Åbo Akademi

FAKULTETSOMRÅDET FÖR
NATURVETENSKAPER OCH TEKNIK

Avhandling Pro Gradu

Generaliserade linjära modeller med tillämpning på premieprissättning

Skribent:
Tony Miiros, 36032

Handledare:
Paavo Salminen

2018

Innehåll

Kapitel 1

Inledning

Generaliserade linjära modeller (GLMs) är som namnet indikerar en generalisering av den vanliga linjära regressionen som har formen $Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$. En linjär regression innebär att om en förklarande variabel (något X_i) ökar med en konstant mängd så ökar eller minskar även responsvariabeln (Y) med en konstant mängd. Detta indikerar att Y följer en normalfördelning, däremot tillåts responsvariabeln i en GLM följa en annan fördelning än normalfördelningen. I GLMs tas detta i beaktande med hjälp av ett par egenskaper, en linjär prediktor och en variansfunktion, som närmare beskrivs senare i avhandlingen.

J. Nelder och R.W.M. Wedderburn utvecklade idén om GLMs år 1972 under den tidsperiod, då statistiska uträkningar allt mer började göras med hjälp av datorer. GLMs visade sig snabbt vara smidigare och mer tidssparande än t.ex. maximum likelihood-metoden. Före Nelders och Wedderburns upptäckt skattades modeller av GLM-typ även med hjälp av Newton-Raphson-metoden.

GLMs används inom demografi, ekonomi, ingenjörskonst, medicin, psykologi och försäkringsmatematik, och kommandon för GLMs finns nuförtiden med i många statistiska mjukvaror, som R, SAS, Genstat och SPSS. Föreliggande arbete inom försäkringsmatematik analyserar hur man med hjälp av GLMs kan modellera nyckeltal som kravfrekvens, kravstorlek och riskpremie, för att erhålla rättvisa premier för dessa.

Denna avhandling presenterar bakomliggande teori kring GLMs och ger exempel med tillämpning inom försäkringsbranschen för icke-livsförsäkringar. Några praktiska uppgifter görs i R för att erhålla så kallade relativiteter, som bestämmer hur premier borde prissättas beroende på olika faktorer. Dessa faktorer kan vara

region, årsmodell eller förarens ålder vid undersökning av till exempel kaskoförsäkringar för bilar. Några praktiska uppgifter ur [1], ”Non-Life Insurance Pricing with Generalized Linear Models”, kommer även att lösas, t.ex. att göra GLM-skattningar för relativiteter. Sist och slutligen tas en mer omfattande ”case study” itu med, som behandlar kaskoförsäkring för motorcyklar. Samtliga uppgifter bygger på data från det svenska försäkringsbolaget Wasa, före dess fusion med Länsförsäkringar Alliance.

Kapitel 2

Inledande sannolikhetsteori

Detta kapitel kommer att innehålla grundläggande definitioner och satser inom sannolikhetsläran, som smidigt kan hänvisas till senare i avhandlingen.

Definition 2.1. Den momentgenererande funktionen till en stokastisk variabel X definieras som

$$M_X(t) := \mathbb{E}[e^{tX}],$$

förutsatt att väntevärdet existerar för $|t| < \delta$ för något $\delta > 0$.

Lemma 2.2. Låt X_1 och X_2 vara två oberoende stokastiska variabler med momentgenererande funktioner $M_{X_1}(t)$ respektive $M_{X_2}(t)$. Då gäller att $M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t)$.

Bevis. Vi har att

$$\begin{aligned} M_{X_1}(t) \cdot M_{X_2}(t) &= \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] = \mathbb{E}[e^{tX_1} e^{tX_2}] \\ &= \mathbb{E}[e^{t(X_1+X_2)}] = M_{X_1+X_2}(t), \quad \square \end{aligned}$$

där den andra likheten gäller ty X_1 och X_2 antas vara oberoende.

Definition 2.3. Den kumulantgenererande funktionen $\Psi_X(t)$ till en stokastisk variabel X definieras som

$$\Psi_X(t) = \log \mathbb{E}[e^{tX}] = \log M_X(t),$$

förutsatt att den momentgenerande funktionen $M_X(t)$ existerar för $|t| < \delta$ för något $\delta > 0$.

Ur den kumulantgenererande funktionen kan man erhålla kumulanterna κ_n, enligt

$$\Psi(t) = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} = \mu t + \sigma^2 \frac{t^2}{2} + ...$$

Denna utveckling är en Maclaurinutveckling och den n:te kumulanten kan fås genom att derivera ovanstående uttryck n gånger och sätta $t = 0$:

$$\kappa_n = \Psi^{(n)}(0).$$

De två första kumulanterna blir således:

$$\kappa_1 = \Psi'(0) = \mu = \mathbb{E}(X)$$

$$\kappa_2 = \Psi''(0) = \sigma^2 = \mathrm{Var}(X)$$

Lemma 2.4. För två oberoende stokastiska variabler X och Y, samt en konstant c, gäller
1) $\Psi_{cX}(t) = \Psi_X(ct)$ och
2) $\Psi_{X+Y}(t) = \Psi_X(t) + \Psi_Y(t)$.

Bevis. Betrakta först 1):

$$\begin{aligned} 1) \quad \Psi_{cX}(t) &= \log(M_{cX}(t)) = \log(\mathbb{E}(e^{t \cdot cX})) = \log(\mathbb{E}(e^{ctX})) = \log(M_X(ct)) \\ &= \Psi_X(ct). \end{aligned}$$

För 2) fås:

$$\begin{aligned} 2) \quad \Psi_{X+Y}(t) &= \log(M_{X+Y}(t)) = \log(\mathbb{E}(e^{t(X+Y)})) = \log(\mathbb{E}(e^{tX+tY})) \\ &= \log(\mathbb{E}(e^{tX}e^{tY})) = \log(\mathbb{E}(e^{tX})\mathbb{E}(e^{tY})) \\ &= \log(\mathbb{E}(e^{tX})) + \log(\mathbb{E}(e^{tY})) = \Psi_X(t) + \Psi_Y(t) \quad \square \end{aligned}$$

Definition 2.5. Skevhet för en stokastisk variabel är det tredje standardiserade momentet γ_1, som definieras som

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

där μ är väntevärdet, σ är standardavvikelsen och κ_i är den i:te kumulanten.

Skevhet är ett mått på hur assymetrisk sannolikhetsfördelningen för en stokastisk variabel är. Om fördelningen har en lång "svans" åt vänster har den en negativ skevhet medan en svans åt höger innebär en positiv skevhet. För normalfördelningen gäller att skevheten är lika med noll.

Definition 2.6. En familj av stokastiska variabler $\{X(t), t \in T\}$, där T är en delmängd av $[0, \infty)$, kallas för en *stokastisk process*. När $T = \mathbb{N}$ eller $T = \mathbb{N}_0$ är $\{X(t), t \in T\}$ en stokastisk process i diskret tid och när $T = [0, \infty)$ är den en stokastisk process i kontinuerlig tid.

Man kan betrakta T som en mängd av tidspunkter. Om T är ett enda tal, t.ex. $T = \{1\}$, är processen $\{X(t), t \in T\} \equiv X(1)$ endast en stokastisk variabel. Om T är ändlig, t.ex. $T = \{1, 2, ..., n\}$, får vi en stokastisk vektor. Stokastiska processer är alltså generaliseringar av stokastiska vektorer och modellerar ofta förändringar för ett system som varierar slumpmässigt över tid.

