Matematisk statistik
Stockholms universitet

En jämförande studie av GLM, Jungs metod och Tweedie-modell för premiesättning av multiplikativ tariff.

Elin Larsson

Examensarbete 2004:15

Postal address:

Matematisk statistik
Dept. of Mathematics
Stockholms universitet
SE-106 91 Stockholm
Sweden

Internet:

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

Matematisk statistik
Stockholms universitet
Examensarbete **2004:15**,
http://www.math.su.se/matstat

En jämförande studie av GLM, Jungs metod och Tweedie-modell för premiesättning av multiplikativ tariff.

Elin Larsson*

Augusti 2004

Abstract

Inom skadeförsäkring finns det olika metoder för att skatta relationstalen i en multiplikativ tariff. I det här examensarbetet har Jungs marginalsummemetod, GLM och en Tweedie-modell undersökts. Syftet med arbetet var att jämföra metoderna i tre avseenden: skattade relationstal, varianser samt konfidensintervall. Jämförelsen har gjorts dels på verkliga data, dels på simulerade data. I de olika simuleringsfallen har olika förhållanden undersökts, som till exempel att skadedata har simulerats från gamma- samt lognormalfördelning.

Resultaten har inte varit entydiga, men den metod som bäst klarade av de förhållanden som undersöktes i det här arbetet var GLM och andra bäst var Tweedie-modellen.

*Postal address: Matematisk statistik, Stockholms universitet, SE-106 91, Stockholm, Sweden. E-mail: elin.larsson@trygghansa.se. Handledare: Rolf Sundberg

En jämförande studie av GLM, Jungs metod och Tweedie-modell för premiesättning av multiplikativ tariff.

Elin Larsson, augusti 2004

Examensarbete i matematisk statistik vid Stockholms universitet

2004:15

Sammanfattning

Inom skadeförsäkring finns det olika metoder för att skatta relationstalen i en multiplikativ tariff. I det här examensarbetet har Jungs marginalsummemetod, GLM och en Tweedie-modell undersökts. Syftet med arbetet var att jämför metoderna i tre avseenden; skattade relationstal, varianser samt konfidensintervall. Jämförelsen har gjorts del på verkliga data, dels på simulerade data. I de olika simuleringsfallen har olika förhållanden undersökts, som till exempel att skadedata har simulerats från gamma- samt lognormalfördelning.

Resultaten har inte varit entydiga, men den metod som bäst klarade av de förhållanden som undersöktes i det här arbetet var GLM och andra bäst var Tweedie-modellen.

Summary

There are several different methods used in non-life insurance to estimate the rating factors in a multiplicative tariff. In this paper three methods are examined: method of marginal totals (Jung's method) and methods based on multiplicative GLM and Tweedie models. The aim is to compare how these methods estimate rating factors, their bias and precision, and their confidence intervals. Comparisons have been made on both real data and simulated data. In the simulations, different conditions have been examined, and for example the claim amounts have been simulated from both the gamma and the lognormal distributions.

The results were not unambiguous, but overall, GLM was the method that performed best in relation to the factors examined in this paper. The Tweedie method was judged second best.

Elin.Larsson@trygghansa.se
Handledare:
Rolf Sundberg, Matem. statistik, Stockholms universitet,
Roland Svensk, Försäkringsbolaget Trygg-Hansa

Innehåll

1. Inledning

Ett sakförsäkringsbolag erbjuder kunder att ersätta eventuella framtida skadekostnader mot en avgift, premien. Premie skall täcka försäkringsbolagets åtaganden mot kunden, riskpremien, samt deras administrativa kostnader och vinst. Hur stor riskpremie kunderna betalar för en försäkring beror dels på hur stor risken är att de råkar ut för en skada, dels på hur stort det eventuella skadebeloppet blir. Olika försäkringstagare motsvarar olika stora risker. När riskpremien beräknas undersöks därför olika premieargument (variabler) som kan tänkas påverka risken för försäkringsbolaget. Ofta antas en multiplikativ modell för väntevärdet av skadekostnaden. Den riskpremie som kunderna betalar ska motsvara den förväntade skadekostnaden för den kategorin av kunder.

1.1. Problemställning och syfte

I en tariff bestäms premien med hjälp av premieargument, där risken i olika klasser inom ett premieargument sätts i relation till varandra. Det finns olika metoder för att bestämma relationstalen i en multiplikativ tariff. En metod som tidigare varit vanlig är marginalsummemetoden, som även kallas Jungs metod. Numera har metoder baserade på generaliserade linjära modeller, GLM, blivit alltmer vanliga. Tweedie-modellerna, som är ett specialfall av GLM, ger upphov till ännu en metod. Syftet med examensarbetet är att jämföra dessa metoder i tre avseenden: skattade relationstal med tillhörande varianser och konfidensintervall.

Jämförelsen görs dels på ”verkliga” data, dels på simulerade data. Frågor som jag hoppas kunna besvara med arbetet är vilken metod som är mest tillförlitlig. Vad händer om vissa modellantaganden inte är uppfyllda, t ex om medelskadan har en annan fördelning än vad metoderna utgår från. Är det någon av metoderna som klara detta bättre än de övriga?

1.2. Avgränsning

Både GLM- och Tweedie-modellklasserna innefattar många olika modeller. Inom GLM kommer den modell att undersökas där antalet skador är Poisson-fördelat och skadebeloppet är gammafördelat. För att skilja denna modell från övriga GLM kallar jag den för Standard-GLM. Inom Tweedie-modellerna undersöks den sammansatta poissonfördelningen direkt på riskpremien. (Se appendix för tabell över de undersökta metoderna.)

När det gäller skattade varianser och konfidensintervall i anslutning till Jungs metod kommer jag att använda den LF-Wasa metod som Stig Rosenlund skriver om i *Evaluation of GLM in non-life insurance*s. Motsvarande beräkningar inom GLM kommer jag att göras enligt den metod som Esbjörn Ohlsson och Björn Johansson tar upp i *Prissättning inom sakförsäkring med Generaliserade linjära modeller*.

2. Bakgrundsteori

2.1. Begrepp

Inom sakförsäkringsmatematiken definieras följande begrepp:

- Riskpremien, Y, total skadekostnad per försäkringsår.
- Medelskadan, M, total skadekostnad dividerat med antal skador.
- Skadefrekvens, S, antal skador per försäkringsår.
- Duration, w, är den tid som försäkringen gällt.

Riskpremien = Medelskadan _ Skadefrekvensen

2.2. Modellantagande

Inom sakförsäkring finns det tre modellantaganden, som gör det möjligt att utforma teorier för riskpremien.

- För olika försäkringsavtal är utfallen (antal skador och skadekostnader) oberoende av varandra.
- I disjunkta tidsintervall är utfallen oberoende av varandra.
- Utfallen för två försäkringsavtal, med samma exponering (antal skador och duration), i samma tariffcell, har samma fördelning.

Det man vill åstadkomma vid indelningen av försäkringarna är att skillnaden mellan olika tariffceller ska vara stor i jämförelse med inom tariffcellerna. [4]

2.3. Multiplikativ modell

Riskpremien ska motsvara den för försäkringsbolaget förväntade skadekostnaden. Vid beräkning av riskpremien har man olika premieargument (variabler) som man tror påverkar bolagets risk eller förväntade skadekostnad. Varje premieargument är i sin tur indelat i olika klasser. Målet är att dela in försäkringsavtal med lika stor risk i samma grupp. På så vis betalar var och en av försäkringstagarna för sin egen risk/förväntade kostnad. Det vanliga är att man ansätter en multiplikativ modell för den förväntade kostnaden för en försäkring med riskpremien $Y_{ijk\ldots}$. (1) (Den multiplikativa modellen används även vid analys av skadefrekvensen och medelskadan.) Modellen parameteriseras vanligen så att man har en bascell och de övriga parametrarna uttrycker hur risken avviker från bascellen.

$$E(Y_{ijk\ldots}) = \gamma_0 \gamma_{1i} \gamma_{2j} \gamma_{3k} \cdots$$

där γ_0 är baspremien och $\gamma_{11} = \gamma_{21} = \gamma_{31} = \cdots = 1$ är bas för respektive premieargument.

γ_{1i} är relationstalen för premieargument 1. (1)

En multiplikativ modell är rimlig eftersom den innebär att förändringar av riskpremien blir relativa. Detta kan lätt inses med ett enkelt exempel. Om man har två

premieargument, ålder och geografisktområde, och premien ändras för ålder så blir den relativa förändringen densamma oberoende av var man bor. Om man istället har en additiv modell och premien ändras för ålder så påverkas premiens procentuella förändring av var man bor.

3. Metoderna

3.1. Jungs marginalsummemetod

En av marginalsummemetodens upphovsmän är svensken Jan Jung, som var med och utvecklade den under slutet 1960-talet. [2] Det metoden säger är att skattningarna av tariffcellernas förväntade skadekostnad ska ha samma marginalsummor som de observerade skadekostnaderna. Ekvationssystemet (4) nedan visar marginalsummemetodens skattningar av relationstalen. Endast två premieargument är med för att få enklare notation. Ekvationerna fick Jung fram genom att maximera poissonfördelningens likelihoodfunktion. Enligt Jung ger ekvationerna alltid väntevärdesriktiga skattningar av marginalsummorna även om antagandet om poissonfördelning inte gäller eller om modellen man antagit inte stämmer. Det är alltså en fördelningsfri metod. [2]

X_{ij} = skadekostnad i tariffcell i j

w_{ij} = duration

$Y_{ij} = X_{ij} / w_{ij}$

Detta ger följande ekvationssystem:

$$\sum_j w_{ij}\gamma_0\gamma_{1i}\gamma_{2j} = \sum_j w_{ij}Y_{ij} \qquad i = 1\ldots p$$
$$\sum_i w_{ij}\gamma_0\gamma_{1i}\gamma_{2j} = \sum_i w_{ij}Y_{ij} \qquad j = 1\ldots q \qquad (2)$$

Om man summerar den första ekvationen över i kan man sedan lösa ut γ_0.

γ_{1i} löser man ut ur den första ekvationen som funktion av γ_{2j} - na, och vice versa.

$$\gamma_0 = \frac{\sum_{ij} w_{ij}Y_{ij}}{\sum_{ij} w_{ij}\gamma_{1i}\gamma_{2j}}$$

$$\gamma_{1i} = \frac{\sum_j w_{ij}Y_{ij}}{\gamma_0 \sum_j w_{ij}\gamma_{2j}} \qquad i = 1\ldots p$$

$$\gamma_{2j} = \frac{\sum_i w_{ij}Y_{ij}}{\gamma_0 \sum_i w_{ij}\gamma_{1i}} \qquad j = i\ldots q \qquad (3)$$

Ekvationssystemet (3) löses iterativt.

3.2. LF-Wasa metoden

Skattning av varianser samt beräkning av tillhörande konfidensintervall har i detta arbete gjorts enligt LF-Wasa metoden, som utgår från Jungs metod. I LF-Wasa metoden används kvadratsummorna av skadekostnaderna vid bildandet approximativa konfidensintervall för riskpremiens relationstal. I metoden antas inget om skadekostnadens fördelning, utan den utgår från att antalet skador är poissonfördelat och att den totala skadekostnaden då blir sammansatt poissonfördelad. [7]

Exempel med två variabler :

Z_{kij} är skadekostnaden för skada $k = 1,2,3,\ldots,N_{ij}$ i tariffcell ij .

N_{ij} antal skador i tariffcell ij med duration w_{ij}

$N_{ij} \sim$ *Poissonfördelad* $(\ddot{e}_{ij} w_{ij})$

$Z_{1ij}, Z_{2ij}, Z_{3ij}, ..., Z_{N_{ij}ij}$ är oberoende och lika fördelade med en fördelning som får bero på ij, men inte på N_{ij} .

$Y_{ij} = \frac{1}{w_{ij}} \sum_{k=1}^{N_{ij}} Z_{kij}$ ~ Sammansatt Poissonfördelning för fixerat ij.

För att få fram variansen för den sammansatta Poissonfördelningen kan den kumulantgenererande funktionen, Ψ(t), deriveras. (Se appendix för motivering till Ψ(t).)

För att notationen ska bli enklare har jag inte tagit med index ij nedan.

