtliga fördelningar.

		Relationstal		**Standardfel**		
		Kvasipoisson Poisson	Negativ Binomial	Poisson	Kvasi-poisson	Negativ Binomial
Basskadefrekvens		0.0018	0.0018	0.1021	0.1350	0.1054
Fordonsålder	1	3.4678	3.5596	0.1035	0.1368	0.1095
	2	1.9319	1.9495	0.0976	0.1290	0.1014
Ägarålder	1	6.5600	6.6997	0.1017	0.1344	0.1058
	2	3.9172	3.9897	0.0967	0.1279	0.1002
	3	1.6335	1.6348	0.1280	0.1692	0.1307
Geografisk zon	1	4.5689	4.6147	0.1023	0.1353	0.1066
	2	2.6355	2.6535	0.1029	0.1360	0.1062
	3	1.5742	1.5807	0.1128	0.1491	0.1158
MC-klass	1-2	1.3842	1.4147	0.1183	0.1563	0.1216
	5	1.6424	1.6857	0.1037	0.1370	0.1074
	6	2.9603	3.0753	0.1007	0.1331	0.1049
	7	1.8285	1.8729	0.4142	0.5474	0.4225

Vi kan se i Tabell 6 att faktorn bonusklass är helt utesluten. Anledningen till detta är att man enligt skattningarna bör få en högre premie när man tillhör en bättre bonusklass (i och med färre skador). Detta är självklart inte en önskvärd egenskap, och med anledning av detta har faktorn uteslutits. Man kan konstatera att det kan vara önskvärt att överlåta bonusklassernas inverkan på premien till att bestämmas av ekonomiska skäl utan statistik som bas.

När vi betraktar Tabell 6 bör vi ha i åtanke att parameterskattningarna är transformerade och därmed γ_{ij}-skattningarna, medan standardfelen är för parameterskattningarna innan transformering, alltså för olika λ. Med skattningarna för kvasipoisson fick vi att faktorn bonusklass ej var signifikant från början, vilket var den enda konsekvensen i klassuppdelning då vi övergick från Poisson till kvasipoisson. Valet att utesluta bonusklass helt leder därför till att vi får samma modell för de båda fördelningarna och skillnaden ligger då enbart i standardfelen. Värt att notera är att vi får större standardfel för varje parameter med kvasipoisson och spridningsparameterskattningen 1.7470. Vi får därför bredare konfidensintervall för samtliga skattningarna med kvasipoisson. AIC och devians för modellerna presenteras i Tabell 7.

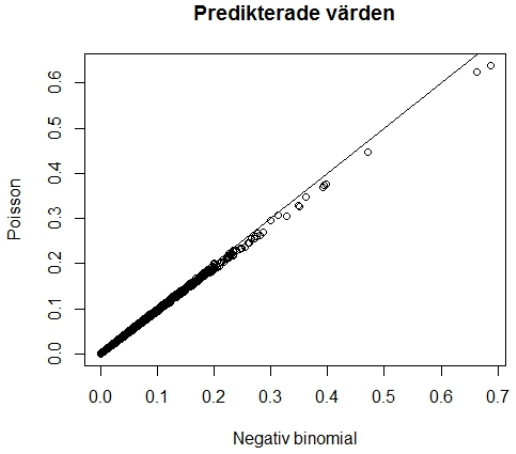
Tabell 7: AIC och devians

	Poisson	Quasipoisson	Negativ binomial
AIC	7160.25	–	7134.56
Devians	5785.68	5785.68	4747.05

Vi kan se att AIC är struken för kvasipoisson, vilket beror på att kvasipoisson inte är en specifik fördelning (enbart en variansfunktion) och vi behöver räkna ut maximerad likelihoodfunktion för att räkna ut AIC. Vi får samma devians för kvasipoisson som för Poisson för att vi arbetar med onormerad devians. Det är naturligt att vi får samma devians för dem då vi inte anpassar för spridningsparametern som är det som skiljer de två åt. Vi kan konstatera att AIC:n är 25.69 enheter mindre för negativ binomialmodellen än för poissonmodellen, vilket ger en stark indikering att negativa binomialfördelningen är ett bättre val.