Definition 2.7. En *Poissonprocess* är en stokastisk process i kontinuerlig tid med de positiva heltalen som värdemängd. Följande egenskaper gäller definitionsmässigt för en Poissonprocess $\{N(t), t \geq 0\}$ med intensitet $\lambda > 0$:

- $N(t)$ är växande och heltalsvärd, med kravet att $N(0) = 0$.
- $\{N(t), t \in T\}$ har oberoende tillskott, dvs. för varje val av $0 \leq t_1 < t_2 < t_3 < t_4$ gäller det att $N(t_2) - N(t_1)$ och $N(t_4) - N(t_3)$ är oberoende.
- $N(s + t) - N(t)$ är Poissonfördelad med parametern $\lambda(t + s) - \lambda t = \lambda s$

Detta är fallet med en *homogen* Poissonprocess, som används i denna avhandling. Detta innebär att för ett tidsintervall med längden s är intensiteten λs, med andra ord att $\mathbb{E}[N(s)] = \lambda s$. Detta ger oss att sannolikheten att den stokastiska variabeln $N(t)$ är lika med n fås som

$$\mathbb{P}(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

En Poissonprocess används för att beskriva med vilken intensitet slumpmässiga händelser sker. Om processen är *inhomogen* byts λ ut mot $\lambda(t)$ och parametern i den tredje punkten ovan (λs) byts ut mot $\int_t^{s+t} \lambda(u)du$.

Definition 2.8. Gammafördelningen är en kontinuerlig sannolikhetsfördelning. En gammafördelad stokastisk variabel $X \sim \text{Gamma}(\alpha, \beta)$ har täthetsfunktionen

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad 0 < x < \infty,$$

där $\alpha > 0$ och $\beta > 0$ är parametrar i fördelningen och Γ står för gammafunktionen, $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. Fördelningen har väntevärdet $\mathbb{E}(X) = \alpha/\beta$ och variansen $\mathrm{Var}(X) = \alpha/\beta^2$, dessa härleds nedan.

En viktig egenskap för gammafunktionen är att $\Gamma(z) = (z-1)\Gamma(z-1)$:

$$\begin{aligned}\Gamma(z) &= \int_0^\infty t^{z-1}e^{-t}dt \\ &= [-t^{z-1}e^{-t}]_0^\infty + \int_0^\infty (z-1)t^{z-2}e^{-t}dt \\ &= (z-1)\int_0^\infty t^{z-2}e^{-t}dt \\ &= (z-1)\Gamma(z-1).\end{aligned}$$

Gammafördelningens väntevärde $\mathbb{E}(X)$ fås enligt

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x f_X(x)dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^\alpha e^{-\beta x}dx \quad \text{(substituerar } \beta x = t \implies dx = dt/\beta) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \big(\frac{t}{\beta}\big)^\alpha e^{-t}\frac{dt}{\beta} \\ &= \frac{1}{\beta\Gamma(\alpha)}\int_0^\infty t^\alpha e^{-t}dt \\ &= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta}.\end{aligned}$$

För att härleda variansen behöver vi $\mathbb{E}(X^2)$,

$$\begin{aligned}\mathbb{E}(X^2) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha+1}e^{-\beta x}dx \quad \text{(substituerar igen } \beta x = t) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \big(\frac{t}{\beta}\big)^{\alpha+1} e^{-t}\frac{dt}{\beta} \\ &= \frac{1}{\beta^2\Gamma(\alpha)}\int_0^\infty t^{\alpha+1}e^{-t}dt \\ &= \frac{\Gamma(\alpha+2)}{\beta^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)\Gamma(\alpha)}{\beta^2\Gamma(\alpha)} = \frac{\alpha(\alpha+1)}{\beta^2}.\end{aligned}$$

Härifrån får vi att

$$\mathrm{Var}(X) = \frac{\alpha(\alpha+1)}{\beta^2} - \big(\frac{\alpha}{\beta}\big)^2 = \frac{\alpha}{\beta^2}.$$

Typiska tillämpningar för gammafördelningen är modellering av regnmängd, hur signalstyrka avtar i trådlös kommunikation eller, mest relevant för denna avhandling, kravstorleken för försäkringar.

Sats 2.9. *För två oberoende gammafördelade stokastiska variabler* $X_1 \sim \text{Gamma}(\alpha_1, \beta)$ *och* $X_2 \sim \text{Gamma}(\alpha_2, \beta)$ *gäller att deras summa,* $Y = X_1 + X_2$*, är gammafördelad enligt* $\text{Gamma}(\alpha_1 + \alpha_2, \beta)$*.*

BEVIS. Vi utnyttjar den momentgenerande funktionen och det faktum att gammafördelningen är en kontinuerlig fördelning, vilket ger att vi kan uttrycka dess momentgenererande funktion med en integral. Den momentgenerande funktionen för X_1, för något t så att $\beta - t > 0$, blir således:

$$\begin{aligned} M_{X_1}(t) = M_{X_1}(t; \alpha_1, \beta) = \mathbb{E}[e^{tX_1}] &= \int_0^\infty e^{tx} f_{X_1}(x; \alpha_1, \beta) dx \\ &= \int_0^\infty e^{tx} \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\beta x} dx \\ &= \frac{\beta^{\alpha}}{\Gamma(\alpha_1)} \int_0^\infty x^{\alpha_1 - 1} e^{-(\beta - t)x} dx \\ &= \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_1)}{(\beta - t)^{\alpha_1}} \\ &= \frac{1}{(1 - t/\beta)^{\alpha_1}}. \end{aligned}$$

Motsvarande momentgenererande funktion för X_2 blir då $1/(1 - t/\beta)^{\alpha_2}$. För $Y = X_1 + X_2$ får vi nu, enligt lemma **??**, att $M_Y(t) = M_{X_1}(t) \cdot M_{X_2}(t) = [1/(1 - t/\beta)^{\alpha_1}] \cdot [1/(1 - t/\beta)^{\alpha_2}] = 1/(1 - t/\beta)^{\alpha_1 + \alpha_2}$. I och med att den momentgenererande funktionen entydigt bestämmer fördelningen för den stokastiska variabeln får vi att $Y \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$. □

Definition 2.10. Täthetsfunktionen $f_{X,Y}$ för en kontinuerlig stokastisk vektor (X, Y) ges av

$$F_{X,Y}(x, y) := \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} dv f_{X,Y}(u, v) \quad \forall x, y.$$

Om $f_{X,Y}$ är kontinuerlig i (x, y) gäller även att

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = f_{X,Y}(x, y).$$

Definition 2.11. Den betingade tätheten av X givet att $Y = y$ är funktionen

$$f_{X|Y}(x|y) = \begin{cases} \dfrac{f_{X,Y}(x, y)}{f_Y(y)} & \text{om } f_Y(y) > 0, \\ 0 & \text{om } f_Y(y) = 0, \end{cases}$$

där f_Y är Y:s täthet.

Från definitionen ovan erhåller vi det betingade väntevärdet av X givet att $Y = y$ som

$$\mathbb{E}(X|Y = y) := \int_{-\infty}^{\infty} x f_{X|Y}(x)dx. \tag{2.0.1}$$

Om X och Y är *oberoende* för varje x och y gäller dessutom att

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Sats 2.12. *För n stycken oberoende stokastiska variabler $X_i, \mathbb{E}(X_i^2) < \infty \ \forall \ i$, med varianser $\mathrm{Var}(X_i)$, gäller att*

$$\mathrm{Var}\big(\sum_{i=1}^{n} X_i\big) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \tag{2.0.2}$$

Bevis. För två stokastiska variabler X_1 och X_2 har vi från variansens definition

$$\mathrm{Var}[X_1 + X_2] = \mathbb{E}[(X_1 + X_2)^2] - \mathbb{E}[X_1 + X_2]^2$$

För två oberoende stokastiska variabler X_1 och X_2 gäller att $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$. Väntevärdet är dessutom en linjär operator, vilket ger att $\mathbb{E}[aX_1] = a\mathbb{E}[X_1]$, där a är en konstant, och $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$. Med hjälp av dessa resultat utvecklar vi uttrycket ovan:

$$\begin{aligned}
\mathrm{Var}[X_1 + X_2] &= \mathbb{E}[X_1^2 + 2X_1X_2 + X_2^2] - \mathbb{E}[X_1 + X_2]\mathbb{E}[X_1 + X_2] \\
&= \mathbb{E}[X_1^2] + \mathbb{E}[2X_1X_2] + \mathbb{E}[X_2^2] - (\mathbb{E}[X_1] + \mathbb{E}[X_2])(\mathbb{E}[X_1] + \mathbb{E}[X_2]) \\
&= \mathbb{E}[X_1^2] + 2\mathbb{E}[X_1]\mathbb{E}[X_2] + \mathbb{E}[X_2^2] - \mathbb{E}[X_1]^2 - 2\mathbb{E}[X_1]\mathbb{E}[X_2] - \mathbb{E}[X_2]^2 \\
&= \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 + \mathbb{E}[X_2^2] - \mathbb{E}[X_2]^2 \\
&= \mathrm{Var}[X_1] + \mathrm{Var}[X_2]
\end{aligned}$$