$\Psi(t) = \Psi_N(\Psi_{Z_1/w}(t))$

$\Psi'(t) = \Psi'_N(\Psi_{Z_1/w}(t))\Psi'_{Z_1/w}(t)$

$\Psi''(t) = \Psi''_N(\Psi_{Z_1/w}(t))(\Psi'_{Z_1/W}(t))^2 + \Psi'_N(\Psi_{Z_1/w}(t))\Psi''_{Z_1/w}(t)$

$$E(Y) = \Psi'(0) = \Psi'_N(\Psi_{Z_1/w}(0))\Psi'_{Z_1/w}(0) = \Psi'_N(0)\Psi'_{Z_1/w}(0) = E(N)E(Z_1)/w = w\lambda E(Z_1)/w$$

$$E(Y) = \lambda E(Z_1)$$

$$Var(Y) = \Psi''(0) = \Psi''_N(\Psi_{Z_1/w}(0))(\Psi'_{Z_1/w}(0))^2 + \Psi'_N(\Psi_{Z_1/w}(0))\Psi''_{Z_1/w}(0)$$
$$= \Psi''_N(0)(\Psi'_{Z_1/w}(0))^2 + \Psi'_N(0)\Psi''_{Z_1/w}(0) = Var(N)E(Z_1/w)^2 + E(N)Var(Z_1/w)$$
$$= w\lambda(E(Z_1)^2 + Var(Z_1))/w^2 = \lambda E(Z_1^2)/w$$

$$Var(Y) = \lambda E(Z_1^2)/w \qquad (4)$$

Nu har vi fått fram variansen för riskpremien, Y, i en tariffcell. Eftersom den totala skadekostnaden i varje del av tariffcellen kan antas vara sammansatt poissonfördelad, kan man enligt LF-Wasa metoden skatta variationskoefficienten, v_{rs}, för observerad riskpremie i premieargument *r* och klass *s,* genom att dividera skadekostnadens kvadratsummor med sammanlagd skadekostnad i kvadrat. (5) Dessa variationsskattningar används även som variansskattning för riskpremiens relationstal, γ_{rs}. [7]

$$\hat{v}_{rs}^2 = \left(\sum Z^2{}_{ij}\right) / \left(\sum Z_{ij}\right)^2 \qquad (5)$$

Summerar över skadekostnader som tillhör premieargument *r* och klass *s*.
[7]

Nu kan man enligt LF-Wasa metoden bilda approximativa konfidensintervall för de skattade relationstalen, $\gamma_{rs,}$, genom att utgå från att logaritmen av γ_{rs} är approximativt normalfördelad då durationen är stor. [7] Riskpremien, Y_{rs}, är approximativt Normalfördelad eftersom den har bildats av en summa. Även riskpremiens relationstal, $\gamma_{sk,}$ och logaritmen av γ_{rs} kan antas vara approximativt normalfördelade eftersom divisions- och logaritmfunktionen är, för korta intervall, approximativt linjära. Vi får här ett 95 procentigt konfidensintervall för den förväntade skadekostnaden i premieargument *r* klass *s*.

$$\log(\gamma_{rs}) \pm 1{,}96\hat{v}_{rs} \quad \Leftrightarrow \quad \gamma_{rs} \exp\left\{\pm 1{,}96\hat{v}_{rs}\right\} \qquad (6)$$

Konfidensintervall för relationstal med klass ett som bas får man genom att anta att logaritmen av γ_{rs}/γ_{r1} är approximativt Normalfördelad med variansen $v^2{}_{rs}+v_{r1}{}^2$. [5]

$$\frac{\hat{\gamma}_{rs}}{\hat{\gamma}_{r1}} \exp\left\{\pm 1{,}96\sqrt{\hat{v}^2{}_{r1} + \hat{v}^2{}_{rs}}\right\} \ (95\%) \qquad (7)$$

3.3.2. Motivering till LF-Wasa metoden.

Vid motiveringen till variansskattningen har Esbjörn Ohlssons *Härledning av ad hoc konfidensintervall¨* använts. Först undersöks fallet med endast ett premieargument. Då motsvaras riskpremiens relationstal av väntevärdet av riskpremien i tariffcellerna. Man kan därför skatta variansen för riskpremien i (4) när man ska skatta relationstalens varianser. För att skatta variansen i ekvation (4) behövs även skattningar för poissonfördelningens parameter, λ, och andra momentet för skadekostnaden.

$\hat{\lambda} = \frac{N}{w}$

$E(Z)$skattas med $\frac{1}{N}\sum_{1}^{N} Z_j$ och $E(Z^2)$ skattas med $\frac{1}{N}\sum_{1}^{N} Z_j{}^2$

Nu får man från (5)att

$E(Y) = \lambda E(Z)$ skattas med $\hat{\lambda} \times \frac{1}{N}\sum_{1}^{N} Z_j = \frac{N}{w} \times \frac{1}{N}\sum_{1}^{N} Z_j = \frac{1}{w}\sum_{1}^{N} Z_j$ och

$Var(Y) = \lambda E(Z^2)w$ skattas med $\hat{\lambda} \times \frac{1}{N}\sum_{1}^{N} Z_j{}^2 \times \frac{1}{w} = \frac{N}{w} \times \frac{1}{N}\sum_{1}^{N} Z_j{}^2 \times \frac{1}{w} = \frac{1}{w^2}\sum_{1}^{N} Z_j{}^2$ (8)

förutsatt att det bara finns ett premieargument.

Den empiriska skattningen, Y, av den förväntade skadekostnaden för en försäkring under ett år, μ, är approximativt normalfördelad. Genom att använda felfortplantningsformlerna (se appendix) får man fram väntevärde och varians för logaritmen av Y. Den skattade variansen i denna approximation ger LF-Wasa metodens variansskattning och konfidensintervall.

Om Y är approximativt $\sim N(\mu, \sigma^2)$, så är $\log(Y)$ approximativt $\sim N(\log(\mu), \sigma^2/\mu^2)$.
I vårt fall skattas $Var(\log(Y)) = Var(Y)/E(Y)^2$ med $\frac{1}{w^2}\sum Z_j^2 \Big/ \left(\frac{1}{w}\sum Z_i\right)^2 = \sum Z^2{}_i \Big/ \left(\sum Z_i\right)^2$,
vilket överensstämmer med variationsskattningen, $\hat{v}^2$, i (6).

Genom tariffcellernas oberoende fås skattade varianser för relationstalen.

$$\log(\hat{y}_k) = \log\left(\frac{Y_k}{Y_1}\right) \sim N\left(\log(\mu_k) - \log(\mu_1), \sqrt{\hat{v}_1^2 + \hat{v}_k^2}\right) \qquad (9)$$

Vi ser i ekvation (9) att då vi endast har ett premieargument får vi samma variansskattning för relationstalen som i LF-Wasa metoden. Den kritik som riktats mot LF-Wasa metoden är att den inte är fullt motiverad då man har fler än ett premieargument. Om man försöker att skatta variansen för relationstalen då man har två premieargument ser det ut på följande sett:

Om man har följande multiplikativa modell

$\mu_{ij} = \alpha_i \beta_j$

Skattar man α_i och β_j enligt Jungs modell får man :

$$\hat{\alpha}_i = \frac{\sum_j w_{ij} Y_{ij}}{\sum_j w_{ij}\hat{\beta}_j} = \frac{w_{i.} Y_{i.}}{\sum_j w_{ij}\hat{\beta}_j} \qquad i = 1...p$$

$$\hat{\beta}_j = \frac{\sum_i w_{ij} Y_{ij}}{\sum_i w_{ij}\hat{\alpha}_i} = \frac{w_{.j} Y_{.j}}{\sum_i w_{ij}\hat{\alpha}_i} \qquad j = 1...q$$

$Y_{i.} = \frac{1}{w_{i.}}\sum_j w_{ij} Y_{ij}$ marginalen av riskpremien i klass i med avseende på premieargument α (10)

Om basklassen i båda premieargumenten är fem får man följande skattningar av relationstalen i klass ett.

$$\hat{\gamma}_{11} = \frac{\hat{\alpha}_1}{\hat{\alpha}_5} = \frac{\sum_j w_{5j}\hat{\beta}_j}{\sum_j w_{1j}\hat{\beta}_j} \times \frac{w_{1.}Y_{1.}}{w_{5.}Y_{5.}} = \frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j} \times \frac{Y_{1.}}{Y_{5.}}$$

$$\hat{\gamma}_{21} = \frac{\hat{\beta}_1}{\hat{\beta}_5} = \frac{\sum_i w_{i5}\hat{\alpha}_i}{\sum_i w_{i1}\hat{\alpha}_i} \times \frac{w_{.1}Y_{.1}}{w_{..5}Y_{..5}} = \frac{\frac{1}{w_{.5}}\sum_i w_{i5}\hat{\alpha}_i}{\frac{1}{w_{.1}}\sum_i w_{i1}\hat{\alpha}_i} \times \frac{Y_{.1}}{Y_{..5}}$$

Om man logaritmerar detta uttryck får man

$$\log(\hat{\gamma}_{11}) = \log\left[\frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j}\right] + \log(Y_{1.}) - \log(Y_{5.})$$

$$\log(\hat{\gamma}_{21}) = \log\left[\frac{\frac{1}{w_{.5}}\sum_i w_{i5}\hat{\alpha}_i}{\frac{1}{w_{.1}}\sum_i w_{i1}\hat{\alpha}_i}\right] + \log(Y_{.1}) - \log(Y_{..5}) \qquad (11)$$

Först undersöks om skattningen i ekvation (11) är konsistent. (Se appendix för definition av konsistent.)

$$E(Y_{ij}) = \mu_{ij} = \alpha_i\beta_j$$

$$E(Y_{5.}) = \frac{1}{w_{5.}}\sum_j w_{5j}E(Y_{5j}) = \frac{\alpha_5}{w_{5.}}\sum_j w_{5j}\beta_j$$

$$E(\log(Y_{5.})) = \log\left(\frac{\alpha_5}{w_{5.}}\sum_j w_{5j}\beta_j\right)$$

$Y_{i.}$ och $Y_{.j}$ konvergerar mot sina väntevärden då informationen växer, vilket leder till att ekvationerna i (11) går mot följande:

$$\log(\hat{\gamma}_{11}) \to \log\left[\frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j}\right] + \log\left(\frac{\alpha_1}{w_{1.}}\sum_j w_{1j}\beta_j\right) + \log\left(\frac{\alpha_5}{w_{5.}}\sum_j w_{5j}\beta_j\right) =$$

$$= \log\left[\frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j}\right] + \log\left(\frac{\frac{\alpha_1}{w_{1.}}\sum_j w_{1j}\beta_j}{\frac{\alpha_5}{w_{5.}}\sum_j w_{5j}\beta_j}\right)$$

$$\log(\hat{\gamma}_{21}) \to \log\left[\frac{\frac{1}{w_{.5}}\sum_i w_{i5}\hat{\alpha}_i}{\frac{1}{w_{.1}}\sum_i w_{i1}\hat{\alpha}_i}\right] + \log\left(\frac{\frac{\beta_1}{w_{.1}}\sum_i w_{i1}\alpha_i}{\frac{\beta_5}{w_{.5}}\sum_i w_{i5}\alpha_i}\right)$$

Lösningen till dessa två ekvationer är $\begin{cases}\hat{\alpha}_i = \alpha_i \\ \hat{\beta}_j = \beta_j\end{cases}$

Asymptotiskt är skattningarna de rätta och alltså är de konsistenta. (12)

Om durationen har en multiplikativ struktur så blir den första termen i ekvation (11) noll. Det visas nedan i (13) för en två gånger två tabell.

Multiplikativ struktur för antal observationer per cell, 2×2 - tabell

$w_{ij} = n_{1i} \times n_{2j} \times c \qquad i, j = 1,2$

w_{ij} durationen i tariffcell ij

n_{1i} durationen i premieargument 1 klass i

n_{2j} durationen i premieargument 2 klass j

c = konstant

Undersöker endast den första termen i högerledet i ekvation (11)

$$\log\left[\frac{\frac{1}{w_{2.}}\sum_j w_{2j}\beta_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\beta_j}\right] = \log\left[\frac{\frac{1}{\sum_j (n_{12}n_{2j}c)}\times(w_{21}\beta_1 + w_{22}\beta_2)}{\frac{1}{\sum_j (n_{11}n_{2j}c)}\times(w_{11}\beta_1 + w_{12}\beta_2)}\right] =$$

$$= \log\left[\frac{\frac{1}{n_{12}\sum_j n_{2j}}\times(n_{12}n_{21}c\beta_1 + n_{12}n_{22}c\beta_2)}{\frac{1}{n_{11}\sum_j n_{2j}}\times(n_{11}n_{21}c\beta_1 + n_{11}n_{22}c\beta_2)}\right] = \log\left[\frac{\frac{1}{n_{12}}\times n_{12}c\times(n_{21}\beta_1 + n_{22}\beta_2)}{\frac{1}{n_{11}}\times n_{11}c\times(n_{21}\beta_1 + n_{22}\beta_2)}\right] = 0 \qquad (13)$$

Då durationen har en multiplikativ struktur är det därför enklare att beräkna väntevärde och varians för uttrycket i (11). Först beräknas väntevärdet för ekvation (11).

$$E(\log(\hat{\gamma}_{11})) = E\left[\log\left[\frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j}\right]\right] + E(\log(Y_{1.})) + E(\log(Y_{5.})) =$$

$$= \log\left[\frac{\frac{1}{w_{5.}}\sum_j w_{5j}\hat{\beta}_j}{\frac{1}{w_{1.}}\sum_j w_{1j}\hat{\beta}_j}\right] + \log\left(\frac{\frac{\alpha_1}{w_{1.}}\sum_j w_{1j}\beta_j}{\frac{\alpha_5}{w_{5.}}\sum_j w_{5j}\beta_j}\right) = \left\{\begin{matrix}\text{om durationen har en}\\ \text{multiplikativ struktur}\end{matrix}\right\} = \log\left(\frac{\alpha_1}{\alpha_5}\right) = \log(\gamma_1) \qquad (14)$$

Variansen för ekvation (11) skattas på följande sätt då durationen har en multiplikativ struktur.

$Var(Y_{ij})$ skattas ju med $\frac{1}{w_{ij}^{\ 2}}\sum_{k=1}^{N_k} Z_{ijk}^{\ 2}$

$Var(Y_{5.}) = \frac{1}{w_{5.}^{\ 2}}\sum_j w_{5j}^{\ 2} Var(Y_{5j})$ skattas därför med $\frac{1}{w_{5.}^{\ 2}}\sum_j\sum_k Z_{5jk}^{\ 2}$

Med samma motivering som då vi hade ett premieargument blir :

$Var(\log(Y_{5.})) \approx Var(Y_{5.})/E(Y_{5.})^2$ som är $\nu_{5.}^{\ 2}$

Då durationen har en multiplikativ struktur kan man bortse från första termen i ekvation (11).

$$Var(\hat{\gamma}_{11}) = \left\{\begin{matrix}\text{om durationen har en}\\ \text{multiplikativ struktur}\end{matrix}\right\} = Var(\log(Y_{5.}) - \log(Y_{1.})) = \nu_{5.}^{\ 2} + \nu_{1.}^{\ 2} \qquad (15)$$

Ekvation (14) visar att om man har en multiplikativ struktur på durationen så blir skattningarna av relationstalen i (11) väntevärdesriktiga. Då man har fler än ett premieargument blir variansskattningen korrekt under multiplikativ duration.

3.3. GLM

I den linjära modellen har man $y_1,\ldots, y_n$ som är observationer av de stokastiska variablerna $Y_1,\ldots, Y_n$. Dessa är oberoende och har gemensam varians. Väntevärdena μ_i, i=1…n, kan uttryckas linjärt med hjälp av ett mindre antal modellparametrar. Och det är detta som är det fundamentala i den linjära modellen. [10] $Y_i=\mu_i+\varepsilon_i$ kan alltså skrivas $Y_i=\mathbf{x}_i'\boldsymbol{\beta}+\varepsilon_i$.