När vi har de färdiga klasserna och gjort vår GLM för poissonfördelningen utför vi överspridningstestet beskrivet i Avsnitt 2.7.1 Poisson. Vi får då $p = 0.00238$ och vi kan förkasta nollhypotesen att det ej finns någon överspridning på 1% signifikansnivå.

Nu övergår vi till att försöka se hur skadefrekvensen faktiskt påverkas av att använda negativa binomialfördelningen istället för Poisson. För att tydligare se skillnaderna har vi plottat de predikterade värdena för en GLM med negativa binomialfördelningen mot de predikterade värdena för en GLM med poissonfördelningen. För att illustrera skillnaden har vi även lagt in den räta linjen $y = x$. Vi får Figur 2.



Figur 2: Predikterade värden för negativ binomial plottade mot predikterade värden för Poisson och linjen $x = y$.

Vi kan konstatera att för den lägre skadefrekvenserna verkar det inte göra någon skillnad mellan negativ binomial och Poisson. När den predikterade skadefrekvensen ligger på 0.3 kan vi se att de predikterade värdena för negativ binomial blir högre än de för Poisson. Det verkar alltså som att skillnaden mellan Poisson och negativ binomial blir mer uttalad när vi kommer till de klasser med en högre skadefrekvens.

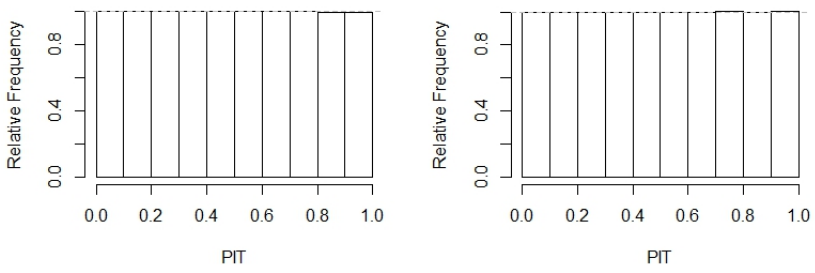
3.4 Aggregerad data

Redan i teoridelen kunde vi konstatera att det aggregerade datamaterialet representerar en tillräcklig statistika för poissonfördelningen, men inte för negativa binomialfördelningen. Då kvasipoisson inte har en fördelningsfunktion kan vi inte visa om aggregering ger en tillräcklig statistika med hjälp av faktoriseringskriteriet. Vi vet att den ger samma parameterskattningar som poissonfördelningen, men genom att utföra en GLM kan vi konstatera att standardfelen skiljer sig. Detta innebär att den aggregerade datan inte ger en tillräcklig statistika. Om man använder sig av kvasipoisson för att bilda premieklasser innebär detta att man skulle kunna få andra klasser om man aggregerar datamaterialet istället för att använda det fullständiga. När det gäller negativa binomialfördelningen skiljer sig både parameterskattningarna och standardfelen åt. För den intresserade läsaren finns relationstalen för negativa binomialfördelningen, standardfel för kvasipoisson och en plott över predikterade värden för negativa binomialfördelningen mot poissonfördelningen med aggregerad data i Appendix A.

Om vi utför test för överspridning på GLM:en med poissonfördelningen för det aggregerade datamaterialet får vi $p = 0.3707$ och kan alltså inte förkasta nollhypotesen att det inte finns någon överspridning på någon rimlig signifikansnivå.

3.5 PIT-histogram

Nu vill vi fortsätta med att kontrollera vilket fördelningsantagande som ger bäst PIT-histogram, och därmed få viss information om hur bra våra de olika fördelningsantagandena är. Kom ihåg att vi för ett bra fördelningsantagande ska få likformiga PIT-histogram. Vi använder oss av medelvärdena vi får av vår anpassade modell som väntevärde och spridningsparameterns skattning som spridningsparameter. Det ger oss de icke-randomiserade PIT-histogrammen i Figur 3.