Detta fall utvidgas lätt till fallet med n st stokastiska variabler, och satsen är bevisad. □

Sats 2.13. *För två oberoende Poissonfördelade stokastiska variabler $X_1 \sim \mathrm{Po}(\lambda_1)$ och $X_2 \sim \mathrm{Po}(\lambda_2)$ gäller att deras summa, $Y = X_1 + X_2$, är Poissonfördelad med parametern $\lambda_1 + \lambda_2$.*

Bevis. Vi utnyttjar lagen om totalsannolikhet och det faktum att X_1 och X_2

är oberoende:

$$\begin{aligned}\mathbb{P}(Y=n) &= \mathbb{P}(X_1+X_2=n) = \sum_{k=0}^{n}\mathbb{P}(X_1+X_2=n|X_2=k)\mathbb{P}(X_2=k)\\ &= \sum_{k=0}^{n}\mathbb{P}(X_1=n-k)\mathbb{P}(X_2=k) = \sum_{k=0}^{n}\frac{\lambda_1^{n-k}}{(n-k)!}e^{-\lambda_1}\frac{\lambda_2^k}{k!}e^{-\lambda_2}\\ &= \frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda_1^{n-k}\lambda_2^k e^{-(\lambda_1+\lambda_2)} = \frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)},\end{aligned}$$

vilket ger att $Y \sim \mathrm{Po}(\lambda_1+\lambda_2)$. Den sista likheten fås med hjälp av binomialsatsen. □

Lemma 2.14. Låt X och Y vara stokastiska variabler, sådana att $\mathbb{E}(X)$ existerar. Då har det betingade väntevärdet $\mathbb{E}(X|Y)$ av X med avseende på Y följande egenskaper:

1) $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$,
2) $\mathbb{E}(X|Y) = \mathbb{E}(X)$, om X och Y är oberoende,
3) $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathbb{E}(X|Y)]$.

Bevis. 1) För en kontinuerlig variabel X gäller

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X|Y)) &= \int_{-\infty}^{\infty}\mathbb{E}(X|Y=y)f_Y(y)dy\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X|Y}(x|y)dxf_Y(y)dy \quad \text{(Från (2.0.1))}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X|Y}(x|y)f_Y(y)dxdy\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dydx \quad \text{(Definition 2.10)}\\ &= \int_{-\infty}^{\infty}xf_X(x)dx\\ &= \mathbb{E}(X).\end{aligned}$$

Fallet med diskreta variabler visas på ett liknande sätt, där integralerna byts ut mot summor.

2) Formeln gäller, ty

$$\begin{aligned}\mathbb{E}(X|Y) &= \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}dx \\ &= \int_{-\infty}^{\infty} x\frac{f_X(x)f_Y(y)}{f_Y(y)}dx = \int_{-\infty}^{\infty} x f_X(x)dx \\ &= \mathbb{E}(X).\end{aligned}$$

3) Detta bevisas på följande sätt:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= \mathbb{E}[\mathbb{E}(X^2|Y)] - [\mathbb{E}[\mathbb{E}(X|Y)]]^2 \\ &= \mathbb{E}\big[\mathrm{Var}(X|Y) + [\mathbb{E}(X|Y)]^2\big] - [\mathbb{E}[\mathbb{E}(X|Y)]]^2 \\ &= \mathbb{E}[\mathrm{Var}(Y|X)] + \big(\mathbb{E}\big[[\mathbb{E}(X|Y)]^2\big] - \big[\mathbb{E}[\mathbb{E}(Y|X)]\big]^2\big) \\ &= \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathbb{E}(X|Y)]. \quad \square\end{aligned}$$

Definition 2.15. Låt X vara en stokastisk variabel med väntevärde $\mathbb{E}(X) = \mu$ och varians $\mathrm{Var}(X) = \sigma^2$. Då definieras variationskoefficienten CV_X av X enligt

$$CV_X = \frac{\sigma}{\mu},$$

där $\sigma = \sqrt{\mathrm{Var}(X)}$ är standardavvikelsen av X.

Kapitel 3

Försäkringsteori för icke-livsförsäkringar

3.1 Värderingsfaktorer, klasser och nyckeltal

För varje typ av försäkring bestäms premien på basen av ett antal variabler, så kallade **värderingsfaktorer**. För att estimera sambanden mellan dessa använder man en statistisk modell, som vi kommer att precisera senare i denna avhandling. Värderingsfaktorerna hör ofta till någon av följande kategorier:

• **Egenskaper hos försäkringstagaren**: ålder eller kön om försäkringstagaren är en privatperson, affärsområde för ett företag, etc.

• **Egenskaper hos det försäkrade objektet**: ålder för en moped eller bil, typ av byggnad, etc.

• **Egenskaper hos den geografiska regionen**: per capita inkomst eller folktäthet för försäkringstagarens bostadsområde, etc.

De variabler som används för premieprissättning är de som finns tillgängliga för försäkringsbolaget, antingen direkt (ålder, kön) eller via datainsamling. Det är svårt att få pålitliga data för andra faktorer, som till exempel ”körbeteende för en person”, ”hur många sängrökare som finns i hushållet”, etc. Dessutom får inte värderingsfaktorerna vara förolämpande mot försäkringstagarna.

Den statistiska undersökningen som en aktuarie utför för att få en tariff kallas för *tariffanalys*. Den baseras på försäkringsbolagets egna data från försäkringskontrakten och kraven för portföljen. Ibland använder bolaget utöver sina egna data även extern data från till exempel en statistikmyndighet.

Olika typer av försäkringar (t.ex. för motorcyklar eller bilar) delas in i *tariff-celler*. Om vi till exempel väljer att göra en tariffanalys för motorcyklar, delas de in i M stycken värderingsfaktorer (motorns storlek, årsmodell, antal växlar), där vi låter m_i stå för antalet klasser i värderingsfaktor i. Antag att vi har n värderingsfaktorer, alltså $M = n$. Då kan vi benämna en tariffcell $(i_1, i_2, \ldots, i_n)$, där i_k betecknar klassen för den k:te värderingsfaktorn.

Exempel 3.1. Tariffanalys för en viss typ av mopeder med två värderingsfaktorer, i = antalet växlar och j = årsmodell. Klasserna blir då 1 växel, 2 växlar, ..., n växlar respektive 1990, 1991, ... Då får vi olika celler enligt (1 växel, 1995), (3 växlar, 2001) och så vidare, och undersöker dessa skilt.

Härnäst definieras några grundkoncept som används i samband med en tariff-analys.

• **Försäkringskontraktets löptid** är tiden som försäkringskontraktet är i kraft, mäts vanligtvis i år. Om man har en grupp av försäkringskontrakt definieras gruppens löptid som summan av de individuella kontraktens löptider.

• **Ett krav** uppstår när försäkringstagaren kräver ekonomisk kompensation för en oförutsedd händelse.

• **Den totala kravkostnaden** i en viss tariffcell är summan av alla individuella kravkostnader.

• För någon grupp av försäkringskontrakt i kraft under en viss tidsperiod, fås **kravfrekvensen** som antalet krav dividerat med löptiden. Detta resulterar i medeltalet för antalet krav under en viss tidsperiod, vanligtvis ett år.

• **Kravstorleken** är totala kravkostnaden dividerat med antalet krav, med andra ord genomsnittskostnaden per krav.