Generaliserade linjära modeller är en utökning av den linjära modellen. Här tillåter man att väntevärdets linjära struktur skapas genom en länkfunktion, $g(\mu_i)=\mathbf{x}_i'\boldsymbol{\beta}$. I detta arbete har man en logaritmisk länk på väntevärdet eftersom man utgår från en multiplikativ modell. Då data är på listform istället för på tabellform innebär den logaritmiska länkfunktionen följande:

$E(Y_i) = \mu_i$ där Y_i är nyckeltal för tariffcell i

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

x_{ij} är en indikator variabel som talar om när β_j ska vara med. (16)

En annan skillnad, från den linjära modellen, är att de stokastiska variablerna inte behöver vara normalfördelade utan kan komma från den familj av sannolikhetsfördelningar som kallas för exponentiella dispersionsmodeller, EDM. Exempel på fördelningar som tillhör EDM är normal-, poisson-, gamma-, lognormal- och paretofördelningen. Alla täthetsfunktioner och sannolikhetsfunktionen, som tillhör EDM, kan beskrivas genom frekvensfunktionen som visas i (17). Vilken EDM-fördelning det är bestäms entydigt av variansfunktionen. [4] Eftersom man har en hel familj av fördelningar kan man ta fram modeller som gäller generellt för alla fördelningar som tillhör EDM.

frekvensfunktionen (17)

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

$E(Y_i) = b'(\theta_i)$

$Var(Y_i) = b''(\theta_i)\frac{\phi}{w_i}$

där $v(\mu_i) = b''(\theta_i)$ kallas variansfunktion

$f_{Y_i}(0) = 0, \quad w_i \geq 0, \quad \phi > 0,$

$b'(\theta_i)$ är två gånger kontinuerligt deriverbar och inverterbar.

3.3.1. Skadefrekvensen

I GLM kan man skatta skadefrekvensen och medelskadan var och en för sig. Genom att faktorisera likelihoofunktionen i en faktor som är likelihooden för frekvensdata och en som är den av frekvensdata betingade likelihooden för skadedata, så kan man separera inferenserna för de två olika grupperna av parametrar.

$$L(s, m) = L(s)L(m|s)$$

Antalet skador som inträffar inom en viss tariffcell antas vara poissonfördelade. Att detta antagande är rimligt kan man se genom att titta på modellantagandena i kapitel två och definitionen av en Poisson-process (se appendix). Enligt modellantagandena är utfallen i disjunkta tidsintervall oberoende av varandra och två avtal med samma exponering inom samma tariffcell har samma fördelning för utfallen. Det sista antagandet innebär att det som är av betydelse är hur länge och inte när ett avtal gällt. Detta innebär att denna räkneprocess har stationära och oberoende inkrement vilket är ett av villkoren för att det ska vara en Poisson-process. Man kan anta att skadorna inträffar en och en vilket stämmer med nästa villkor för Poisson-processen. [4] I en Poisson-process är antalet skador som inträffar under en viss tidsperiod poissonfördelat. [8] Att antalet skador som inträffar under en viss tidsperiod är Poissonfördelat är alltså ett rimligt antagande.

Skadefrekvensen, $S_i = N_i / w_i$ (18)

N_i antal skador i tariffcell i med duration w_i

$N_i \sim Poisson(w_i\lambda_i)$

$$f_{S_i}(s_i; \lambda_i) = P(S_i = s_i) = P(N_i = s_i w_i) = \exp\{-w_i\lambda_i\}\frac{(w_i\lambda_i)^{s_i w_i}}{(s_i w_i)!}$$

$E(S_i) = E(N_i)/w_i = \lambda_i = b'(\theta_i)$

$Var(S_i) = Var(N_i/w_i) = \lambda_i / w_i = b''(\theta_i)\frac{\phi}{w_i}$

$Där\ \phi = 1 \quad och \quad v(\lambda_i) = \lambda_i$

3.3.2. Medelskadan

När man skattar medelskadan utgår man från att de enskilda skadebeloppen är gammafördelade. Eftersom variansfunktionen entydigt bestämmer vilken EDM man har är det en bra metod att undersöka om variansfunktionen är rimlig, då man bestämmer fördelningen. För gammafördelningens variansfunktion är väntevärdet proportionellt mot standardavvikelsen, vilket är ett rimligt antagande. [4] Enligt Ohlsson (2003) sats 2.2 är att alla EDM är reproducerbara vilket innebär att om man slår ihop två tariffceller där medelvärdet inte skiljer sig mycket åt så stannar man inom samma EDM-familj. Även detta är ett rimligt antagande för skadefördelningen.

Medelskadan, $M_i = X_i / n_i$ (19)

X_i kostnad för skador i tariffcell i med duration w_i

$$X_i \sim Gamma(w_i a_i, b_i)$$

$$f_{M_i}(m_i) = w_i f_{X_i}(w_i m_i) = \frac{(w_i b_i)^{w_i a_i}}{\Gamma(w_i a_i)} y^{w_i a_i - 1} e^{b_i w_i m_i}$$

$$E(M_i) = a_i / b_i$$

$$Var(M_i) = (w_i a_i)/(w_i b_i)^2 = \left(\frac{a_i}{b_i}\right) \Big/ w_i a_i = \left(\frac{a_i}{b_i}\right) \frac{w_i}{\phi}$$

$$där \quad \phi = 1/a_i \quad och \quad v(\mu_i) = \mu_i^2$$

3.3.3. Skattningar i GLM

Eftersom både poisson- och gamma-fördelningen tillhör EDM-familjen kommer skattningarna att beskrivas utifrån frekvensfunktionen (17). I GLM skattas parametrarna genom att maximera log-likelihoodfunktionen. För att göra detta deriveras log-likelihoodfunktionen med avseende på parametrarna (se appendix). Man kan se att skattningarna av relationstalen inte beror på ϕ.

log - likelihood funktionen (20)

$$l(\grave{\mathbf{e}}, \phi, \mathbf{y}) = \frac{1}{\phi} \sum_i w_i (y_i \theta_i - b(\theta_i)) + \sum_i c(y_i, \phi, w_i)$$

$$\frac{\partial l}{\partial \beta_j} = 0 \quad \text{ger ML - ekvationerna :} \quad \sum_i w_i \frac{(y_i - \mu_i) x_{ij}}{v(\mu_i) g'(\mu_i)} = 0$$

När varianserna skattas används att ML-skattningarna är asymptotiskt normalfördelade och asymptotiskt väntevärdesriktiga.

$$\hat{\mathbf{a}} \approx N\left(\hat{\mathbf{a}}, \mathbf{I}^{-1}\right) \qquad (21)$$

$$\mathbf{I} = -E(\mathbf{H}) = -E\left(\frac{\partial^2 l}{\partial\beta_j \partial\beta_k}\right) = -\sum_i x_{ij}\left[\frac{w_i}{\phi v(\mu_i) g'(\mu_i)^2}\left(1 + \frac{y_i g''(\mu_i)}{g'(\mu_i)} - \frac{\mu_i v'(\mu_i)}{v(\mu_i)}\right)\right] x_{ik}$$

I (21) ser man att variansskattningen beror på parametern ϕ. För skadefrekvensen behövs dock ingen skattning av ϕ eftersom den är ett. ϕ kan skattas på tre olika sätt, genom ML-skattning, Pearsons och Devians. I det här arbetet används Pearsons skattning.

Pearsons χ^2

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi}\sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \text{ som är approximativt } \chi^2 \text{ (n - r) - fördelad.}$$

$$\Rightarrow \hat{\phi} = \frac{\phi\chi^2}{n\text{-}r} = \frac{1}{n-r}\sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \qquad (22)$$

Med hjälp av de skattade parametrarna kan konfidensintervall bildas. (23) visar ett 95-procentigt konfidensintervall för medelskadans och skadefrekvensens betavariabler. När man vill ha konfidensintervall för gammavariablerna använder man sambandet som visas i (23).

95% - igt KI för β : $\quad \hat{\beta} \pm 1{,}96\, Se(\hat{\beta})$

95% - igt KI för γ : $\quad (e^{\hat{\beta} - 1{,}96\, Se(\hat{\beta})},\ e^{\hat{\beta} + 1{,}96\, Se(\hat{\beta})})$ $\qquad (23)$

När varianser och konfidensintervall för riskpremien beräknas bortser man från att skattningarna av medelskadan och skadefrekvensen parametrar är beroende. Att detta kan göras motiveras av att likelihoofunktionen kan faktoriseras på det sätt som beskrivs i avsnitt 3.3.1. Variansen för riskpremien kan då beräknas enligt (24).

$$Y_i = S_i \times M_i = \frac{\text{antal skador}}{\text{duration}} \times \frac{\text{summan av skadekostnaden}}{\text{antal skador}}$$

$$\hat{\beta}^{Y}{}_{ijk} = \hat{\beta}^{S}{}_{ijk} + \hat{\beta}^{M}{}_{ijk}$$

$$Var(\hat{\beta}^{Y}{}_{ijk}) = Var(\hat{\beta}^{S}{}_{ijk}) + Var(\hat{\beta}^{M}{}_{ijk}) \qquad (24)$$

3.4. Tweedie-modellen

De EDM som har variansfunktionen $v(\mu)=\mu^p$ kallas för Tweedie-modeller. Dessa modeller tillhör de generaliserade linjära modellerna. Det som är karaktäristiskt med dessa är att de är skalinvarianta. Med det menas att en stokastisk variabel, som tillhör EDM, som multipliceras med en konstant, tillhör fortfarande samma EDM familj. Tweedie-modellen där 1<p<2 är den sammansatta Poissonfördelningen. I denna sammansatta Poissonfördelning är skadeantalet Poissonfördelat och skadekostnaden Gammafördelad, vilket gör den lämplig till att användas direkt på riskpremien. Eftersom sammansatt Poissonfördelningen används i försäkringssammanhang, är det intressant att koppla den till EDM så att generaliserade linjära modeller kan användas.

Tabell 1.

Några olika fördelningar för olika p	
p=0	Normalfördelning
p=1	Poissonfördelning
1<p<2	Sammansatt Poissonförd.
p=2	Gammafördelning
p=3	Invers normal förd.

EDM-funktionen för Tweedie-modellen med 1<p<2.

$$f(y,\theta_0,\phi)=\exp\left\{\frac{y\theta_0-b(\theta_0)}{1/\lambda w_i}+c(y,\lambda,p)\right\}$$

$$där\quad \theta_0=\theta\lambda^{p-1},\quad b(\theta)=\frac{-1}{p-2}\left(-(p-1)\theta\right)^{p-2/p-1},\quad b'(\theta)=\left(\theta(1-p)\right)^{1/1-p}\quad \text{och}$$

$$c(y,\phi,p)=\begin{cases}\log\left(\sum_{n=1}^{\infty}\left(\frac{(w\lambda)^{\frac{1}{p-1}}b\left(\frac{-1}{y}\right)}{\Gamma(n(2-p)/(p-1))n!\,y}\right)\right) & y>0\\ 0 \qquad \text{annars}\end{cases}$$

Denna fördelning får man genom följande:

Låt Z_i vara skadekostnaderna och $N(w)$ antal skador. Dessa är oberoende av varandra.

$$Z_i\sim Gamma(-\theta,-\alpha)\quad \text{med täthetsfunktion}\ \frac{(-\theta)^{-\alpha}}{\Gamma(-\alpha)}z^{-\alpha-1}e^{\theta z}\quad \text{och}\quad N(w)\sim Poisson(wm)$$

$$Z(w)=\sum_{i}^{N(w)}Z_i\qquad \text{och}\qquad Y(w)=\tfrac{1}{w}\sum_{i}^{N(w)}Z_i$$

$$\text{Parametrana fås genom;}\ \lambda>0\ \ m=\lambda^{\frac{1}{p-1}}b(\theta),\quad \alpha=\frac{p-2}{p-1}\qquad (25)$$

[3]

I det här arbetet har p=1,5 använts.

3.5. SAS proc genmod

I det här arbetet har SAS använts. I SAS finns det färdiga procedurer, proc genmod, för beräkning av relationstal, varianser och konfidensintervall för generaliserade linjära modeller. Att beräkna skattningar av relationstalen i Jungs metod är samma sak som att använda sig av proc genmod med Poissonantagande. (Se appendix för bevis.)

Proc genmod anpassar data till en generaliserad linjär modell med hjälp av maximum-likelihood metoden. För att maximera likelihoodfunktionen använder SAS Newton-Raphson algoritmen. Genom länkfunktionen, link, anger man vilken typ av modell man vill ha, i vårt fall där modellen är multiplikativ anger man logaritmfunktionen. Responsen antas ha en EDM-fördelning. Man får ange den typ av EDM-fördelning man vill anpassa data till. Så här kan man skriva i SAS när man vill skatta betaparametrarna för medelskadan med hjälp av proc genmod. [11]

```
proc genmod data=medelskada;
          class &var;
          model medelskada=&var /dist=gamma
          link=log
          pscale;
          weight skador;
          ods output ParameterEstimates=r_medelskada;
          run;
```

I SAS kan man skatta parametern ϕ genom maximum-likelihood, Pearsons chi-två och Devians. Om man inte anger någon metod så används ML- skattningen då man har gammafördelning, men för poissonfördelningen fixeras ϕ vid ett. För att få så kallad överspridning då man har poissonfördelning anger man vilken metod ϕ ska skattas med. (23) I det här arbetet har jag använt mig av Pearsons chi-två vid skattning av ϕ i gammafördelningen, och ingen överspridning i poissonfördelningen.

Då man har överspridning i Poissonfördelningen ser variansfunktionen ut som förljande:

$$V(\lambda_i) = \phi\lambda_i \quad där\ \phi > 1$$

Om $\phi < 1$ så kallas det för underspridning [11] (26)

Det man ska tänka på när man skattar relationstalen för medelskadan i SAS är att det bör göras på icke-aggregerade data för att inte missa variationen inom cellerna. När det gäller skadefrekvensen kan man arbeta med aggregerade data när man inte antar någon överspridning.