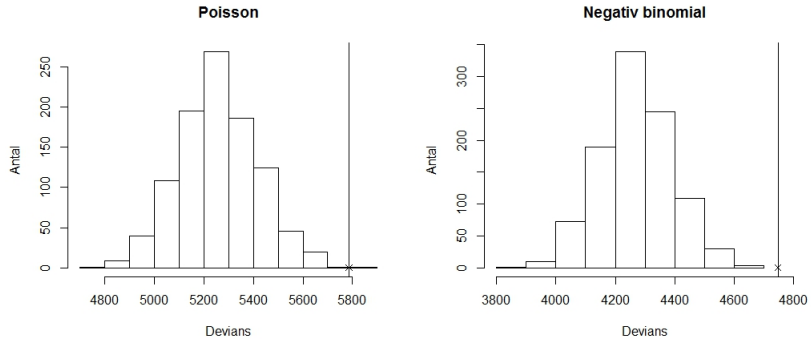


Figur 3: PIT-histogram, till vänster för Poisson och höger för negativ binomial.

Vi kan se att vi får nästintill perfekt likformiga histogram för båda fördelningarna.

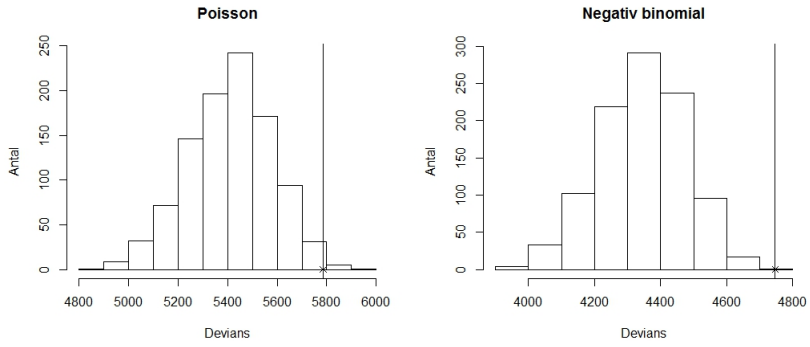
3.6 Devians

Vi fortsätter med att ta reda på om de devianser vi fått för de olika modellerna är realistiska för våra specifika fördelningsantaganden. För att göra detta har vi simulerat 1000 dataset med en ny responsvariabel, baserat på våra anpassade modeller. Sedan modellerar vi varje dataset med poissonfördelningen och negativa binomialfördelningen. Vi får då 1000 devianser för de två modellvalen, vilka kan ge oss vägledning i om våra observerade devianser avviker från de simulerade. Vi kommer enbart presentera resultaten för poissonfördelningen och negativa binomialfördelningen, eftersom vi får samma devians (och därmed samma resultat) för kvasipoisson som för Poisson. Vi börjar med att göra detta när vi simulerar nya poissonfördelade stickprov, baserat på vår GLM för poissonfördelningen och får Figur 4.



Figur 4: Devianser från poissonanpassade och negativ binomialanpassade modeller för poissonsimulerad data

Vi genomgår sedan samma procedur, fast de nya simulerade stickproven för responsvariabeln är negativt binomialfördelade baserade på vår GLM för negativa binomialfördelningen. Detta ger Figur 5.



Figur 5: Devianser från poissonanpassade och negativ binomialanpassade modeller för negativ binomialsimulerad data

De vertikala linjerna i histogrammen är markeringar för vart de observerade devianserna befinner sig. Vi kan se att i histogrammet för nya, simulerade, poissonfördelade responsvariabler ligger de observerade devianserna i utkanten av de simulerade eller utanför. Även när vi använder negativa binomialfördelningen är devianserna i utkanten av materialet.