• **Riskpremie** är totala kravkostnaden dividerat med löptiden, alltså genomsnittskostnaden per försäkringsår (eller annan tidsperiod). Riskpremien kan därför skrivas om som produkten av kravfrekvensen och kravstorleken, eftersom:

$$\begin{aligned}\text{Riskpremie} &= \text{totala kravkostnaden/löptiden} \\ &= (\text{antalet krav/löptiden}) \times (\text{totala kravkostnaden/antalet krav}) \\ &= \text{kravfrekvensen} \times \text{kravstorleken}\end{aligned}$$

• **Förtjänad premie** är mängden av premieinkomster som täcker risken under en viss undersökt tidsperiod.

• **Förlustkvoten** är kravkostnaden dividerat med förtjänad premie. Relaterat till detta är kombinerade kvoten, som definieras som kravkostnaden plus administrativa kostnader dividerat med förtjänad premie.

Exempel 3.2. *Mopedförsäkring*
Vi tar ett exempel på mopedförsäkringar, vars data kommer från det svenska försäkringsbolaget Wasa, före dess fusion med Länsförsäkringar Alliance. Tariffen som Wasa använder baserar sig på tre värderingsfaktorer, som visas i tabellen nedan.

Värderingsfaktor	Klass	Klassbeskrivning
Fordonsklass	1	Vikt över 60 kg och mer än två växlar
	2	Övrig
Fordonets ålder	1	Åtminstone 1 år
	2	2 år eller mer
Geografisk zon	1	Centrala och semi-centrala delar av Sveriges tre största städer
	2	Förorter och medelstora städer
	3	Mindre städer, bortsett från de i 5 eller 7
	4	Småstäder och landsbygden, bortsett från de i 5 eller 7
	5	Norra städer
	6	Norra landsbygden
	7	Gotland

Detta exempel är taget från [1], som använder sig av programmet SAS för att göra estimeringar. Författarna (Ohlsson & Johansson) har på länken http://staff.math.su.se/esbj/GLMbook/moppe.sas pulicerat SAS-kod för läsaren att använda fritt, t.ex. erhålla data eller göra egna tester. Denna avhandling kommer dock att fokusera på att använda programmet R, så första uppgiften blir att konvertera SAS-koden till R-format, för framtida bruk.

```
webb <- url("http://www2.math.su.se/~esbj/GLMbook/moppe.sas")
rdata <- readLines(webb, n = 200L, warn = FALSE, encoding = "unknown")
close(webb)
## Hittar dataintervallet (rdata)
## ^cards är ett uttyck på länken ovan som är det sista som står
```

```
## före tabellen, programmet börjar hämta data från och med
## raden efter (1L). Sista biten är från "^;", tar allt därefter.
## Sen skapas tabell1.2 och kolumnnamnen fylls i
rdata.start <- grep("^cards;", rdata) + 1L
rdata.end   <- grep("^;", rdata[rdata.start:999L]) + rdata.start - 2L
tabell.1.2  <- read.table(text = rdata[rdata.start:rdata.end],
                          header = FALSE, sep = "", quote = "",
                          col.names = c("premiekl", "moptva", "zon", "dur",
                              "medskad", "antskad", "riskpre", "helpre", "cell"),
                          na.strings = NULL,
                          colClasses = c(rep("factor", 3), "numeric",
                              rep("integer", 4), "NULL"),
                          comment.char = "")
## (Rensar bort onödiga delar)
rm(webb, rdata, rdata.start, rdata.end)
comment(tabell.1.2) <-
   c("Titel: Partial casco moped insurance from Wasa insurance, 1994--1999",
      "Källa: http://www2.math.su.se/~esbj/GLMbook/moppe.sas",
      "Upphovsrätt: http://www2.math.su.se/~esbj/GLMbook/")
## Skapar en variabel; skadefrekvens
tabell.1.2$skadfre = with(tabell.1.2, antskad / dur)
## Sparar resultatet för senare användning
save(tabell.1.2, file = "tabell.1.2.RData")
## Skriver ut tabellen:
print(tabell.1.2)
```

	premiekl	moptva	zon	dur	medskad	antskad	riskpre	helpre	skadfre
1	1	1	1	62.9	18256	17	4936	2049	0.27027027
2	1	1	2	112.9	13632	7	845	1230	0.06200177
3	1	1	3	133.1	20877	9	1411	762	0.06761833
4	1	1	4	376.6	13045	7	242	396	0.01858736
5	1	1	5	9.4	0	0	0	990	0.00000000
6	1	1	6	70.8	15000	1	212	594	0.01412429
7	1	1	7	4.4	8018	1	1829	396	0.22727273
8	1	2	1	352.1	8232	52	1216	1229	0.14768532
9	1	2	2	840.1	7418	69	609	738	0.08213308
10	1	2	3	1378.3	7318	75	398	457	0.05441486
11	1	2	4	5505.3	6922	136	171	238	0.02470347
12	1	2	5	114.1	11131	2	195	594	0.01752848
13	1	2	6	810.9	5970	14	103	356	0.01726477
14	1	2	7	62.3	6500	1	104	238	0.01605136
15	2	1	1	191.6	7754	43	1740	1024	0.22442589
16	2	1	2	237.3	6933	34	993	615	0.14327855
17	2	1	3	162.4	4402	11	298	381	0.06773399
18	2	1	4	446.5	8214	8	147	198	0.01791713
19	2	1	5	13.2	0	0	0	495	0.00000000
20	2	1	6	82.8	5830	3	211	297	0.03623188
21	2	1	7	14.5	0	0	0	198	0.00000000
22	2	2	1	844.8	4728	94	526	614	0.11126894
23	2	2	2	1296.0	4252	99	325	369	0.07638889
24	2	2	3	1214.9	4212	37	128	229	0.03045518
25	2	2	4	3740.7	3846	56	58	119	0.01497046
26	2	2	5	109.4	3925	4	144	297	0.03656307
27	2	2	6	404.7	5280	5	65	178	0.01235483
28	2	2	7	66.3	7795	1	118	119	0.01508296

Figur 3.1.1: Tabell 1.2, härledd ur R

Kolumnen ”helpre” representerar premien som de facto användes i tariffen år 1999. Kravfrekvensen (skadfre) har även lagts till, som blir antalet krav (antskad) dividerat med löptiden (dur).

Nedan ser vi en tabell över vanliga nyckeltal:

Exponering ω	Respons X	Nyckeltal $Y = X/\omega$
Löptid	Antalet krav	Kravfrekvensen
Löptid	Kravkostnad	Riskpremie
Antalet krav	Kravkostnad	Genomsnittskravet
Införtjänade premier	Kravkostnad	Förlustkvot
Antalet krav	Antalet stora krav	Storkravsförhållandet

Samtliga nyckeltal är uppbyggda på samma sätt: ett förhållande mellan en stokastisk variabels utfall (X) och ett volymmått ω, som vi kallar för *exponering*. Observera att ω inte betraktas som en stokastisk variabel utan som en deterministisk. Den stokastiska variabeln X kallas för *responsen*.

3.2 Grundläggande modellantaganden

Här nedan listar vi några antaganden om försäkringskontrakt, som används för konstruerandet av en statistisk modell.

Antagande 3.3. *(Försäkringskontrakten är oberoende av varandra)*
Antag att vi har n stycken kontrakt, för någon responstyp i tabell 1 och låt X_i beteckna responsen för kontrakt i. Då är $X_1, X_2, ..., X_n$ oberoende.

Detta antagande kan ifrågasättas. Om till exempel två bilar som krockar med varandra har likadana försäkringskontrakt kan man inte påstå att X_i är oberoende. Vi förbiser dock detta problem då det är i en ganska liten skala som dylika olyckor inträffar samt för att uträkningarna blir lättare om vi antar att kontrakten är oberoende.

Antagande 3.4. *(Tidsoberoende)*
Antag att vi har k stycken disjunkta tidsintevall $T_i \equiv (t_{i-1}, t_i], i = 1, \ldots, k$. För någon responsstyp i tabell 1, låt X_{T_i} beteckna denna respons under tidsintervall $(t_{i-1}, t_i]$. Då är $X_{T_1}, X_{T_2}, ..., X_{T_k}$ oberoende.