När det gäller Tweedie-modellen finns det ingen färdig procedur i SAS som kan användas. Man får istället använda sig av programmeringssteg i proc genmod, där man själv talar om vad deviansen och variansen ska vara. I (27) har jag beräknat deviansen för Tweedie-modellen.

Beräkning av deviansen för Tweedie - modellen

$$Devians, D = 2\left[l(\mathbf{y}) - l(\hat{\mu})\right]$$

$$\log - likelihood, l(\grave{e}; \phi, \mathbf{y}) = \frac{1}{\phi}\sum_i w_i\left(y_i\theta_i - b(\theta_i)\right) + \sum_i c(y_i, w_i, \phi)$$

$$\mu_i = b'(\theta_i) \Rightarrow \theta_i = h(\mu_i) \Rightarrow$$

$$D = \frac{2}{\phi}\sum_i w_i\left(y_i h(y_i) - b(h(y_i)) - y_i h(\mu_i) - b(h(\mu_i))\right)$$

i Tweedie :

$$b(\theta_i) = -\frac{1}{p-2}\left(-(p-1)\theta_i\right)^{\frac{p-2}{p-1}}$$

$$b'(\theta_i) = \left((1-p)\theta_i\right)^{\frac{-1}{p-1}}$$

$$h(\mu_i) = \frac{\mu_i^{-(p-1)}}{1-p}$$

$$\Rightarrow D = \frac{2}{\phi}\sum_i w_i\left(\frac{y_i^{(2-p)}}{(2-p)(1-p)} - \frac{y_i\mu_1^{1-p}}{1-p} + \frac{\mu_i^{2-p}}{2-p}\right) \qquad (27)$$

3.6. Fördelar och nackdelar med metoderna.

Det finns både för och nackdelar med de tre olika metoderna som skattar relationstalen i en multiplikativ tariff. En fördel, med Standard-GLM och Tweedie-modellen, är att de utgår från fördelningar vilket gör det möjligt att beräkna varianser och konfidensintervall. Det är även en fördel att man vet under vilka förutsättningar som metoderna gäller eftersom det gör det möjligt att avgöra när de bör användas. I Standard-GLM analyseras skadefrekvensen och medelskadan separat vilket gör det möjligt att ha olika klassindelningar i frekvens- respektive medelskadeanalysen. Detta gör det enklare att finna signifikanta klassindelningar. Att medelskadan och skadefrekvensen antas vara oberoende vid beräkning av riskpremiens variansskattning kan dock vara en nackdel. En annan fördel med GLM är att det finns färdiga procedurer i SAS som man kan använda sig av.

Det kan både finnas för- och nackdelar med att Jungs metod är fördelningsfri. Det som är negativt är att man inte vet när man bör eller inte bör använda metoden. Men å andra sidan är det en fördel att den alltid ger väntevärdesriktiga marginalsummor oberoende av fördelning. När det gäller LF-Wasa-metoden, som används vid variansskattning av Jungs relationstal, så är det en nackdel att metoden inte är motiverad då man har fler än ett premieargument.

4. Verkliga data

Beräkningarna har gjorts på skadestatistik från Trygg-Hansas bussförsäkringar mellan åren 1998 och 2003. Till att börja med undersöktes vilka variabler som var möjliga att ha med i modell, alltså variabler som det fanns uppgifter om i de flesta av försäkringarna. Utav dessa användes fem variabler. Därefter gjordes klassindelning och de klasser som hade störst duration fick vara basklasser. Sedan beräknades relationstal, varianser samt tillhörande konfidensintervall enligt de tre metoderna.

4.1. Resultat från verkliga data

För de skattade relationstalen gav metoderna likartade resultat. När det gällde de skattade standardavvikelserna så var Tweedies skattningarna något högre än Standard-GLM och Jungs skattningar för i stort sett alla klasser. (Se tabell 2.)

I tabell 3 ser man att även om relationstalen skiljer sig något åt mellan de olika metodernas skattningar så är skattningarna inom varandras konfidensintervall.

En nackdel med att undersöka verkliga data är att man inte har något facit att jämföra med, vilket gör det svårt att avgöra vilken av metoderna som är bättre än de andra..

Tabell 2.

		Relationstal					standaravvikelse				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
Parameter	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
variabel 1	1	1,29	0,78	1,00	0,98	1,01	0,04	0,10	0,13	0,12	0,17
variabel 1	2	1,00	1,00	1,00	1,00	1,00					
variabel 1	3	0,78	1,28	0,99	0,96	1,00	0,02	0,14	0,12	0,11	0,13
variabel 2	1	0,69	1,01	0,69	0,64	0,71	0,03	0,19	0,14	0,12	0,17
variabel 2	2	1,08	0,93	1,00	0,96	0,98	0,03	0,13	0,14	0,12	0,15
variabel 2	3	1,00	1,00	1,00	1,00	1,00					
variabel 2	4	1,58	1,34	2,11	1,98	2,06	0,05	0,17	0,27	0,27	0,29
variabel 3	1	1,00	1,00	1,00	1,00	1,00					
variabel 3	2	0,78	0,99	0,77	0,77	0,75	0,02	0,10	0,08	0,08	0,09
variabel 3	3	0,44	0,94	0,41	0,40	0,40	0,01	0,12	0,06	0,05	0,06
variabel 4	1	1,00	1,00	1,00	1,00	1,00					
variabel 4	2	0,97	1,30	1,26	1,30	1,31	0,02	0,13	0,13	0,15	0,15
variabel 5	1	1,00	1,00	1,00	1,00	1,00					
variabel 5	2	0,71	0,92	0,66	0,66	0,68	0,02	0,10	0,07	0,08	0,08
variabel 5	3	0,97	0,80	0,78	0,76	0,79	0,03	0,10	0,10	0,08	0,12

Tabell 3.

		95%-igt KI									
		Standard-GLM						Jung		Tweedie	
Parameter	Klass	frekvens		medelskada		riskpremie		riskpremie		riskpremie	
variabel 1	1	1,21	1,37	0,61	1,00	0,78	1,30	0,78	1,23	0,73	1,41
variabel 1	2										
variabel 1	3	0,74	0,82	1,03	1,59	0,79	1,25	0,76	1,20	0,78	1,28
variabel 2	1	0,63	0,75	0,70	1,45	0,48	1,01	0,44	0,92	0,46	1,10
variabel 2	2	1,02	1,15	0,71	1,21	0,77	1,32	0,75	1,23	0,73	1,33
variabel 2	3										
variabel 2	4	1,49	1,67	1,05	1,70	1,65	2,71	1,52	2,57	1,56	2,71
variabel 3	1										
variabel 3	2	0,75	0,82	0,82	1,21	0,64	0,95	0,62	0,94	0,59	0,95
variabel 3	3	0,41	0,46	0,73	1,21	0,32	0,53	0,31	0,51	0,30	0,52
variabel 4	1										
variabel 4	2	0,93	1,02	1,07	1,58	1,03	1,54	1,05	1,62	1,04	1,63
variabel 5	1										
variabel 5	2	0,68	0,75	0,75	1,14	0,53	0,81	0,53	0,83	0,54	0,87
variabel 5	3	0,92	1,03	0,63	1,01	0,61	0,99	0,62	0,94	0,59	1,06

5. Simulerade data

Nästa steg, i jämförelsen mellan Jungs metod, Standard-GLM och Tweedie-modellen var att simulera fram data. Här följer en beskrivning av hur det första simuleringsfallet gick till. För de övriga simuleringsfallen beskrivs endast vilka förändringar som gjorts. (Se appendix för tabell över de olika simuleringsfallen.)

5.1. Simuleringsfall 1; Standardmodellen

1. Först bestämdes hur modellen skulle se ut, det vill säga om den skulle vara fullständigt multiplikativ samt hur många premieargument, klasser och observationer per cell den skulle ha. Relationstalen för skadefrekvensen och medelskadan fixerades, dels för att använda dem vid simuleringen, dels som ”facit” att jämföra de skattade parametrarna med.

   I det första simuleringsfallet hade väntevärdena för nyckeltalen (riskpremien, skadefrekvensen, skadebeloppet) en fullständigt multiplikativ modell.

$$E(N_{ijkl}) = \lambda_{ijk\ldots} = \tilde{a}_0^N \tilde{a}_{1i}^N \tilde{a}_{2j}^N \tilde{a}_{3k}^N \tilde{a}_{4l}^N = \exp(\hat{a}_0^N + \hat{a}_{1i}^N + \hat{a}_{2j}^N + \hat{a}_{3k}^N + \hat{a}_{4l}^N) \qquad (28)$$

N_{ijkl} = skadefrekvensen i tariffcell $ijkl$

$\tilde{a}_0$ = bascell

$\tilde{a}_{1i}$ = relationstal för de olika klasserna för premieargument 1. $i = 1,\ldots,4$

$\tilde{a}_{2j}$ = relationstal för de olika klasserna för premieargument 2. $j = 1,\ldots,5$

$\tilde{a}_{3k}$ = relationstal för de olika klasserna för premieargument 3. $k = 1,\ldots,5$

$\tilde{a}_{4l}$ = relationstal för de olika klasserna för premieargument 4. $l = 1,\ldots,4$

För medelskada på motsvarande sätt.

2. Nästa steg var att bestämma vilka fördelningar som antalet skador och skadekostnader skulle simuleras ifrån. I det första fallet simulerades antalet skador från poissonfördelningen och skadekostnaden från gammafördelningen.

   Först simulerades antalet skador. För varje *ijkl* (tariffcell) simulerades antalet skador, N_{ijkl}, w_{ijkl} antal gånger med väntevärdet från (28). Datasättet hade nu en observation per försäkringsår och w_{ijkl} försäkringsår (duration) per tariffcell.

   Ett alternativt sätt vid poissonfördelningen är att simulera antalet skador i varje tariffcell med väntevärdet från (28) multiplicerat med w_{ijkl}. Detta går att göra eftersom summan av w stycken poissonfördelade stokastiska variabler med väntevärde λ är poissonfördelade med väntevärde $w\lambda$.

   För varje skada simulerades sedan skadebeloppet fram. För att få fram gammafördelningens parametrar (a, b) beräknades medelvärdet och den skattade standardavvikelsen från den riktiga skadestatistiken, bussdatasättet. Detta gjordes för att få en rimlig skadefördelning och inte för att få en fördelning som exakt motsvarade det riktiga bussdatasättet. Parametrarna skattades på följande sätt.

$\bar{x}$ = medelvärdet av skadekostnaden från bussdata

s^2 = den skattade variansen från bussdata

$\mu = ab$ = gammafördelningens väntevärde

$\sigma^2 = ab^2$ = gammafördelningens varians

Vilket ger skattningarna

$\hat{b} = s^2 / \bar{x}$ och $\hat{a} = \bar{x} / \hat{b}$

Skattnig av a och b i varje tariffcell:

$\hat{a}_{ijk\ldots} = \hat{a}$

$\hat{b}_{ijk\ldots} = \hat{b}\gamma_0^M \gamma_{1i}^M \gamma_{2j}^M \gamma_{3k}^M \cdots$

där $\gamma_0^M \gamma_{1i}^M \gamma_{2j}^M \gamma_{3k}^M \ldots$ fixerats från början.

Nu får man en gammafördelning med följande struktur på väntevärde och varians.

$$\hat{\mu}_{ijk\ldots} = \hat{a}\hat{b}_{ijk\ldots} = \hat{a}\hat{b}\gamma_0^M \gamma_{1i}^M \gamma_{2j}^M \gamma_{3k}^M \cdots$$

$$\hat{\sigma}^2{}_{ijk\ldots} = \hat{a}\hat{b}^2{}_{ijk\ldots} = \hat{\mu}^2{}_{ijk\ldots} / \hat{a} \qquad (29)$$

3. Relationstal, varianser och konfidensintervall beräknades därefter utifrån Standard-GLM, Jungs metod och Tweedie-modellen.

4. Simuleringen upprepades k antal gånger. Av varje parameter, varians och konfidensintervall fanns det nu k stycken skattningar. För att få ett hanterbart datamaterial, som det gick att utföra beräkningar på, las efter varje simulering alla skattningar på en enda rad. Detta gav ett datasätt med k stycken rader, med alla skattningar av samma parameter i samma kolumn.