3.7 Monte Carlo

Vi kommer nu räkna ut ett p-värde med hjälp av Monte Carloapproximation. Detta gör vi för att testa om poissonantagandet eller alternativet, negativ binomailfördelningen, ger den bättre modellen. Vi ställer upp hypoteserna

H_0: Antal skadeanmälningar, Y, är poissonfördelat.
H_1: Antal skadeanmälningar är negativt binomialfördelat.

Vi vill testa detta genom likelihoodkvotstatistikan. Vi inför beteckningen $L(\theta)$ för maximerade loglikelihoodfunktionen med index 0 och 1 under H_0 respektive H_1. Vi vill alltså ha statistikan

$$T = -2(L(\theta)_0 - L(\theta)_1).$$

Vi får denna statistika med hjälp av våra färdiga modellers AIC. I AIC:n har vi en term för antalet parametrar i modellen. Vi har en parameter mer i modellen med negativa binomialfördelningen än poissonfördelningen, vilket beror på att utöver alla parametrar som finns för modellen i poissonfördelningen finns även spridningsparametern i negativa binomialfördelningen. Vi inför beteckningen s för antalet parametrar i poissonmodellen och får

AIC_0: $-2(L(\theta)_0 - s)$
AIC_1: $-2(L(\theta)_1 - (s+1))$.

Vi får att

$$\text{AIC}_1 - 2 = -2(L(\theta)_1 - s).$$

För att få T tar vi

$$\text{AIC}_0 - (\text{AIC}_1 - 2) = -2((L(\theta)_0 - s) - (L(\theta)_1 - s)) = -2(L(\theta)_0 - L(\theta)_1) = T.$$

Detta ger det observerade värdet på statistikan $t = 27.68$. Vi simulerar nya stickprov ($S = 100$) för responsvariabeln, där de nya simulerade stickproven är poissonfördelade (överensstämmer med nollhypotesen) och baseras på vår anpassade modell. För varje dataset utför vi, alla med samma faktorer och klasser, en ny GLM med poissonfördelningen och en med negativa binomialfördelningen. Vi använder dessa modeller för att beräkna de simulerade värdena för $T, t_1^*, ..., t_{100}^*$. Vi använder detta för att beräkna p-värdet, vilket ger $p_{mc} = 0$. Vi kan därmed förkasta nollhypotesen och vi får att negativa binomialfördelningen är ett lämpligare antagande.

3.8 Huvudresultat

Vi sammanfattar nu kort de viktigaste resultaten i modelleringen. Vi kan konstatera att skillnaden i de skattade värdena för skadefrekvensen när man använder poissonfördelningen eller negativa binomialfördelningen främst märks vid de högre skadefrekvenserna, där relationstalen blir högre för negativa binomialfördelningen. I de ej randomiserade PIT-histogrammen fick vi i stort sett perfekta resultat för de olika modellerna. Vi fick lägre AIC för GLM:en med negativa binomialfördelningen, och vid likelihoodkvottestet med nollhypotesen att responsvariabeln är poissonfördelad mot att den är negativt binomialfördelad fick vi $p_{mc} = 0$. Detta tyder på att negativa binomialfördelningen är bättre anpassad för att modellera responsvariabeln. När vi tittar på histogrammen över devianser kan vi se att då vi simulerar från poissonfördelningen ligger vår observerade devians från Poisson GLM:en i utkanten och från negativ binomial GLM:en helt utanför våra observerade värden. När vi istället simulerar från negativa binomialfördelningen ligger båda observerade devianserna inom histogrammen, men fortfarande i utkanten. En möjlig tolkning är att även detta tyder på att negativa binomialfördelningen är ett mer korrekt antagande än poissonfördelningen. Detta innebär dock inte att det är ett korrekt, eller ens ett bra, antagande.