Här kan vi tänka oss att X är antalet krav eller kravkostnaden, och att de då alltså är oberoende mellan olika tidsperioder. Detta antagande kan också ifrågasättas en aning: om någon nyligen haft inbrott kanske hen installerar ett inbrottsalarm, eller om en bilförare krockar så kör hen mer försiktigt i fortsättningen. I allmänhet kan man dock utgå från att tidsoberoende är ett rimligt antagande. Man vill dock kanske argumentera för att X skulle minska med tiden (färre krav eller lägre kravkostnader).

Antagande 3.5. *(Homogenitet)*
Antag att vi har två kontrakt i samma tariffcell, med samma responstyp och samma exponering ω. För någon responsstyp i tabell 1, låt X_i beteckna responsen för kontrakt i. Då följer X_1 och X_2 samma sannolikhetsfördelning.

Antagandet om homogenitet är inte heller alltid uppfyllt. I praktiken försöker ett försäkringsbolag placera försäkringskontrakten i *relativt* homogena grupper och kräva samma premie inom samma tariffcell. Detta är ofta rätt svårt, men ju fler värderingsfaktorer som beaktas, desto jämnare grupper erhålls.

Ett sätt att hantera icke-homogenitet inom samma tariffcell är olika *bonussystem* för privatpersoner eller *erfarenhetsrankingar* för stora företag.

Antagande 3.5 påstår dessutom att två försäkringskontrakt som är lika långa men som är i kraft under olika tidsperioder följer samma fördelning. Dessa kan såvida vara kontrakt som är i kraft direkt efter varandra av samma person för samma försäkrade objekt; en så kallad *förnyelseprocess*. Antagandet antyder alltså att fördelningen är likadan under olika säsonger, vilket inte känns särskilt trovärdigt (bilförsäkring under vintern jämfört med under sommaren). Försäkringsbolag sätter dock ofta en löptid på *hela år* för kontrakten vilket eliminerar detta problem.

Då man betraktar fallet att X är kravstorleken, kan det uppstå variationer på grund av till exempel inflation. Detta åtgärdas ofta genom att räkna om kravkostnaderna till nuvarande priser via något index. Överlag påverkar inte olika trender premienivån, så länge förhållandet mellan de olika tariffcellerna hålls stabilt över åren.

3.2.1 Väntevärden och varianser

I en tariffanalys försöker man alltid att skatta väntevärdet av variabeln i en tariffcell, men för att få korrekta estimat behöver vi också veta en hel del om *varianserna*. Med hjälp av antaganden 3.3 - 3.5 ovan kan vi undersöka dessa mått för olika responser och nyckeltal i tabell 1. Betrakta ett godtyckligt nyckeltal $Y = X/\omega$ för en viss kontraktgrupp i en tariffcell med exponering ω och respons X. Notera här att X är en stokastisk variabel, medan ω *inte* är det.

Vi utgår först från situationen där ω är antalet krav, vi kan då skriva om X som summan av ω stycken individuella responser $Z_1, ..., Z_\omega$. Om t.ex. X är kravkostnaden blir alltså Z_k kravkostnaden för det k:te kravet. Observera att detta inte är samma sak som totala kravkostnaden, eftersom vi undersöker en kontraktgrupp, som endast är en delmängd av en tariffcell. Vi har från antagande 3.3 - 3.4 att de olika Z_k är oberoende, eftersom kraven kommer att tillhöra olika kontrakt eller komma från olika tidsintervall. Antagande 3.5 ger dessutom att identiska fördelningar gäller för de olika Z_k, ur vilket erhålls att $\mathbb{E}(Z_k) = \mu$ och $\text{Var}(Z_k) = \sigma^2$, de beror alltså inte av k. Från detta fås att väntevärdet och variansen för responsen respektive nyckeltalet blir:

$$E(X) = \omega\mu \text{ och } \text{Var}(X) = \omega\sigma^2, \text{ samt} \tag{3.2.1}$$

$$E(Y) = \mu \text{ och } \text{Var}(Y) = \sigma^2/\omega \tag{3.2.2}$$

Variansen i (3.2.2) följer av att:

$$\mathrm{Var}(Y) = \mathrm{Var}(X/\omega) = (1/\omega^2)\mathrm{Var}(X) = (1/\omega^2)\omega\sigma^2 = \sigma^2/\omega$$

I följande lemma förutsätter vi att dessa resultat gäller även för nyckeltal vars exponering är löptid eller förtjänad premie.

Lemma 3.6. Under antaganden 3.3 - 3.5, om X är någon respons i tabell 1 med $\omega > 0$ och $Y = X/\omega$, ges väntevärdet och variansen för X respektive Y enligt ekvationer (3.2.1) - (3.2.2). Här är μ och σ^2 väntevärdet respektive variansen för en respons med exponering $\omega = 1$.

Bevis. Vi har redan sett att resultatet håller när ω är antalet krav. Vi har två fall utöver att ω är antalet krav: löptid och förtjänad premie. Vi antar nu att ω är ett rationellt tal, $\omega = m/n$, och vi kan då dela upp exponeringen i m delar som har storleken $1/n$. Om vi har fallet med löptid delar vi upp m i lika stora tidsintervall, om ω är förtjänad premie gör vi tidsintervallen tillräckligt långa för att få premien till $1/n$ var, under antagandet att premien betalas kontinuerligt över tiden. Responserna i dessa tidsintervall är benämnda $Z_1, ..., Z_m$, vilket är en samling av oberoende och identiskt fördelade stokastiska variabler. Om vi lägger tilll n stycken sådana responser Z_k får vi en variabel Z med exponering ω. Antagande 3.5 ger att alla dylika Z har har samma väntevärde och varians och vi kan därför skriva $\mathbb{E}(Z) = \mu$ och $\mathrm{Var}(Z) = \sigma^2$. Härifrån får vi att $\mathbb{E}(Z) = n\mathbb{E}(Z_1)$ och $\mathrm{Var}(Z) = n\mathrm{Var}(Z_1)$, vilket medför att $\mathbb{E}(Z_k) = \mathbb{E}(Z_1) = \mu/n$ och $\mathrm{Var}(Z_k) = \mathrm{Var}(Z_1) = \sigma^2/n$. Eftersom $X = \sum_{k=1}^{m} Z_k$ fås att $\mathbb{E}(X) = \mu m/n$ och $\mathrm{Var}(X) = \sigma^2 m/n$, alltså gäller (3.2.1) och därifrån följer (3.2.2).

Detta bevisar lemmat när exponeringen är ett rationellt tal. Vi tillägger utan bevis, att för ett godtyckligt $\omega > 0$ bör en rationell approximation användas. □

Ett resultat av detta lemma är att vi alltid borde använda viktade varianser i alla modeller för nyckeltal.

3.3 Multiplikativa modeller

En tariffanalys baseras ofta på försäkrarens egna data. Om ett försäkringsbolag skulle ha tillräckligt med data om kraven i varje tariffcell så skulle man kunna bestämma premien för cellen genom att estimera den förväntade kostnaden för

den observerade riskpremien. Detta är i praktiken sällan fallet eftersom vissa celler ibland har obetydliga eller icke-existerande data. Det är därför nödvändigt att utveckla modeller som kan skatta en riskpremie som varierar mer jämnt över cellerna, med god skattning av cellestimaten. Det är dessutom viktigt att premierna är relativt stabila över tid och att de inte påverkas av stora slumpmässiga fluktueringar.

Vi återgår till att diskutera värderingsfaktorerna för olika försäkringskontrakt. Vi antar för enkelhetens skull att vi har två värderingsfaktorer ($M = 2$) och benämner därför en tariffcell med (i, j), där i och j betecknar klassen för den första respektive den andra värderingsfaktorn. Vi kan nu beteckna exponeringen för cell (i, j) med ω_{ij} och responsen med X_{ij}, vilket ger oss nyckeltalet $Y_{ij} = X_{ij}/\omega_{ij}$. Enligt lemma 3.6 har vi att $\mathbb{E}(Y_{ij}) = \mu_{ij}$, där μ_{ij} är väntevärdet under enhetsexponeringen $\omega_{ij} = 1$.