5.2. Simuleringsfall 2; Felspecificerad skadefördelning

Här har samma simuleringsmönster följts som för fall ett. Det som skiljer simuleringarna åt är att i punkt två har fördelningen, som skadekostnaden simulerats ifrån, ändrats från gammafördelning till lognormalfördelning. Även denna gång användes bussdatasättets medelvärde och empiriska varians vid beräkningen av fördelningens parametrar. Parametrarna skattades på följande sätt.

lognormalfördelningen(a, b^2)

med väntevärdet, $\mu = e^{\left\{a + \frac{b^2}{2}\right\}}$ och variansen, $\sigma^2 = e^{2a}\left(e^{2b^2} - e^{b^2}\right)$

medelvärdet av skadekostnaden i bussdata $= \bar{x}$

stickprovsvariansen i bussdata $= s^2$

$\hat{\mu} = \bar{x}$

$\hat{\sigma}^2 = s^2$

Skattar parametrarna a, b i lognormalfördelningen

$$e^{\hat{a}} = \bar{x}e^{\frac{-b^2}{2}}$$

$$e^{\hat{b}^2} = \frac{s^2}{\bar{x}^2} + 1 = \frac{s^2 + \bar{x}^2}{\bar{x}^2}$$

$$e^{\hat{a}} = \bar{x}\left(\frac{s^2 + \bar{x}^2}{\bar{x}^2}\right)^{\frac{-1}{2}} = \frac{\bar{x}^2}{\sqrt{s^2 + \bar{x}^2}}$$

Anpassar parametrarna så att väntevärde och varians blir olika i de olika tariffcellerna

$$e^{\hat{a}_{ijk\ldots}} = e^{\hat{a}}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M\cdots = \frac{\bar{x}^2}{\sqrt{s^2 + \bar{x}^2}}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M\cdots$$

$$\hat{a}_{ijk\ldots} = \log\left(\frac{\bar{x}^2}{\sqrt{s^2 + \bar{x}^2}}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M\cdots\right) = \log\left(\frac{\bar{x}^2}{\sqrt{s^2 + \bar{x}^2}}\right) + \beta_0^M + \beta_{1i}^M + \beta_{2j}^M + \beta_{3k}^M + \cdots$$

$$\hat{b}^2 = \log\left(\frac{s^2 + \bar{x}^2}{\bar{x}^2}\right) \qquad (30)$$

Väntevärdet och variansen i lognormalfördelningen blir

$$\hat{\mu}_{ijk\ldots} = e^{\left\{\hat{a}_{ijk\ldots}+\frac{\hat{b}_{ijk\ldots}^2}{2}\right\}} = e^{\hat{a}}e^{\frac{\hat{b}}{2}}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M\cdots = \bar{x}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M\cdots$$

$$\hat{\sigma}^2{}_{ijk\ldots} = e^{2\hat{a}_{ijk\ldots}}\left(e^{2\hat{b}_{ijk\ldots}^2} - e^{\hat{b}_{ijk\ldots}^2}\right) = (\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M)^2 e^{2\hat{a}}\left(e^{2\hat{b}_{ijk\ldots}^2} - e^{\hat{b}_{ijk\ldots}^2}\right) =$$

$$= (\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M)^2 e^{2\hat{a}+\hat{b}^2}\left(e^{\hat{b}^2}-1\right) = \hat{\mu}^2{}_{ijk\ldots}\left(e^{\hat{b}^2}-1\right)$$

Vilket ger samma variansfunktionen som vi hade i simuleringsfall 1, $v(\mu) = \mu^2$. (31)

5.3. Simuleringsfall 3; Felspecificerad skadefördelning och variansfunktion

Det som undersöktes i simuleringsfall tre, var hur bra metodernas skattningar blev när skadekostnaden simulerats från en fördelning med en annan variansfunktion än $v(\mu)=\mu^2 k$. Här skattades parametrarna i lognormalfördelningen på följande sätt.

Anpassar parametrarna i (31) så att väntevärde och varians blir olika i de olika tariffcellerna

$$\hat{b}^2_{ijk\ldots} = \log\left(\frac{s^2+\bar{x}^2}{\bar{x}^2}(\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M)^2\right) \quad \hat{a} = \log\left(\frac{\bar{x}^2}{\sqrt{s^2+\bar{x}^2}}\right)$$

$$\mu_{ijk\ldots} = e^{\left\{\hat{a}_{ijk\ldots}+\frac{b^2_{ijk\ldots}}{2}\right\}} = e^{\hat{a}}e^{\frac{b^2_{ijk\ldots}}{2}} = \frac{\bar{x}^2}{\sqrt{s^2+\bar{x}^2}}\sqrt{\frac{s^2+\bar{x}^2}{\bar{x}^2}(\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M)^2} = \bar{x}\gamma_0^M\gamma_{1i}^M\gamma_{2j}^M\gamma_{3k}^M$$

Vilket ger följande varians

$$\sigma^2_{ijk\ldots} = e^{2a_{ijk\ldots}}\left(e^{2b_{ijk\ldots}^2} - e^{b_{ijk\ldots}^2}\right) = e^{2a}\left(e^{2b_{ijk\ldots}^2} - e^{b_{ijk\ldots}^2}\right) = \mu^2_{ijk\ldots}(e^{b^2_{ijk\ldots}}-1) \quad (32)$$

5.4. Simuleringsfall 4; Avvikelse från multiplikativitet.

I det fjärde simuleringsfallet undersöktes hur bra de tre metodernas skattningar blev när modellen som simuleringarna gjordes från inte längre var fullständigt multiplikativ. De fixerade relationstalen kunde inte längre användas som ”facit”. Istället undersöktes hur bra skattningarna stämde i de olika tariffcellerna. Först fixerades relationstalen för skadefrekvensen och medelskadan, därefter beräknades väntevärdena i de olika tariffcellerna enligt ekvation (33). Därefter simulerades skadefrekvensen och medelskadan på samma sätt som i simuleringsfall ett från och med punkt två.

En modell som inte har fullständigt multiplikativ struktur

$$\gamma_{ijkl} = 0{,}75\times(\gamma_0\gamma_{1i}\gamma_{2j}\gamma_{3k}\gamma_{4l}) + 0{,}25\times k$$

Där k är ett viktade medelvärdet av tariffcellernas fixerade väntevärden. (33)

Ingen av metoderna kan skatta relationstalen rätt eftersom de utgår från en multiplikativ modell. För att få ett mått på hur bra metodernas parameterskattningar var viktades felen som metoderna gav i de olika tariffcellerna med durationen. Detta gjordes eftersom det är bättre att modellerna skattar fel när durationen är liten jämfört med då durationen är stor.

Metodernas skattningar undersöktes genom att kvadratsumman av skattningarnas fel jämfördes.

$$\sum_{ijkl} w_{ijkl}\left(Y_{ijkl} - \hat{\gamma}_0\hat{\gamma}_{1i}\hat{\gamma}_{2j}\hat{\gamma}_{3k}\hat{\gamma}_{4l}\right)^2 \qquad (34)$$

Summering över alla simuleringar och celler.

LF-Wasa metoden har inte något sätt att skatta varianserna för de skattade parametrarna i tariffcellerna och därför har inte variansskattningarna jämförts här.

6. Hur kan man undersöka metoderna mot varandra utifrån det simulerade materialet?

6.1. Vilken av metoderna ger ”mest” väntevärdesriktiga skattningar?

Först jämfördes de skattade parametervärdena från Jungs metod, Standard-GLM och Tweedie-modellen med de ”riktiga” parametervärdena, de som fixerats från början. Eftersom skattningarna såg symmetriska ut i histogram jämfördes medelvärdena av skattningarna med de riktiga värdena. Här såg man om det fanns något systematiskt fel dvs om parameterskattningarna var väntevärdesriktiga, alltså om villkoret (35) var uppfyllt. (β_{11} har tagits som exempel, samma beräkningar har utförts för alla variabler.)

$$\mathrm{E}\left(\frac{1}{k}\sum_s \hat{\beta}^{s}{}_{11}\right) = \beta_{11} \qquad \textit{skattning från simulering}, s = 1\ldots k \qquad (35)$$

För att undersöka om de skattade variablerna var väntevärdesriktiga bilades konfidensintervall för medelvärdet av parameterskattningarna. Medelvärdet av variablerna är, enligt centrala gränsvärdes satsen (se appendix), approximativt Normalfördelat.

$\frac{1}{k}\sum_s \hat{\beta}^{s}{}_{11}$ är approximativt $N(\mu_{11}, \sigma^2{}_{11}/k)$ för stora k.

$E(\beta_{11}) = \mu_{11}$

$Var(\beta_{11}) = \sigma^2{}_{11}$

k = antal skattningar (36)

Eftersom det var många skattningar av varje parameter kunde även variansen skattas med stickprovsvariansen, S^2. Skattningarna ansågs vara väntevärdesriktiga om testet i (37) inte kunde hitta att skattningarna inte var väntevärdesriktiga, alltså om H_0 inte kunde förkastas.

$$\overline{\beta}_{11} = \frac{1}{k}\sum_s \hat{\beta}^s{}_{11}$$

$$S^2{}_{11} = \frac{1}{k-1}\sum_s \left(\hat{\beta}^s{}_{11} - \overline{\beta}_{11}\right)^2 \qquad \text{Summering över simuleringarna, } s = 1\ldots k$$

$$H_0 : \overline{\beta}_{11} = \beta_{11} \qquad H_1 : \overline{\beta}_{11} \neq \beta_{11}$$

β_{11} = riktiga parametervärdet

$$KI : \overline{\beta}_{11} \pm 1{,}96 \times \sqrt{S^2{}_{11}/k} \quad \text{med konfidensgrad 95\%} \qquad (37)$$

6.2. Vilken metod ger minst varians/standardavvikelse?

Skattningarnas empiriska varians (stickprovsvariansen), som är en skattning av den sanna variansen, undersöktes för att se vilken metod som gav minst varians.

$$S^2{}_{11} = \frac{1}{k-1}\sum_s \left(\hat{\beta}^s{}_{11} - \overline{\beta}_{11}\right)^2 \qquad \text{Summering över simuleringarna, } s = 1\ldots k$$

6.3. Beräknar metoderna varianser och konfidensintervall korrekt?

Den empiriska standardavvikelsen jämfördes med kvadratroten ur medelvärdet av de skattade varianserna för att se om de skattade varianserna beräknats korrekt.

$$S_{11} \text{ jämfördes med } \mathrm{Se}(\overline{\hat{\beta}}_{11}) = \sqrt{\frac{1}{k}\sum_s \mathrm{var}(\hat{\beta}^s{}_{11})} \quad \text{simulering, } s = 1\ldots k \qquad (38)$$

För att undersöka om konfidensintervallen beräknats korrekt beräknades konfidensintervall för de skattade parametrarna med hjälp av de skattade varianserna, alltså de varianser som beräknats enligt metodernas formler med de skattade parametrarna insatta. Därefter undersöktes om de ”riktiga” parametervärdena låg inom intervallet. För att hålla reda på när villkoret var uppfyllt skapades en bernoulli variabel som var ett om vilkoret var uppfyllt och noll annars.

μ_{ij} = det sanna parametervärdet

$\hat{\mu}_{ij}$ = skattning av parametern enligt Jungs eller GLM – metoden

$$Z_{ij}(s) = \begin{cases} 1 & \text{, om } \mu_{ij} \text{ ligger inom det x – procentiga konfidensintervallet för } \hat{\mu}_{ij}. \\ 0 & \text{, annars} \end{cases}$$

Simulering, $s = 1 \ldots k$

Z_{ij} är Bernoullifördelad$(x/100)$ (39)

Efter att *k* stycken simuleringar gjorts, och från dessa hade *k* stycken skattningar av varje parameter beräknats, så hade man *k* stycken Bernoullivariabler. Summan av dessa är Binomialfördelade. Under hypotesen att metodens 95 procentiga konfidensintervall beräknats korrekt blev:

$$\sum_s Z_{ij}(s) \quad \text{binomialfördelad}(k, 0.95) \quad \text{simulering } s = 1 \ldots k$$

För att undersöka om konfidensintervallen beräknats korrekt, beräknades först medelvärdet av Z-variablerna och sedan konfidensintervall för dessa. Medelvärdet visade hur ofta de riktiga variablerna låg inom metodernas skattade konfidensintervall.

$$\overline{Z}_{ij} = \frac{1}{k} \sum_s Z_{ij} \qquad simulering \quad s = 1 \ldots k$$

$$\overline{Z}_{ij} \pm 1{,}96 \times \frac{\sqrt{k \times 0{,}95 \times (1 - 0{,}95)}}{k} \quad för\ 95\% - igt\ konfiden\sin tervall \qquad (40)$$

6.4. Hur ofta lyckas metoderna identifiera skillnader mellan klasser?

Hur ofta förkastar metoderna en icke sann hypotes? Ett ensidigt konfidensintervall för de skattade variablerna beräknades för att undersöka hur ofta metoderna förkastade att det inte fanns en skillnad mellan variabeln och bascellen. Därefter skapades en bernoullivariabel som höll reda på hur ofta hypotesen förkastats.

$H_0 : \beta_{ij} = \beta_{ib} \quad j \neq b$

$H_1 : \beta_{ij} < \beta_{ib}$ där $\hat{\beta}_{ib}$ är bascell $\Leftrightarrow \beta_{ij} < 0$

$95\% - \text{igt KI} : \left(\hat{\beta}_{ij}, \hat{\beta}_{ij} + 1{,}6449 Se(\hat{\beta}_{ij})\right)$

$$z_j = \begin{cases} 1 & \text{om 0 finns i konfidensintervallet} \\ 0 & \text{annars} \end{cases} \quad \text{simulering, } j = 1 \ldots k \qquad (41)$$

6.5. Vad kan undersökas genom simuleringarna?

Vid simulering av data finns det möjligheter att undersöka hur bra metoderna fungerar under olika omständigheter. I de olika simuleringsfallen undersöktes olika aspekter som kunde tänkas påverka metodernas skattningar. I simuleringarna användes fyra premieargument, med fyra klasser i två av dem och fem i de andra två, vilket innebar 400 tariffceller. Varje simulerat datasätt motsvarade ett försäkringsår och i varje försäkringsår fanns det 334 139 stycken ettåriga försäkringar. Antal skador som inträffade bland dessa var ungefär 30 000 stycken. Samma fixerade relationstal och lika stor duration i de olika tariffcellerna användes i de olika simuleringsfallen. Däremot hade premieargumenten varierande mönster vad gäller relationstal och duration.

6.5.1. Durationens inverkan

Durationens påverkan, på hur bra metodernas skattningar blev, undersöktes. Det som främst undersöktes var om LF-Wasa: s variansberäkning påverkades av att durationen inte hade en multiplikativ struktur. För att undersöka detta hade två premieargument, *j* och *k*, exakt lika relationstal både för skadefrekvensen och för medelskadan, medan antalet försäkringsår skiljde sig åt. För premieargumentet *k* var det lika stor duration i de olika klasserna men för *j* varierade antalet. Alla premieargument, utom för *j* och *l*, hade en multiplikativ struktur när det gällde durationen. Den summerade durationen över premieargumenten *j* och *l* såg ut på följande sätt. (tabell 4)

Tabell 4

Durationen per cell i tusental, n_{jkl}				
$n_{.j.l}$	$n_{.j.1}$	$n_{.j.2}$	$n_{.j.3}$	$n_{.j.4}$
$n_{.1.l}$	50	25	15	5
$n_{.2.l}$	35	17,5	10,5	3,5
$n_{.3.l}$	25	12,5	7,5	2,5
$n_{.4.l}$	15	7,5	4,5	1,5
$n_{.5.l}$	5	15	25	50

6.5.2. Relationstalen inverkan

Även relationstalens utseende varierades för att se om någon av metoderna bättre klarade av vissa omständigheter än vad de andra gjorde. Hur ofta modellerna förkastade att två relationstal var lika, när de var lika undersöktes, enligt (41). För detta ändamål hade premieargument *i* två klasser med lika relationstal och två klasser där relationstalen endast skiljde sig lite åt från varandra.