4 Diskussion

Då kvasipoisson gav samma klasser som poissonfördelningen och man får samma parameterskattningar med båda metoderna bedömdes vidare analys med kvasipoisson som överflödig. Om vi inte valt att utesluta bonusklass ur den statistiska analysen hade de båda metoderna dock gett olika klasser (då bonusklass ej var signifikant i kvasipoisson, men var signifikant för Poisson), och vid ett annat datamaterial hade de båda metoderna definitivt kunnat ge olika klasser. Man hade då kunnat få mer intressanta resultat vid fortsatt analys med kvasipoisson.

Något överraskande fick vi två nästintill perfekta PIT-histogram, vilka alltså gör att det ser ut som att både negativa binomialfördelningen och poissonfördelningen är ideala för att modellera skadefrekvensen. Detta hade varit mer troligt om vi haft en negativ binomialfördelning där spridningsparametern gick mot noll, eftersom det ger en modell vi inte kan särskilja från en poissonfördelning. Vi använder dock den spridningsparameter vi får i R, vilken inte går mot 0. Detta tyder därför inte på att båda fördelningarna är ideala för att beskriva skadefrekvensen, utan det finns något annat som påverkar. En möjlig slutsats från detta är att PIT-histogrammen inte är relevanta i vår analys, vilket är en slutsats övriga analysverktyg stöder. Möjligt är att problemen i PIT-histogrammen beror på att vår responsvariabel enbart antar tre värden varav ytterst få som ej är noll. Det skulle kunna leda till att det finns för få olika observationer att basera PIT-värdena på, vilket därmed ger i stort sett perfekta histogram.

Inferensen blir svår i ett datamaterial med så pass få skador. Antalet tvåor är så pass få relaterat till hela datamaterialet att man nästan kan bortse från dem, vilket skulle ge en responsvariabel med enbart två utfall - 0 och 1. Vi skulle då få en responsvariabel med samma egenskaper som en binär responsvariabel, trots att den egentligen kan anta fler värden. Att vi har så pass få observerade skador, och i vissa klasser inga observerade skador alls, leder till en osäkrare inferens.

Stödet i slutsatsen mot PIT-histogrammen är tydligt i våra histogram över devianser. Tydligast kan vi se det i devianserna när vi simulerat datamaterial från poissonfördelningen, där deviansen för GLM:en med poissonfördelningen hamnar långt i utkanten och deviansen för GLM:en med negativ binomialfördelningen till och med ligger helt utanför histogrammet (se Figur 4). Detta är starka indikationer på att responsvariabeln inte är poissonfördelad och att det nästintill perfekta PIT-histogrammet inte fångar upp problemen med antagandet. När vi simulerar från negativa binomialfördelningen får vi histogram som ser något troligare ut än de tidigare (se Figur 5), men trots detta kan vi konstatera att om man gör en GLM med poissonfördelningen på detta data ligger vår observerade devians åtminstone bland de 50 sista av de simulerade, och i en GLM med negativa binomialfördelningen ligger den bland de 5 sista. Detta är ett mer möjligt resultat än vid poissonsimulerad data, men tyder på att inte heller negativa binomialfördelningen är helt korrekt.

Trots att vi tidigt uteslöt kvasipoisson ur fortsatt analys kunde vi konstatera att alla standardfel blev högre vid denna än vid poissonfördelningen. När vi tillät en spridningsparameter fick vi alltså direkt högre standardfel för alla

parameterskattningar. Detta tyder på att standardfelen är för låga när vi antar poissonfördelningen. Vi har även mer statistiskt säkerställda skäl till att betrakta negativa binomialfördelningen som en starkare kandidat till skadefrekvensens fördelning. Först kan vi konstatera att vi fick det statistiskt signifikanta resultatet för förkastning av en nollhypotes att det inte finns någon överspridning och därmed acceptera mothypotesen att det faktiskt finns överspridning. Vi fick dessutom det resultatet när mothypotesen var uppställd i samma variansfunktion som negativa binomialfördelningens, vilket ger ännu starkare skäl att överväga negativa binomialfördelningen. När vi sedan gör ett likelihoodkvottest och med hjälp av Monte Carloapproximation för p-värdet kan konstatera att vi kan förkasta poissonmodellen till förmån för negativ binomialmodellen så kan vi konstatera att negativa binomialfördelningen i vår fallstudie beskriver skadefrekvensen bättre än poissonfördelningen. Vi har dock även andra komponenter att ta hänsyn till.