Definition 3.7. En *multiplikativ modell* (med två relativiteter) är av formen

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}, \tag{3.3.1}$$

där $\{\gamma_{1i}; i = 1, ..., m_1\}$ och $\{\gamma_{2j}; j = 1, ..., m_2\}$ kallas för relativiteter och motsvarar parametrarna för de olika klasserna för värderingsfaktor 1 respektive värderingsfaktor 2, medan γ_0 kallas för ett basvärde (se nedan). Som tidigare nämnt representerar m_1 och m_2 antalet klasser för respektive värderingsfaktor.

För att få parametrarna unika börjar vi med att specifiera en sorts referenscell, som vi kallar för *bascell*. Vi sätter cellen $(1, 1)$ lika med bascellen, då får vi att $\gamma_{11} = \gamma_{21} = 1$. Nu kan γ_0 tolkas som ett sorts basvärde, som blir lika med nyckeltalet för bascellen. De andra parametrarna mäter nu den relativa skillnaden i förhållande till bascellen. Till exempel, om $\gamma_{12} = 1,25$ är väntevärdet för cell $(2, 1)$ 25 % högre än i cell $(1, 1)$, på motsvarande sätt är väntevärdet för cell $(2, 2)$ 25 % högre än för cell $(1, 2)$. Vi förtydligar med en tabell nedan:

Faktor 1	Faktor 2
$\mu_{11} = \gamma_0$	$\mu_{21} = 1,25\gamma_0$
$\mu_{12} = \gamma_0\gamma_{22}$	$\mu_{22} = \gamma_0\gamma_{12}\gamma_{22} = 1,25\mu_{12}$

Detta antagande om multiplikativitet innebär att det inte existerar något samband mellan de två värderingsfaktorerna. I praktiken betyder det att om

vi analyserar värderingsfaktor 1 för något nyckeltal, är förhållandet mellan två klasser samma oberoende av de andra värderingsklasserna. Till exempel om värderingsfaktor 1 är åldersklass, värderingsfaktor 2 är geografisk zon och nyckeltalet är riskpremie, är förhållandet mellan två olika åldersklasser samma inom samtliga geografiska zoner.

Orsaken till att multiplikativa modeller är att föredra är för att de ger ett rättvist förhållande mellan klasser i en given värderingsfaktor. Om en ökning på 20 % skulle ske för någon värderingsfaktor skulle själva nyckeltalet (t.ex. riskpremie) också öka med 20 %. Om vi hade en additiv modell av formen $\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}$ skulle nyckeltalet öka med en fix mängd (20 % av den värderingsfaktorn som höjs), vilket skulle kunna vara orättvist mellan t.ex. olika regioner.

Den multiplikativa modellen med $M = 2$ kan lätt utvidgas till fallet med godtyckligt många värderingsfaktorer enligt:

$$\mu_{i_1,i_2,\ldots,i_M} = \gamma_0\gamma_{1i_1}\gamma_{2i_2}\ldots\gamma_{Mi_M}. \tag{3.3.2}$$

Den multiplikativa modellen är användbar vid bestämmandet av hur den krävda premien borde delas upp bland kontrakten. Den allmänna nivån justeras genom basvärdet γ_0, medan de andra parametrarna kontrollerar hur mycket som ska tas betalt för ett kontrakt, givet detta basvärde. I praktiken bestäms relativiteterna γ_{ki_k} före basvärdet.

Kapitel 4

Grunderna i prissättning med GLM

Målet med en tariffanalys är att bestämma hur ett eller flera nyckeltal $Y_i, i = 1, \ldots, n$, varierar med ett antal värderingsfaktorer. Detta är kopplat till att analysera hur den beroende variabeln Y_i varierar med kovariaterna (de oberoende variablerna) $X_{ij}, j = 1, ..., p$ i en multipel linjär regression. Vi erinrar oss den generella linjära modellen i definitionen nedan.

Definition 4.1. En generell linjär modell är av formen

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i, \tag{4.0.1}$$

där Y_i är modellerad genom en linjär funktion av förklarande variabler $X_{ij}, j = 1, \ldots, p$, samt en felterm $\epsilon_i, i = 1, \ldots, n$, där ϵ_i är stokastiska variabler som representerar felen i ekvation (4.0.1), se [2].

"Generell" betyder i det här sammanhanget att Y_i kan bero av fler än en förklarande variabel. Vi antar att feltermerna ϵ_i är oberoende och ofta antas de vara identiskt fördelade enligt $\epsilon_i \sim N(0, \sigma^2)$.

Linjär regression, eller den mer omfattande generella linjära modellen, är dock inte lämplig i icke-livsförsäkringsprissättning av två orsaker:

1) Om Y_i:s definitionsmängd inte är hela $\mathbb{R}$, t.ex. om den är diskret, håller inte antagandet att feltermen är normalfördelad. Om Y_i är antalet krav följer den en diskret fördelning på de icke-negativa heltalen, dessutom är krav-kostnader icke-negativa och brukar vara skeva åt höger.

2) I linjära modeller är väntevärdet en linjär funktion av kovariaterna, dock har vi konstaterat att multiplikativa modeller är mer rimliga vid försäkringsprissättning. Dessutom kan det blir problematiskt om variansen för Y_i beror av väntevärdet, som inte beaktas i generella linjära modeller.

Generaliserade linjära modeller (GLMs) är en rik klass av statistiska verktyg, som generaliserar den generella linjära modellen i två riktningar, som båda tar hand om de ovan nämnda problemen:

- *Sannolikhetsfördelningar:* I stället för att utgå från att responsen Y_i är normalfördelad fungerar GLMs med en generell klass av fördelningar, i synnerhet normal-, Poisson- och gammafördelningarna.
- *Modell för väntevärdet:* I linjära modeller är väntevärdet en linjär funktion av väntevärden av kovariaterna X_{ij}. I GLMs är någon monoton transformation av väntevärdet en linjär funktion av väntevärden av X_{ij}:na, där den linjära och multiplikativa modellen är specialfall.

Anmärkning 4.2. *Om en stokastisk variabel skrivs med indexet i är det hädanefter underförstått att $i = 1, \ldots, n$, om inget annat anges.*

4.1 Exponentiella fördelningsfamiljen

Ett viktigt verktyg inom GLM-teorin är funktioner som tillhör *exponentiella fördelningsfamiljen* (exponential family of distributions, mer kända som EDM:s, exponential dispersion models), vilka generaliserar normalfördelningen som används i linjära modeller. De flesta statistiska fördelningarna som normal-, binomial-, och Poissonfördelningarna tillhör exponentiella fördelningsfamiljen.

Definition 4.3. De stokastiska variablerna $Y_i, i = 1, \ldots, n$ hör till samma EDM-klass om frekvensfunktionen av Y_i, som i det kontinuerliga fallet är en täthetsfunktion och i det diskreta fallet en elementär sannolikhet ges av

$$f_{Y_i}(y; \theta_i, \phi, \omega_i) = \begin{cases} \exp\big(\dfrac{y\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y, \phi, \omega_i)\big), & \text{om } y \text{ är ett möjligt utfall till } Y_i; \\ 0 & \text{annars,} \end{cases} \tag{4.1.1}$$

Härvid antas att

- θ_i, den kanoniska parameterna, beror av i och tillhör en öppen delmängd av $\mathbb{R}$.
- $\phi > 0$, spridningsparametern, beror inte av i.
- $b(\cdot)$, den s.k. kumulantfunktionen, beror inte av i och är två gånger kontinuerligt deriverbar med inverterbar första derivata.
- $c(\cdot, \cdot, \cdot)$ är en (utfyllnings)funktion som inte beror av θ_i.