7. Resultat av simuleringarna

7.1. Resultat av simuleringsfall 1; Standardmodellen

Först undersöktes hur nära medelväderna av parameterskattningarna var de sanna värdena. För Standard-GLM var medelvärdena av skadefrekvensens parameterskattningar väldigt nära de sanna värdena. Däremot för Standard-GLM:s skattning av medelskadan och riskpremien var skillnaden något större. När Jungs parameterskattningar av riskpremien undersöktes, såg man att skattningarna var ganska nära de riktiga parametervärdena, men att de var något sämre än Standard-GLM:s skattningar. Medelvärdena av Tweedie-metodens parameterskattningar var väldigt lika Standard-GLM: s och alltså nära de sanna värdena. (Se tabell 5) När väntevärdesriktighet undersöktes, enligt testet i (37), kunde testet inte förkasta att några av skattningarna var väntevärdes riktiga.

Tabell 5.

Simulering. 1.		medelvärdet av parameterskattningarna					Riktiga betavärdena			differens mellan r-premie och riktiga		
		Standard-GLM			Jung	Tweedi						
arg	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	Stand-GLM	Jung	Tweedie
i	1	-0,020	-0,034	-0,054	-0,055	-0,054	-0,020	-0,037	-0,057	0,003	0,002	0,003
i	2	-0,011	-0,022	-0,033	-0,033	-0,033	-0,010	-0,018	-0,028	-0,005	-0,010	-0,005
i	3	-0,001	0,001	0	0	0	0	0	0	0	0	0
i	4	0	0	0	0	0	0	0	0	0	0	0
j	1	-0,406	-0,469	-0,874	-0,875	-0,874	-0,406	-0,467	-0,873	-0,001	-0	-0,001
j	2	-0,288	-0,420	-0,708	-0,710	-0,708	-0,288	-0,416	-0,703	-0,005	-0,010	-0,005
j	3	-0,182	-0,407	-0,589	-0,589	-0,589	-0,182	-0,406	-0,588	-0,001	-0	-0,001
j	4	-0,088	-0,072	-0,159	-0,162	-0,160	-0,087	-0,069	-0,156	-0,003	-0,010	-0,004
j	5	0	0	0	0	0	0	0	0	0	0	0
k	1	-0,406	-0,464	-0,870	-0,870	-0,870	-0,406	-0,467	-0,873	0,003	0,003	0,003
k	2	-0,288	-0,415	-0,703	-0,702	-0,702	-0,288	-0,416	-0,703	0	0,001	0,001
k	3	-0,182	-0,404	-0,586	-0,585	-0,586	-0,182	-0,406	-0,588	0,002	0,003	0,002
k	4	-0,087	-0,069	-0,155	-0,156	-0,155	-0,087	-0,069	-0,156	0,001	0	0,001
k	5	0	0	0	0	0	0	0	0	0	0	0
l	1	-0,619	-0,508	-1,127	-1,126	-1,127	-0,619	-0,511	-1,130	0,003	0,004	0,003
l	2	-0,368	-0,367	-0,735	-0,734	-0,735	-0,368	-0,368	-0,735	0	0,001	0
l	3	-0,167	-0,266	-0,433	-0,433	-0,433	-0,167	-0,262	-0,429	-0,004	-0	-0,004
l	4	0	0	0	0	0	0	0	0	0	0	0

Hur bra metodernas variansskattningar var undersöktes genom att medelvärdena av de skattade varianserna jämfördes med stickprovsvarianserna. Eftersom varianserna var små undersöktes roten ur medelvärdet av variansskattningarna och stickprovsvariansen för att jämförelsen skulle bli enklare. För Standard-GLM stämde medelvärdena av de skattade varianserna bra överens med stickprovsvarianserna. Detta resultat tyder på att

det går att skatta varianserna för skadefrekvensen och medelskadan var och en för sig, trots att de är beroende, när man skatta variansen för riskpremien, vilket bekräftar resonemanget i 3.3.1. För Jungs metod stämde medelvärdet av de skattade varianserna bra överens med stickprovsvariansen för de premieargument som hade en multiplikativ struktur på durationen, *i* ock *k*. Däremot för premieargument *j* och *l* var medelvärdet av de skattade varianserna lägre än stickprovsvariansen. Detta resultat beror på att LF-Wasa metoden bortser från en del av variansen då durationen inte har en multiplikativ struktur. När Tweedie metodens skattade varianser undersöktes var medelvärdet av dem ganska lika stickprovsvarianserna. (Se tabell 6)

Tabell 6.

Simulering 1.		Roten ur medelvärdet av de skattade varianserna, m-std					S (empirisk)				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	0,020	0,091	0,093	0,100	0,092	0,021	0,091	0,092	0,101	0,092
i	2	0,016	0,070	0,072	0,078	0,071	0,016	0,072	0,074	0,078	0,073
i	3	0,014	0,062	0,063	0,069	0,062	0,014	0,062	0,063	0,068	0,063
i	4										
j	1	0,018	0,081	0,083	0,073	0,082	0,018	0,080	0,083	0,084	0,083
j	2	0,019	0,086	0,088	0,079	0,088	0,018	0,087	0,089	0,090	0,088
j	3	0,020	0,092	0,094	0,086	0,095	0,020	0,091	0,093	0,094	0,092
j	4	0,023	0,105	0,108	0,102	0,105	0,023	0,105	0,108	0,110	0,107
j	5										
k	1	0,018	0,083	0,085	0,090	0,083	0,018	0,084	0,087	0,090	0,087
k	2	0,018	0,080	0,082	0,087	0,081	0,018	0,077	0,080	0,085	0,080
k	3	0,017	0,078	0,080	0,084	0,080	0,017	0,075	0,077	0,081	0,077
k	4	0,017	0,076	0,078	0,082	0,075	0,016	0,074	0,076	0,079	0,076
k	5										
l	1	0,019	0,087	0,089	0,072	0,087	0,019	0,088	0,090	0,092	0,090
l	2	0,018	0,083	0,085	0,077	0,084	0,019	0,082	0,084	0,087	0,084
l	3	0,017	0,076	0,078	0,076	0,077	0,016	0,077	0,080	0,082	0,079
l	4										

Alla tre metoderna gav parameterskattningar som var nära de riktiga värdena, men Tweedie och Standard-GLM:s skattningarna var något bättre än Jungs (se tabell 6 & 7). Även den empiriska variansen var lika för alla metoderna, men den metod som hade lägst empirisk varians var Tweedie och den som hade högst var Jung (se tabell 6). Standard-GLM var den metod vars skattade varianser bäst överensstämde med de empiriska varianserna (se tabell 8).

Tabell 7.

Kvadratsumma av riktiga - medelvärdet av de skattade, delat med antal (10^{-7})				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
2,37	65,98	75,96	108,79	73,65

Tabell 8.

Kvadratsumman av kvoten mellan m-std och S				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
1,005	1,010	1,008	0,908	0,985

Tabell 9 nedan visar huruvida de 95 procentiga konfidensintervallen beräknats korrekt. De kursiva värdena i tabellen markerar de som inte beräknats korrekt enligt villkoret i (40). Alla tre metoderna beräknade i stort sett alla konfidensintervallen korrekt, bortsätt från där durationen inte hade en multiplikativ struktur i Jungs metod. Tweedie modellen beräknade de 95 procentiga konfidensintervallen korrekt för alla premieargument. (Se tabell 9)

Hur ofta modellerna inte förkastade att parametrarna inte skiljde sig från bascellen undersöktes enligt testet i (41). För premieargument *i* skiljde sig de riktiga parametrarna i klass ett och två endast lite åt och i klass tre och fyra var de exakt lika. Man kunde se att för klass tre förkastade inte metoderna att det inte fanns en skillnad i ungefär 95 procent av fallen, vilket stämde eftersom det var 95 procentiga konfidensintervall som jag hade utgått ifrån. För klass ett och två däremot, fanns det ju en liten skillnad mellan klasserna och där förkastade inte metoderna att det inte fanns en skillnad ofta. Även för klass tre i premieargumenten *j* och *k* förkastade inte metoderna ibland att det inte fanns en skillnad trots att det fanns en skillnad. Ju mindre skillnad mellan klasserna det var desto större risk var det att metoderna inte förkastade det inte fanns någon skillnad. Detta är dock ett rimligt resultat och alla tre metoder gav lika resultat. (Se tabell 9)

Tabell 9.

Simulering 1.		Beräknas de 95%-iga KI korrekt?					Hur ofta förkastas inte att parametern inte skiljer sig från bascellen?				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	0,946	0,948	0,947	0,956	0,951	0,726	0,911	0,868	0,859	0,860
i	2	0,950	0,944	0,947	0,956	0,941	0,827	0,913	0,883	0,885	0,878
i	3	0,945	0,946	0,959	0,951	0,948	0,943	0,954	0,947	0,953	0,943
i	4										
j	1	0,950	0,958	0,951	***0,911***	0,941	0,000	0,000	0,000	0,000	0,000
j	2	0,960	0,947	0,948	***0,912***	0,952	0,000	0,000	0,000	0,000	0,000
j	3	0,954	0,953	0,955	***0,927***	0,959	0,000	0,004	0,000	0,000	0,000
j	4	0,959	0,949	0,945	***0,923***	0,939	0,010	0,840	0,586	0,526	0,568
j	5										
k	1	0,961	0,945	0,945	0,944	0,938	0,000	0,000	0,000	0,000	0,000
k	2	0,946	0,952	0,951	0,957	0,945	0,000	0,000	0,000	0,000	0,000
k	3	0,952	0,958	***0,964***	***0,965***	0,962	0,000	0,001	0,000	0,000	0,000
k	4	0,953	0,955	0,956	0,954	0,941	0,001	0,774	0,339	0,384	0,330
k	5										
l	1	0,953	0,956	0,957	***0,877***	0,942	0,000	0,000	0,000	0,000	0,000
l	2	0,940	0,952	0,951	***0,913***	0,947	0,000	0,000	0,000	0,000	0,000
l	3	0,952	0,940	0,940	***0,930***	0,937	0,000	0,029	0,000	0,001	0,000

Att Standard-GLM metoden gav säkra skattningar i det första simuleringsfallet är inte konstigt med tanke på att jag utgått från att antalet skador var poissonfördelat och att skadebeloppet var gammafördelat.

7.2. Resultat av simuleringsfall 2; Felspecificerad skadefördelning

Standard-GLM:s skattningar av skadefrekvensens, medelskadans och riskpremiens parametrar var även här väldigt bra, medelvärdet av parameterskattningarna var nära de riktiga värdena. Samma resultat gäller även för Jungs metod och Tweedie metoden (se tabell 10). När väntevärdes riktighet undersöktes, enligt testet (37), såg man att alla skattningarna var väntevärdes riktiga, dvs att testet inte förkastade att de var väntevärdes riktiga, med ett undantag för Standard-GLM:s skattning av skadefrekvensen för premieargument *j* klass två.

Tabell 10.

Simulering 2.		medelvärdet av parameterskattningarna					riktiga betavärdena			differens mellan r-premie och riktiga		
		Standard-GLM			Jung	Tweedie						
arg.	klass	frek	m -skada	r-premie	r-premie	r-premie	frek	m-skada	r -premie	Stand-GLM	Jung	Tweedie
i	1	-0,021	-0,037	-0,058	-0,059	-0,059	-0,020	-0,037	-0,057	-0,001	-0,002	-0,002
i	2	-0,01	-0,017	-0,027	-0,028	-0,027	-0,010	-0,018	-0,028	0,001	0	0,001
i	3	-0,001	0,002	0,001	0	0,001	0	0	0	0,001	0	0,001
i	4	0	0	0	0	0	0	0	0	0	0	0
j	1	-0,406	-0,467	-0,873	-0,872	-0,873	-0,406	-0,467	-0,873	0	0,001	0
j	2	-0,286	-0,416	-0,703	-0,702	-0,702	-0,288	-0,416	-0,703	0	0,001	0,001
j	3	-0,182	-0,407	-0,589	-0,589	-0,589	-0,182	-0,406	-0,588	-0,001	-0,001	-0,001
j	4	-0,087	-0,073	-0,16	-0,16	-0,16	-0,087	-0,069	-0,156	-0,004	-0,004	-0,004
j	5	0	0	0	0	0	0	0	0	0	0	0
k	1	-0,406	-0,468	-0,874	-0,874	-0,874	-0,406	-0,467	-0,873	-0,001	-0,001	-0,001
k	2	-0,288	-0,418	-0,706	-0,706	-0,706	-0,288	-0,416	-0,703	-0,003	-0,003	-0,003
k	3	-0,182	-0,402	-0,584	-0,585	-0,584	-0,182	-0,406	-0,588	0,004	0,003	0,004
k	4	-0,087	-0,067	-0,154	-0,155	-0,154	-0,087	-0,069	-0,156	0,002	0,001	0,002
k	5	0	0	0	0	0	0	0	0	0	0	0
l	1	-0,619	-0,514	-1,133	-1,134	-1,134	-0,619	-0,511	-1,130	-0,003	-0,004	-0,004
l	2	-0,368	-0,37	-0,738	-0,74	-0,739	-0,368	-0,368	-0,735	-0,003	-0,005	-0,004
l	3	-0,167	-0,262	-0,429	-0,43	-0,429	-0,167	-0,262	-0,429	0	-0,001	0
l	4	0	0	0	0	0	0	0	0	0	0	0

Medelvärdet av de skattade varianserna skiljde sig inte mycket från stickprovsvariansen för Standard-GLM:s skattningar. Även här tyder resultatet på att det går att skatta varianserna för skadefrekvensen och medelskadan var och en för sig trots att de är beroende, vilket bekräftar resonemanget i 3.3.1. När det gällde skattningarna av varianserna för Jungs metod så var de inte lika bra, medelvärdet av dem var lägre än stickprovsvariansen i alla olika klasser. Skillnaden var dock större för de klasser där

durationen inte hade en multiplikativ struktur. För Tweedie var medelvärdena av de skattade varianserna lägre än stickprovsvarianserna i nästan alla klasser. (Se tabell 12)

Tabell 11.