Vi kan dra två viktiga slutsatser angående fördelningarna redan från teoridelen. Först vet vi att poissonfördelningen är oberoende av skala, vilket negativa binomialfördelningen inte är. Vi vet även att vi får en tillräcklig statistika när vi aggregerar datamaterialet om vi arbetar med poissonfördelningen, och att vi därmed inte förlorar någon viktig information. Inte heller detta stämmer dock för negativa binomialfördelningen, vilket ger oss två fördelar med poissonfördelningen.

Att avgöra vilken fördelning som är bästa alternativet är därför inte helt rakt på sak och beror till viss del på preferenser. Vad är viktigast? Ger negativa binomialfördelningen en så pass mycket bättre modellering av data att det är värt att förlora de önskvärda egenskaperna att fördelningen är oberoende av skala och att aggregerat datamaterial ger en tillräcklig statistika? Är dessa önskvärda egenskaper så viktiga att man kan bortse från förbättringen i modellering? Eller är det helt enkelt så att man betraktar de skillnader man får i skattningarna av skadefrekvensen som så obetydliga att det är självklart att bortse från dem? Detta är svårt att avgöra med den information vi har nu och bättre överlämnat till de med större branschinformation. Det är lättare för dem att bedöma precis hur stora skillnader i skadefrekvensen som är viktiga, och precis hur viktigt det är att aggregerad data utgör en tillräcklig statistika och att modellen är oberoende av skala. Något som skulle kunna hjälpa vore möjligtvis något slags konfidensintervall för skattningarna man får i Figur 2, då dessa skulle kunna ge indikationer till om skillnaden mellan skadefrekvensen för de olika fördelningarna är statistiskt säkerställd.

Ur ett rent statistiskt perspektiv skulle jag rekommendera negativa binomialfördelningen. Detta eftersom vi kunde konstatera en skillnad i relationstalen som gör valet mellan fördelningar relevant och eftersom alla test vi gjort tyder på att den är bättre för att modellera skadefrekvensen. Ur en branschsynpunkt kan det dock, som tidigare nämnt, vara värt att bortse från dessa fördelar med negativa binomialfördelningen till förmån för möjligheten att aggregera datan och skalinvariansen vi får med poissonfördelningen.

I vårt datamaterial har vi, vilket kan ses i Tabell 4, få observationer som inte är noll, och inga som är större än två. Kanske finns sakförsäkringar över annat

än motorcyklar där det kan finnas fler skador och vara större skillnad mellan poissonfördelningen och negativa binomialfördelningen.

Slutligen bör vi även ha i åtanke att då histogrammen över devianserna (Figur 4 och 5) indikerar att då båda fördelningarna är något bristfälliga i arbetet med skadefrekvensen kan det finnas anledning att arbeta vidare och försöka hitta fördelningar bättre än de båda, eller rent av bättre metoder. Ett alternativ skulle även kunna vara att arbeta vidare med premieklasserna och se om det skulle ge någon skillnad. Något annat man skulle kunna undersöka vidare är exempelvis så kallade zero-inflated models. Dessa är ett alternativ i dataset där observationerna är överrepresenterade av nollor till så hög grad att det inte kan förklaras ens av väldigt låg intensitet i fördelningen man väljer för att modellera materialet (Hilbe, 2014).

Referenser

Agresti, Alan. 2013. *Categorical Data Analysis*. Third edn. Hoboken, New Jersey: John Wiley & Sons.

Alm, Sven Erick, & Britton, Tom. 2008. *STOKASTIK Sannolikhetsteori och statistikteori med tillämpningar*. First edn. Kina: Liber.

Cameron, A Colin, & Trivedi, Pravin K. 1990. Regression-based tests for overdispersion in the Poisson model. *Journal of econometrics*, **46**(3), 347–364.