Exempel 4.4. Låt Y vara en normalfördelad stokastisk variabel, $Y \sim N(\mu, \sigma^2)$. Då kan dess täthetsfunktion skrivas som

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{\sigma^2} \right) \\ &= \exp\big(\frac{y_i\mu - \mu^2/2}{\sigma^2} + c(y, \sigma^2)\big), \end{aligned}$$

där vi har sorterat ut delen av tätheten som inte beror på μ,

$$c(y, \sigma^2, \omega) = -\frac{1}{2}\big(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\big).$$

Detta är en EDM med $\theta = \mu$ och $b(\theta) = \theta^2/2$. Härav följer att normalfördelningen är en EDM. Det går även bra att vikta variansen med avseende på en exponering ω, i fallet ovan antas ω vara lika med 1.

Exempel 4.5. På samma sätt som i det föregående exemplet undersöker vi hur Poissonfördelningen tillhör EDM-familjen. Låt Y vara en Poissonfördelad stokastisk variabel, $Y \sim \text{Po}(\mu)$. Frekvensfunktionen är, för icke-negativa heltal y:

$$\begin{aligned} f_Y(y; \mu) = P(Y = y) &= e^{-\mu}\frac{\mu^y}{y!} \\ &= \exp\{y\log(\mu) - \mu + c(y)\}, \end{aligned}$$

där $c(y) = -\log(y!)$. Detta är en EDM, vilket kan ses genom att substituera $\theta = \log(\mu)$,

$$f_Y(y; \theta) = \exp\{y\theta - e^\theta + c(y)\}. \tag{4.1.2}$$

Detta är formen given i (4.1.1), med $\phi = 1$ och kumulantfunktionen $b(\theta) = e^\theta$. I och med att detta är det oviktade fallet är $\omega = 1$. Parameterrummet är $\mu > 0$, med andra ord den öppna mängden $-\infty < \theta < \infty$.

Vi fortsätter att undersöka Poissonfördelningens tillämpningar inom EDM-teorin. Låt $N(t)$ vara antalet krav för ett individuellt kontrakt under tidsintervallet $[0, t]$, med $N(0) = 0$. Den stokastiska processen $\{N(t), t \geq 0\}$ kallas för kravprocessen. Under antaganden 3.4 - 3.5 samt ett antagande att krav inte klustrar sig är kravprocessen en *Poissonprocess*, se definition 2.6. Detta medför att vi kan anta att antalet krav för ett individuellt kontrakt under en viss tidsperiod är Poissonfördelade. Tack vare antagande 3.3 (oberoende mellan kontrakten) får vi även en Poissonfördelning för en grupp av försäkringskontrakt i samma tariffcell. Detta leder in oss på en transformation av den vanliga Poissonfördelningen.

Exempel 4.6. Låt X_i vara antalet krav i en tariffcell med löptid ω_i och låt μ_i beteckna väntevärdet när $\omega_i = 1$. Enligt lemma 3.6 har vi att $\mathbb{E}(X_i) = \omega_i \mu_i$, så X_i följer en Poissonfördelning med frekvensfunktion

$$f_{X_i}(x_i; \mu_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, ... \tag{4.1.3}$$

Denna transformation av Poissonfördelningen brukar kallas för den *relativa Poissonfördelningen.*

Den relativa Poissonfördelningen används ofta för att skatta fördelningen för kravfrekvensen $Y_i = X_i/\omega_i$ (X_i = antalet krav, ω_i = löptid). Den relativa Poissonfördelningen används hellre istället för den vanliga Poissonfördelningen, eftersom den inte tillräckligt bra kan skatta Y. Detta är eftersom Poissonfördelningen har de hela talen som värdemängd, medan kravfrekvensen kan anta reella tal.

Frekvensfunktionen är, för sådana y_i för vilka $\omega_i y_i$ är ett icke-negativt heltal:

$$\begin{aligned} f_{Y_i}(y_i; \mu_i) &= P(Y_i = y_i) = P(X_i = \omega_i y_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i y_i}}{(\omega_i y_i)!} \\ &= \exp\{\omega_i [y_i \log(\mu_i) - \mu_i] + c(y_i, \omega_i)\}, \end{aligned} \tag{4.1.4}$$

där $c(y_i, \omega_i) = \omega_i y_i \log(\omega_i) - \log(\omega_i y_i!)$. Detta är en EDM, vilket kan ses genom att substituera $\theta_i = \log(\mu_i)$,

$$f_{Y_i}(y_i; \theta_i) = \exp\{\omega_i (y_i \theta_i - e^{\theta_i}) + c(y_i, \omega_i)\}. \tag{4.1.5}$$

Detta är formen given i (4.1.1), med $\phi = 1$ och kumulantfunktionen $b(\theta_i) = e^{\theta_i}$. Parameterrummet är $\mu_i > 0$, med andra ord den öppna mängden $-\infty < \theta_i < \infty$.

Lemma 4.7. Låt Y_1 och Y_2 vara kravfrekvensen i två celler med exponeringar (löptider) ω_1 respektive ω_2, och låt båda följa en relativ Poissonfördelning med

parametern μ. Om dessa två celler sammanslås kommer kravfrekvensen i den nya cellen att bli det viktade medelvärdet

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2} = \frac{X_1 + X_2}{\omega_1 + \omega_2},$$

där X_1 och X_2 är antalet krav.

BEVIS. Eftersom $\omega_1 Y_1 + \omega_2 Y_2$ är summan av två oberoende Poissonfördelade stokastiska variabler är den själv Poissonfördelad, detta enligt sats 2.12. Därav fås att Y följer en relativ Poissonfördelning med exponering $\omega_1+\omega_2$ och parameter μ. □

En fördelning som är sluten under denna typ av medelvärdesstruktur kallas för *reproduktiv*. Detta är ett naturligt krav för samtliga EDMs.

Vi tittar härnäst på hur man kan modellera kravstorleken för krav i respektive tariffcell. Vi tar och lämnar bort indexet i för enkelhetens skull. Exponeringen för kravstorlek, som är antalet krav, är därför skriven som ω. I dessa analyser kommer vi att betinga på antalet krav så att exponeringens vikt är icke-slumpmässig, som sig bör. Idén är att vi först analyserar kravfrekvens med antalet krav som utfallet för en stokastisk variabel; när detta är gjort betingar vi på antalet krav för analyserandet av kravstorlek. Här är totala kravkostnaden i cellen X och kravkostnaden är $Y = X/\omega$.

Vilken fördelning är optimal för att modellera kravstorleken? Svaret är inte lika klart som i fallet med kravfrekvensen, som man kan tänka sig är Poissonfördelad. I detta fall borde fördelningen vara positiv och ha en positiv skevhet (vara skev åt höger), vilket alltså eliminerar normalfördelningen. En fördelning som blivit populär inom GLM-analys är gammafördelningen. Som vi visar i avsnitt 4.3.2 antyder gammafördelningen att standardavvikelsen är proportionell mot väntevärdet μ, vilket innebär att vi har en konstant variationskoefficient (se definition **??**). Med tanke på att det är kravstorleken vi försöker modellera är detta till fördel.