Simulering 2		Roten ur medelvärdet av de skattade varianserna, m-std					S (empirisk)				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	0,020	0,084	0,086	0,094	0,084	0,020	0,083	0,086	0,099	0,088
i	2	0,016	0,065	0,066	0,075	0,065	0,016	0,068	0,070	0,080	0,072
i	3	0,014	0,057	0,058	0,067	0,057	0,014	0,057	0,058	0,068	0,059
i	4										
j	1	0,018	0,074	0,076	0,071	0,075	0,017	0,076	0,077	0,083	0,079
j	2	0,019	0,079	0,081	0,076	0,080	0,018	0,078	0,080	0,086	0,081
j	3	0,020	0,084	0,087	0,082	0,087	0,020	0,083	0,084	0,091	0,085
j	4	0,023	0,097	0,100	0,093	0,096	0,023	0,095	0,097	0,103	0,098
j	5										
k	1	0,018	0,076	0,078	0,086	0,076	0,019	0,077	0,079	0,088	0,080
k	2	0,018	0,073	0,076	0,083	0,074	0,018	0,072	0,074	0,082	0,075
k	3	0,017	0,071	0,073	0,082	0,073	0,017	0,073	0,074	0,085	0,076
k	4	0,017	0,070	0,072	0,080	0,068	0,017	0,072	0,074	0,083	0,075
k	5										
l	1	0,019	0,080	0,082	0,070	0,080	0,018	0,080	0,081	0,087	0,082
l	2	0,018	0,076	0,079	0,075	0,077	0,017	0,078	0,079	0,085	0,081
l	3	0,017	0,070	0,072	0,073	0,071	0,016	0,070	0,072	0,077	0,073
l	4										

I denna simulering var medelvärdet av parameterskattningar nära de sanna värdena för alla tre metoderna, skillnaden var dock minst för Standard-GLM och störst för Tweedie. Standard-GLM var den metod som gav lägst stickprovsvarians och det var även den metod där de skattade varianser stämde bäst överens med stickprovsvarianserna. Jung däremot var den metod som gav högst empirisk varians och även den metod där den skattade variansen sämst överensstämde med den empiriska variansen. (Se tabell 12 och 13)

Tabell 12.

Kvadratsumma av riktiga - medelvärdet av de skattade, delat med antal (10^{-7})				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
3,88	43,46	46,56	54,80	52,50

Tabell 13.

Kvadratsumman av kvoten mellan m-std och S				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
1,027	0,982	0,997	0,859	0,920

I tabell (14) nedan, som visar om de 95 procentiga konfidensintervallen beräknats korrekt, betecknar de kursiva värdena de fall där konfidensintervallen inte beräknats korrekt. Standard-GLM beräknade alla konfidensintervall korrekt med ett undantag för riskpremien (premieargument *i* klass ett). När det gällde Jungs metod såg man samma mönster som för simuleringsfall ett. För de premieargument som inte hade en multiplikativ struktur för durationen beräknades inte konfidensintervallen korrekt. För de övriga två argumenten var det endast på ett ställe som metoden inte beräknade konfidensintervallet korrekt. Tweedie, som i förra simuleringen beräknade konfidensintervallen korrekt, gav här ett mycket sämre resultat. Metoden gav en lägre konfidensgrad än vad de skulle göra. Självklart finns det ett samband med att metodens skattade standardavvikelser var för låga och därmed blev konfidensintervallen för ”smala”. (Se tabell 14)

När hur ofta modellerna lyckades identifiera skillnader mellan klasser undersöktes, kunde man se liknade resultat som för simuleringsfall ett. För premieargument *i* klass tre, där det i själva verket inte var någon skillnad, accepterade Jung och Tweedie mer sällan än vad Standard-GLM att det inte fanns en skillnad. Vilket betyder att Jung och Tweedie sa att parametern hade betydelse för ofta. När det gällde klass ett och två, där det fanns en liten skillnad mellan klasserna och bascellen, accepterade Standard-GLM oftare att parametern inte skiljde sig från bascellen än vad Tweedie och Jung gjorde. (Se tabell 14)

Tabell 14.

Simulering 2		Beräknas de 95%-iga KI korrekt?					Hur ofta förkastas inte att parameter inte skiljer sig från bascellen?				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
Arg.	Klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	0,951	0,954	0,945	***0,934***	0,939	0,712	0,895	0,841	0,795	0,814
i	2	0,952	0,941	***0,932***	0,936	***0,920***	0,852	0,921	0,897	0,867	0,870
i	3	0,944	0,954	0,948	0,947	***0,931***	0,949	0,953	0,951	0,939	0,935
i	4										
j	1	0,963	0,942	0,950	***0,894***	0,940	0,000	0,000	0,000	0,000	0,000
j	2	0,958	0,952	0,955	***0,918***	0,940	0,000	0,002	0,000	0,001	0,000
j	3	0,940	0,946	0,952	***0,921***	0,941	0,000	0,004	0,000	0,004	0,000
j	4	0,955	0,939	0,945	***0,909***	***0,934***	0,011	0,823	0,494	0,405	0,460
j	5										
k	1	0,941	0,950	0,950	0,952	***0,933***	0,000	0,000	0,000	0,000	0,000
k	2	0,944	0,953	0,962	0,958	0,941	0,000	0,000	0,000	0,000	0,000
k	3	0,951	0,948	0,948	0,943	***0,926***	0,000	0,003	0,000	0,002	0,000
k	4	0,951	0,940	0,943	0,946	***0,920***	0,000	0,740	0,289	0,350	0,278
k	5										
l	1	0,958	0,954	0,962	***0,865***	0,959	0,000	0,000	0,000	0,000	0,000
l	2	0,964	0,944	0,950	***0,917***	***0,933***	0,000	0,005	0,000	0,000	0,000
l	3	0,959	0,950	0,947	***0,931***	***0,931***	0,000	0,028	0,000	0,000	0,001
l	4										

7.3. Resultat av simulering 3; Felspecificerad skadefördelning och variansfunktion

Medelvärdet av Standard-GLM:s parameterskattningar stämde mycket bra med de riktiga parametrarna för skadefrekvensen. För alla andra parameterskattningar stämde medelvärdena sämre överens med de riktiga parametrarna om man jämför med de andra simuleringsfallen. Den metod som var något bättre än de övriga var Tweedie (se tabell 15). När väntevärdes riktighet undersöktes, enligt formel (37), så var det många parametrar som inte var väntevärdes riktiga (se tabell 16 där en etta betyder att testet i (37) inte kunde förkastas). Jung var den metod som hade flest väntevärdes riktiga parameterskattningar.

Tabell 15.

Simulering 3.		medelvärdet av parameterskattningarna					riktiga betavärdena			differens mellan r-premie och riktiga		
		Standard-GLM			Jung	Tweedie						
arg.	klass	frek	m- skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	Stand-GLM	Jung	Tweedie
i	1	-0,02	-0,044	-0,064	-0,07	-0,065	-0,020	-0,037	-0,057	-0,007	-0,013	-0,008
i	2	-0,01	-0,019	-0,029	-0,035	-0,03	-0,010	-0,018	-0,028	-0,001	-0,007	-0,002
i	3	0	0	-0,001	-0,003	-0,001	0	0	0	-0,001	-0,003	-0,001
i	4	0	0	0	0	0	0	0	0	0	0	0
j	1	-0,407	-0,461	-0,868	-0,876	-0,871	-0,406	-0,467	-0,873	0,005	-0,003	0,002
j	2	-0,288	-0,411	-0,699	-0,707	-0,702	-0,288	-0,416	-0,703	0,004	-0,004	0,001
j	3	-0,183	-0,403	-0,586	-0,593	-0,588	-0,182	-0,406	-0,588	0,002	-0,005	0
j	4	-0,088	-0,073	-0,161	-0,165	-0,162	-0,087	-0,069	-0,156	-0,005	-0,009	-0,006
j	5	0	0	0	0	0	0	0	0	0	0	0
k	1	-0,405	-0,457	-0,862	-0,87	-0,866	-0,406	-0,467	-0,873	0,011	0,003	0,007
k	2	-0,287	-0,4	-0,688	-0,692	-0,69	-0,288	-0,416	-0,703	0,015	0,011	0,013
k	3	-0,183	-0,392	-0,575	-0,579	-0,577	-0,182	-0,406	-0,588	0,013	0,009	0,011
k	4	-0,087	-0,058	-0,146	-0,147	-0,146	-0,087	-0,069	-0,156	0,010	0,009	0,010
k	5	0	0	0	0	0	0	0	0	0	0	0
l	1	-0,619	-0,502	-1,121	-1,129	-1,125	-0,619	-0,511	-1,130	0,009	0,001	0,005
l	2	-0,367	-0,364	-0,731	-0,738	-0,734	-0,368	-0,368	-0,735	0,004	-0,003	0,001
l	3	-0,167	-0,26	-0,426	-0,431	-0,428	-0,167	-0,262	-0,429	0,003	-0,002	0,001
l	4	0	0	0	0	0	0	0	0	0	0	0

.

Tabell 16.

Är parameterskattningarna väntevärdes riktiga?						
		Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie
i	1	1	1	1	1	1
i	2	1	1	1	1	1
i	3	1	1	1	1	1
i	4	1	1	1	1	1
j	1	1	0	1	1	1
j	2	1	1	1	1	1
j	3	1	1	1	1	1
j	4	1	1	1	1	1
j	5	1	1	1	1	1
k	1	1	0	0	1	1
k	2	1	0	0	0	0
k	3	1	0	0	1	0
k	4	1	0	0	1	0
k	5	1	1	1	1	1
l	1	1	0	0	1	1
l	2	1	1	1	1	1
l	3	1	1	1	1	1
l	4	1	1	1	1	1

Medelvärdet av Standard-GLM:s skattade varianser stämde mycket bra med stickprovsvariansen för skadefrekvensen, däremot för medelskadan var skillnaden större. För riskpremien, i premieargument *j* och *l,* stämde inte variansskattningarna bra överens med stickprovsvariansen. För Jungs metod var medelvärdet av de skattade varianserna lägre än stickprovsvarianserna för alla premieargument. Samma resultat gällde för Tweedies skattningarna av varianserna, de var lägre än stickprovsvariansen för alla premieargument. (Se tabell 17)

Tabell 17.

Simulering 3.		Roten ur medelvärdet av de skattade varianserna , m-std					S (empiriskt)				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	0,020	0,126	0,128	0,162	0,114	0,020	0,128	0,129	0,184	0,136
i	2	0,016	0,097	0,099	0,136	0,087	0,016	0,104	0,105	0,152	0,111
i	3	0,014	0,086	0,087	0,127	0,077	0,014	0,088	0,089	0,136	0,096
i	4										
j	1	0,018	0,112	0,113	0,113	0,100	0,018	0,101	0,103	0,130	0,109
j	2	0,019	0,119	0,121	0,117	0,108	0,019	0,104	0,106	0,130	0,111
j	3	0,020	0,127	0,129	0,126	0,117	0,020	0,112	0,113	0,141	0,119
j	4	0,023	0,146	0,148	0,162	0,128	0,024	0,155	0,158	0,198	0,165
j	5										
k	1	0,018	0,114	0,116	0,144	0,102	0,019	0,115	0,116	0,158	0,121
k	2	0,018	0,111	0,112	0,147	0,099	0,018	0,114	0,115	0,158	0,120
k	3	0,017	0,107	0,109	0,145	0,098	0,017	0,112	0,113	0,158	0,119
k	4	0,017	0,105	0,106	0,157	0,092	0,017	0,123	0,125	0,173	0,131
k	5										
l	1	0,019	0,120	0,122	0,121	0,107	0,020	0,119	0,120	0,152	0,128
l	2	0,018	0,115	0,117	0,133	0,103	0,018	0,125	0,128	0,163	0,136
l	3	0,017	0,106	0,107	0,141	0,095	0,017	0,120	0,122	0,156	0,130
l	4										

Alla tre metoders parameterskattningar var ganska lika varandra, men Tweedies skattningar var mest och Standard-GLM var minst lika de riktiga parametrarna (se tabell 18). Den metod som gav lägst empirisk varians var Standard-GLM och den som gav högs var Jungs (se tabell 17). När det gällde skattningarna av standardavvikelserna var Standard-GLM:s skattningar mest lika och Twwedies minst lika den empiriska standardavvikelsen (se tabell 19).

Tabell 18.

Kvadratsumma av riktiga - medelvärdet av de skattade, delat med antal (10^{-7})				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
4	624	613	448	416

Tabell 19.

Kvadratsumman av kvoten mellan m-std och S				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
0,983	0,983	0,982	0,781	0,694

För Standard-GLM beräknades konfidensintervallen ibland korrekt. För Jung såg man samma mönster som tidigare, att för de argument där durationen inte hade en multiplikativ struktur var konfidensintervallen lägre än de skulle vara. Tweedie

modellens 95 procentiga konfidensintervall hade egentligen lägre konfidensgrad än vad den skulle ha. (Se tabell 20)

Tabell (20) visar även hur ofta metoderna lyckades identifiera att det fanns en skillnad mellan klasserna. Standard-GLM och Jungs metod lyckades oftare inte förkasta att det inte fanns en skillnad mellan basklassen och klass tre i premieargument *i,* än vad Tweedie metoden gjorde. Tweedie sa att parametern hade betydelse för ofta. För klass ett och två, där det fanns en liten skillnad, var Jung och Tweedie de metoder som oftast sa att det fanns en skillnad.

Tabell 20.