Czado, Claudia, Gneiting, Tilmann, & Held, Leonhard. 2009. Predictive model assessment for count data. *Biometrics*, **65**(4), 1254–1261.

Davison, AC, & Hinkley, DV. 2013. *Bootstrap methods and their application*. First edn. E-book: Cambridge University Press.

Höhle, Michael, Meyer, Sebastian, & Paul, Michaela. 2015. *surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena*.

Hilbe, Joseph M. 2014. *Modeling Count Data*. First edn. E-book: Cambridge University Press.

Johansson, Björn, & Ohlsson, Esbjörn. 2010. *Non-Life Insurance Pricing With Generalized Linear Models*. First edn. E-book: Springer Science & Business Media.

Kleiber, Christian, & Zeileis, Achim. 2008. *Applied Econometrics with R*. New York: Springer-Verlag. ISBN 978-0-387-77316-2.

Olive, David J. 2014. *Statistical Theory and Inference*. First edn. E-book: Springer.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Svensk, Försäkring. 2011a. *Försäkringens historia*. http://www.svenskforsakring.se/Huvudmeny/I-fokus/Artiklar1/Kategorier/2011/Kommissionens-riktlinjer-for-konsneutrala-premier/. [Online; hämtad 22-April-2015].

Svensk, Försäkring. 2011b. *Kommissionens riktlinjer för könsneutrala premier*. http://www.svenskforsakring.se/Huvudmeny/I-fokus/Artiklar1/Kategorier/2011/Kommissionens-riktlinjer-for-konsneutrala-premier/. [Online; hämtad 22-April-2015].

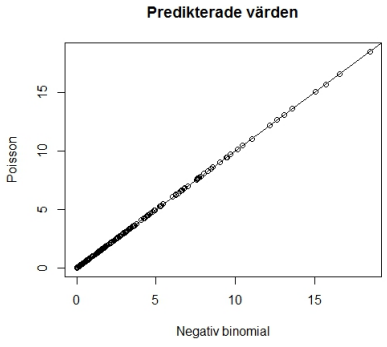
Venables, William N, & Ripley, Brian D. 2002. *Modern Applied Statistics with S*. Fourth edn. E-book: Springer Science & Business Media.

Appendix A

Tabell 8: Relationstal (γ_{ij}) för negativa binomialfördelningen och standardfel för kvasipoisson, med ej aggregerad och aggregerad data

		Negativ binomial		**Kvasipoisson**	
		Aggregerad	Ej aggregerad	Aggregerad	Ej aggregerad
Basskadefrekvens		0.0018	0.0018	0.1061	0.1350
Fordonsålder	1	3.4682	3.5596	0.1076	0.1368
	2	1.9318	1.9495	0.1014	0.1290
Ägarålder	1	6.5710	6.6997	0.1057	0.1344
	2	3.9228	3.9897	0.1005	0.1279
	3	1.6355	1.6348	0.1330	0.1692
Geografisk zon	1	4.5779	4.6147	0.1064	0.1353
	2	2.6387	2.6535	0.1069	0.1360
	3	1.5766	1.5807	0.1172	0.1491
MC-klass	1-2	1.3837	1.4147	0.1229	0.1563
	5	1.6438	1.6857	0.1077	0.1370
	6	2.9557	3.0753	0.1046	0.1331
	7	1.8284	1.8729	0.4304	0.5474

Först kan vi se skillnaden i de skattade relationstalen, i formen γ_{ij}. Sedan ser vi skillnaden i standardfelen för de icke transformerade skattningarna med kvasipoisson. I Figur 6 kan vi se de predikterade värdena för negativ binomialmodellen plottade mot de för poissonmodellen för aggregerat datamaterial. I Figur 6 finns även en linje $y = x$ för att illustrera skillnader.



Figur 6: Predikterade värden för negativ binomial plottade mot predikterade värden för Poisson vid aggregerad data.