Simulering 3.		Beräknas de 95%-iga KI korrekt?					Hur ofta förkastas inte att parametern inte skiljer sig från bascellen?				
		Standard-GLM			Jung	Tweedie	Standard-GLM			Jung	Tweedie
arg.	klass	frek	m-skada	r-premie	r-premie	r-premie	frek	m-skada	r-premie	r-premie	r-premie
i	1	***0,965***	0,948	0,946	***0,929***	***0,908***	0,742	0,919	0,891	0,789	0,815
i	2	0,95	0,941	0,937	0,941	***0,875***	0,836	0,925	0,907	0,878	0,846
i	3	0,948	0,957	0,958	0,954	***0,877***	0,939	0,955	0,949	0,952	0,907
i	4										
j	1	0,945	0,959	0,956	***0,903***	***0,932***	0	0,006	0	0,004	0,001
j	2	0,954	0,965	***0,967***	***0,916***	0,939	0	0,024	0	0,004	0,002
j	3	0,951	***0,967***	***0,976***	***0,907***	0,944	0	0,051	0,003	0,023	0,007
j	4	0,944	***0,934***	***0,931***	***0,887***	***0,881***	0,015	0,853	0,669	0,581	0,568
j	5										
k	1	0,942	0,945	0,945	0,947	***0,892***	0	0,01	0,001	0,005	0,001
k	2	0,944	0,944	0,941	0,94	***0,878***	0	0,029	0	0,019	0,001
k	3	0,958	0,938	***0,936***	0,944	***0,891***	0	0,035	0,003	0,03	0,004
k	4	0,946	***0,918***	***0,914***	0,949	***0,834***	0	0,82	0,562	0,674	0,495
k	5										
l	1	0,942	0,953	0,952	***0,85***	***0,896***	0	0,008	0	0	0
l	2	0,953	***0,925***	***0,925***	***0,891***	***0,868***	0	0,065	0,001	0,008	0,003
l	3	0,937	***0,918***	***0,923***	0,939	***0,854***	0	0,208	0,025	0,079	0,026
l	4										

7.4. Resultat av simulering 4; Avvikelser från multiplikativitet

När avvikelser från multiplikativitet undersöktes kunde inte samma undersökningar göras som i de andra simuleringsfallen, eftersom jämförelsen nu gjordes i tariffcellerna. Först undersöktes de skattade parametervärdena enligt formel (34). Resultatet blev att alla tre metoderna skattade väntevärdet i tariffcellerna lika bra eller lika dåligt, se tabell 21. För att kunna jämföra stickprovsvariansen togs ett medelvärde av dem över alla tariffcellerna. Man bör tänka på att parameterskattningarna har systematiska fel som påverkar stickprovsvariansen. Detta beror på att parameterskattningarna inte kan skattas rätt då vi utgått från en modell som inte är fullt multiplikativ. Tweedie var den metod som gav lägst stickprovsvarians och Jung den som gav högst, se tabell 22. Hur bra variansskattningarna stämde överens med stickprovsvariansen undersöktes för

Standard-GLM och Tweedie. I tabell 23 ser man att bådas variansskattningar är bra men att Tweedies är något bättre.

Tabell 21

Kvadratsumman av de observerade värdena minus de skattade, viktat med durationen				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
2100504	9,75E+14	1,27E+13	1,27E+13	1,27E+13

Tabell 22

Medelvärdet av stickprovsvariansen över alla celler				
Standard-GLM			Jung	Tweedie
frek	m-skada	r-premie	r-premie	r-premie
0,024	0,109	0,112	0,114	0,111

Tabell 23

Kvadratsumman av stickprovsvariansen dividerat med medelvärdet av den skattade variansen dividerat med antal celler.			
Standard-GLM			Tweedie
frek	m-skada	r-premie	r-premie
0,977	0,971	0,969	0,981

Alltför stora avvikelser från multiplikativitet har inte undersökts eftersom det är rimligt att anta att man inte ansätter en multiplikativ modell om man inte tror att det är multiplikativitet som råder.

8. Slutsatser och diskussion

I arbetet har skattade relationstal, varianser samt konfidensintervall undersökts för de olika metoderna. När de verkliga skadedata från bussförsäkringar undersöktes var det svårt att avgöra vilken av metoderna som skattade bäst. Däremot när skadedata simulerades fram var det lättare att jämföra metoderna, eftersom skattningarna kunde jämföras med de fixerade relationstalen. Det gav också en möjlighet att undersöka olika förhållanden, som till exempel olika fördelningar för skadedata.

Om man först tittar på hur bra metoderna skattade relationstalen, så var Standard-GLM den metod som var mest tillförlitlig. Det var den dels för att parameterskattningar oftast stämde bäst överens med de fixerade relationstalen, dels för att dess skattningarna oftast varierade minst. Även när det gällde varianser var Standard-GLM:s skattningar mest tillförlitliga, eftersom de bäst överensstämde med den empiriska variansen. När det gäller vilken metod som oftast skattade konfidensintervallen korrekt så var det ingen av metoderna som utmärkte sig. I det tredje simuleringsfallet skattades dock konfidensintervallen ofta fel för alla tre metoderna. I ett avseende var Jungs metod och Tweedie-metoden bättre än Standard-GLM och det var hur ofta som metoderna identifierade att det fanns en skillnad mellan två klasser. Detta beror på att de skattade varianserna var för låga, vilket bör betyda att de oftare säger att det finns en skillnad då skillnad inte existerar.

Den metod som bäst klarade av de förhållanden som undersöktes i detta arbete, var Standard-GLM. Det beroendet mellan skadefrekvensen och medelskadan som metoden bortser ifrån vid beräkning av riskpremiens variansskattning, har inte påverkat skattningarna nämnvärt. När det gäller LF-Wasa:s variansskattning så har den påverkats negativt då durationen inte haft en multiplikativ struktur. I det här arbetet har inte ett optimalt p undersökts för Tweedie metoden vilket man bör ha i åtanke eftersom Tweddies skattningar förmodligen skulle blivit annorlunda om detta gjorts.

9. Referenser

[1] Gut, A. (1995), *An Intermediate Course in Probability,* Springer-Verlag New York, Inc

[2] Jung, J (1968) *On automobile insurance ratemaking, Estimating relativities in a multiplicative model*, Astin Bulletin, 5, 41-48

[3] Jørgensen, B och Souza, M (1994), *Fitting Tweedie's Compound Poisson Model to Insurance Claims Data,* SAJ

[4] Ohlsson, E & Johansson, B. (2003), *Prissättning inom sakförsäkring med Generaliserade linjära modeller,* Kompendium, Matematisk statistik, Stockholms Universitet

[5] Ohlsson, E (1998), *Felfortplantningsformlerna,* Kompendium, Matematisk statistik, Stockholms Universitet

[6] Ohlsson, E (2002), *Härledning av ad hoc konfidensintervall,* PM 2002-11-19

[7] Rosenlund, S (2002), *Evaluation of GLM in non-life insurance,*
PM 2002-01-22, Länsförsäkringsbolagen

[8] Ross, S. (2000), *Introduction to Probability Models* Seventh edition, Academic Press, USA

[9] SAS institute Inc, Technical Report P-243 (1993) *SAS/STAT Software: The GENMOD Procedure*, SAS institute Inc, USA

[10] Sundberg, R. (1997), *Kompendium i Tillämpad Matematisk Statistik,* Matematisk statistik, KTH

[11] Tamhane, A & Dunlop, D (2000), *Statistics and data analysis, from elementary to intermediate,* Prentice Hall, Inc. Upper Saddle River, USA

10. Appendix

- Tabell över de undersökta metoderna.

Modellerna	**Fördelning**	**Variansfunktion**
Jungs metod	fördelningsfri	
GLM	Skadefrekvensen~Poisson	$v(\mu)=\mu$
	Medelskadan~Gamma	$v(\mu)=\mu^2$
Tweedie-modellerna	Riskpremien~Sammansatt Poisson	$v(\mu)=\mu^{1,5}$

- Tabell över de olika simuleringsfallen.

	Simuleringsfallen	**Skadefrekvensen simuleras från**	**Medelskadan simuleras från**	**Varians-funktionen**
1	Standard modellen	Poissonförd.	Gammaförd.	$v(\mu)=\mu^2$
2	Felspecificerad skadefördelning	Poissonförd.	Lognormalförd.	$v(\mu)=\mu^2$
3	Felspecificerad skadefördelning och variansfunktion	Poissonförd.	Lognormalförd.	$v(\mu)=k(\mu)\mu^2$
4	Avvikelse från multiplikativitet	Poissonförd.	Gammaförd.	$v(\mu)=\mu^2$

- Den kumulantgenererande funktionen för den sammansatt poissonfördelningen:

För att se att Y blir sammansatt Poissonfördelad undersöks den momentgenererande funktionen, M(t) :

$$M(t) = E(e^{tY}) = \sum_n E\left(e^{t(Z_1+\cdots+Z_n)/w} \middle| N=n\right) P(N=n)$$

$$= \sum_n E\left(e^{t(Z_1+\cdots+Z_n)/w}\right) P(N=n) = \{iid\}$$

$$= \sum_n E\left(e^{t\frac{z_1}{w}^n}\right) P(N=n)$$

$$= \sum_n E\left(e^{tZ_1/w}\right)^n e^{-\lambda w} P(N=n)$$

$$= \sum_n \left(M_{Z_1/w}(t)\right)^n P(N=n)$$

$$= \sum_n e^{n\log M_{Z_1/w}(t)} P(N=n)$$

$$= E(e^{N\log M_{Z_1/w}(t)}) = M_N(\log(M_{Z_1/w}(t))$$

$$\log(\mathrm{M}(t)) = \log(M_N(\log(M_{Z_1/w}(t))) \quad \Leftrightarrow \quad \Psi(t) = \Psi_N(\Psi_{Z_1/w}(t))$$

- Sats 1 i felforplantningsformlerna.

Låt $\mathbf{X}_n$, $n = 1,2,3...$, vara en följd av stokastiska variabler som är asymtotiskt normalfördelade

$$\sqrt{n}(\mathbf{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Låt $g(x)$ vara en deriverbar funktion vars derivata är kontinuerlig i μ och skilld från noll där. Då gäller att :

$$\sqrt{n}(g(\mathbf{X}_n) - g(\mu)) \xrightarrow{d} N(0, [g'(\mu)]^2 \sigma^2)$$

[5]

- Centrala gränsvärdessatsen
Låt $X_1, X_2, \ldots$ vara oberoende, lika fördelade stokastiska variabler med ändligt väntevärde, μ och varians, σ^2.

$$S_n = \sum_{i=1}^{n} X_i$$

Då gäller följande : $\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$ då $n \to \infty$

[1, sid 172]

- Definition av Poisson-process enligt Ross (2000) sid 258.
En räkneprocess, {N(t),t>0}, är en Poisson-process om antal händelser vid tiden noll är noll och om processen har stationära och oberoende inkrement. Även P{N(h)=1}=λh+o(h) och P{N(h)>=2}=o(h) ska vara uppfyllt. Detta kan tolkas som att händelserna kommer en och en.
[8]

- Definition av konsistent skattning:

En skattning $\hat{\theta}_n$, som är definierad för varje n, är en konsistent skattning av θ om den går mot θ i sannolikhet, dvs för $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

- Skattningar av Jungs parametrar är samma som GLM under Poissonantagande.

Antag att man har en relativ Poissonfördelning som i (16)

X_{ij} är skadekostnad i tariffcell ij med duration w_{ij}, $X_{ij} \sim Poisson(w_{ij}\mu_{ij})$

$Y_{ij} = X_{ij} / w_{ij}$

$$P(Y_{ji} = y_{ij}) = P(X_{ij} = w_{ij} y_{ij}) = \exp(-w_{ij}\mu_{ij}) \frac{(w_{ij}\mu_{ij})^{y_{ji} w_{ji}}}{(w_{ij}\mu_{ij})} = \exp\{w_{ij}(y_{ij} \log(\mu_{ij}) - \mu_{ij}) + c(y_{ij}, w_{ij})\}$$

$$l = \sum_i \sum_j w_{ij}(y_{ij} \log(\mu_{ij}) - \mu_{ij}) = w_{ij}(y_{ij}(\log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j})) - \gamma_0\gamma_{1i}\gamma_{2j})$$

$$\frac{\partial l}{\partial \gamma_0} = \sum_i \sum_j w_{ij}\left(\frac{y_{ij}}{\gamma_0} - \gamma_{1i}\gamma_{2j}\right) = 0 \quad \Rightarrow \qquad \gamma_0 = \frac{\sum_i \sum_j w_{ij} y_{ij}}{\sum_i \sum_j w_{ij}\gamma_{1i}\gamma_{2j}}$$

$$\frac{\partial l}{\partial \gamma_{1i}} = \sum_j w_{ij}\left(\frac{y_{ij}}{\gamma_{1i}} - \gamma_0\gamma_{2j}\right) = 0 \qquad \Rightarrow \gamma_{1i} \qquad = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}\gamma_0\gamma_{2j}} \qquad i = 1,2,3,....$$

$$\frac{\partial l}{\partial \gamma_{2j}} = \sum_i w_{ij}\left(\frac{y_{ij}}{\gamma_{2j}} - \gamma_0\gamma_{1i}\right) = 0 \qquad \Rightarrow \gamma_{2j} \qquad = \frac{\sum_i w_{ij} y_{ij}}{\sum_i w_{ij}\gamma_0\gamma_{1i}} \qquad j = 1,2,3,....$$

Vilket är samma som Jungsekvationer i (3).

- Derivatan av log-likelihood funktionen för EDM:s frekvensfunktion.

log - likelihood funktionen för frekvensfunktionen

$$l(\grave{e}, \phi, \mathbf{y}) = \frac{1}{\phi}\sum_i w_i(y_i\theta_i - b(\theta_i)) + \sum_i c(y_i, \phi, w_i)$$

$$\mu_i = b'(\theta_i) \qquad g(\mu_i) = \eta_i = \sum_j x_{ij}\beta_j$$

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \sum_i \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} =$$

$$= \frac{1}{\phi}\sum_i (w_i(y_i - b'(\theta_i)))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \quad \Rightarrow \quad \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{v(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_i (y_i - b'(\theta_i))\frac{w_i x_{ij}}{v(\mu_i)g'(\mu_i)}$$

[4]