Vi antar för stunden att kostnaden för ett individuellt krav är gammafördelad; detta är fallet då $\omega = 1$. För en gammafördelad stokastisk variabel X har vi, för indexparametern $\alpha > 0$ och skalaparametern $\beta > 0$, frekvensfunktionen

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}; \quad x > 0, \tag{4.1.6}$$

detta enligt defintion 2.7. Väntevärdet för denna fördelning är som sagt α/β och variansen är α/β^2. Enligt sats 2.8 är summan för oberoende gammafördelade

stokastiska variabler också gammafördelad, med samma skalaparameter och en indexparameter som är summan av de individuella indexparametrarna. Så om X är summan av ω stycken oberoende gammafördelade stokastiska variabler, konstaterar vi att $X \sim \text{Gamma}(\omega\alpha, \beta)$. Frekvensfunktionen för $Y = X/\omega$ är då

$$f_Y(y) = \omega f_X(\omega y) = \frac{(\omega\beta)^{\omega\alpha}}{\Gamma(\omega\alpha)} y^{\omega\alpha-1} e^{-\omega\beta y}; \quad y > 0, \tag{4.1.7}$$

så $Y \sim \text{Gamma}(\omega\alpha, \omega\beta)$ med väntevärde α/β. Innan vi transformerar denna fördelning på EDM-form, är det fördelaktigt att omparametrisera den genom $\mu = \alpha/\beta$ och $\phi = 1/\alpha$.
Det nya parameterrummet är givet av $\mu > 0$ och $\phi > 0$. Frekvensfunktionen är

$$\begin{aligned} f_Y(y) = f_Y(y; \mu, \phi) &= \frac{1}{\Gamma(\omega/\phi)} \left(\frac{\omega}{\mu\phi}\right)^{\omega/\phi} y^{(\omega/\phi)-1} e^{-\omega y/(\mu\phi)} \\ &= \exp\left\{\frac{-y/\mu - \log(\mu)}{\phi/\omega} + c(y, \phi, \omega)\right\}; \quad y > 0, \end{aligned} \tag{4.1.8}$$

där $c(y, \phi, \omega) = \log(\omega y/\phi)\omega/\phi - \log(y) - \log\Gamma(\omega/\phi)$. Vi har att $\mathbb{E}(Y) = \omega\alpha/(\omega\beta) = \mu$ och $\text{Var}(Y) = \omega\alpha/(\omega\beta)^2 = \phi\mu^2/\omega$, vilket stämmer överens med lemma 3.6.

För att visa att gammafördelningen är en EDM ändrar vi den första parametern i (4.1.7) till $\theta = -1/\mu$; den nya parametern tar värden i den öppna mängden $\theta < 0$. Om vi återgår till att indexera parametrarna med index i, medan ϕ fortfarande är oberoende av i, fås frekvensfunktionen för kravstorleken Y_i enligt

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i/\theta_i + \log(-1/\theta_i)}{\phi/\omega_i} + c(y_i, \phi, \omega_i)\right\}. \tag{4.1.9}$$

Vi konstaterar att dessa gammafördelningar bildar en EDM-klass med $b(\theta_i) = -\log(-\theta_i)$, och därav kan vi använda den i en GLM.

Anmärkning 4.8. *Man kan börja tänka sig att vilken fördelning som helst kan tänkas vara en EDM med rätt parametrisering. Så är dock inte fallet; lognormalfördelningen kan inte på något sätt konstrueras som en EDM. Observera att funktionen c* ***inte*** *får bero av θ!*

I kölvattnen av dessa exempel är det lämpligt att ta upp ett bekant matematiskt verktyg, nämligen den kumulantgenererande funktionen $\Psi(\cdot)$, se definition 2.3. Kumulantfunktionen $b(\cdot)$ i en EDM har fått sitt namn från sambandet som den har till den kumulantgenererande funktionen, som vi kommer att se i följande sats.

Sats 4.9. *Den kumulantgenererande funktionen $\Psi(t)$ för en EDM existerar och ges av*

$$\Psi(t) = \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega}. \tag{4.1.10}$$

BEVIS. Vi har från definition 2.2 att $M_Y(t) = \mathbb{E}(e^{tY})$, om funktionen är ändlig för varje t så att $|t| < \delta$ för något $\delta > 0$. Vi har för enkelhets skull lämnat bort indexet i för att undvika för många parametrar. För kontinuerliga EDM:s får vi, enligt (4.1.1):

$$\begin{aligned}
\mathbb{E}(e^{tY}) &= \int_{-\infty}^{\infty} e^{ty} f_Y(y;\theta,\phi)dy \\
&= \int_{-\infty}^{\infty} \exp\left\{\frac{y\cdot(\theta+t\phi/\omega)-b(\theta)}{\phi/\omega} + c(y,\phi,\omega)\right\} dy \\
&= \int_{-\infty}^{\infty} \exp\left\{\frac{y\cdot(\theta+t\phi/\omega)-b(\theta)+b(\theta+t\phi/\omega)-b(\theta+t\phi/\omega)}{\phi/\omega} + c(y,\phi,\omega)\right\} dy \\
&= \exp\left\{\frac{b(\theta+t\phi/\omega)-b(\theta)}{\phi/\omega}\right\} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{\frac{y\cdot(\theta+t\phi/\omega)-b(\theta+t\phi/\omega)}{\phi/\omega} + c(y,\phi,\omega)\right\} dy.
\end{aligned}$$

Vi har sedan tidigare antagit att parameterrummet för θ i en EDM måste vara öppet. Härav följer att för varje t i närheten av 0 att $\theta + t\phi/\omega$ är i parameterrummet. Detta medför att den sista integralen är lika med 1 och att den föregående termen representerar den momentgenererande funktionen, som alltså existerar för varje t tillräckligt nära 0. Till sist tar vi logaritmen av uttrycket och erhåller högerledet i (4.1.9). I det diskreta fallet byter vi ut integralerna mot summor (och indexerar med avseende på Y:s utfallsrum) och räknar på liknande sätt. □

Exempel 4.10. Vi verifierar att sats 4.8 gäller för normalfördelningen. Den kumulantgenererande funktionen för en normalfördelad stokastisk variabel Y ges av

$$\Psi_Y(t) = \mu t + \frac{\sigma^2 t^2}{2},$$

eftersom $M_Y(t) = e^{\mu t} e^{\sigma^2 t^2/2}$. Från exempel 4.4 har vi att $b(\theta) = \theta^2/2$. Vi sätter

in detta i (4.1.9) och erhåller

$$\begin{aligned}\Psi_Y(t) &= \frac{b(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega} = \frac{\frac{(\theta+t\phi/\omega)^2}{2} - \frac{\theta^2}{2}}{\phi/\omega} \\ &= \frac{\frac{\theta^2}{2} + \frac{\theta t\phi}{\omega} + \frac{t^2\phi^2}{2\omega^2} - \frac{\theta^2}{2}}{\phi/\omega} = \frac{\theta t\phi}{\omega} \cdot \frac{\omega}{\phi} + \frac{t^2\phi^2}{2\omega^2} \cdot \frac{\omega}{\phi} \\ &= \theta t + \frac{\phi}{\omega} \cdot \frac{t^2}{2} = \mu t + \sigma^2 \cdot \frac{t^2}{2},\end{aligned}$$

då vi drog slutsatsen i exemplet innan att $\theta = \mu$ och $\phi/\omega = \sigma^2$.

Vi använder resultaten ovan för att härleda väntevärdet för en EDM via första kumulanten (vi antar som sagt att $b(\cdot)$ är två gånger deriverbar):

$$\Psi'(t) = b'(\theta + t\phi/\omega) \text{ och } \mathbb{E}(Y) = \Psi'(0) = b'(\theta) \tag{4.1.11}$$

Andra kumulanten, variansen, ges av

$$\Psi''(t) = b''(\theta + t\phi/\omega)\phi/\omega \text{ och } \mathrm{Var}(Y) = \Psi''(0) = b''(\theta)\phi/\omega. \tag{4.1.12}$$

I allmänhet är det mer behändigt att betrakta variansen som en funktion av väntevärdet μ. Vi har just sett att $\mu = \mathbb{E}(Y) = b'(\theta)$, och eftersom vi har antagit att b' är en inverterbar funktion, kan vi skriva det inversa sambandet enligt $\theta = b'^{-1}(\mu)$. Detta kan vi nu sätta in i $b''(\theta)$, vilket leder in oss på nästa definition.

Definition 4.11. Låt Y vara en stokastisk variabel som följer en EDM-fördelning med $\mathbb{E}(Y) = \mu$. Då definieras Y:s *variansfunktion* enligt

$$\nu(\mu) = b''(b'^{-1}(\mu)). \tag{4.1.13}$$

Vi kan alltså uttrycka $\mathrm{Var}(Y)$ som en produkt av variansfunktionen $\nu(\mu)$ och en (viktad) spridningsparameter ϕ/ω. Nedan följer ett par exempel på variansfunktioner.

Exempel 4.12. Variansfunktionen för den relativa Poissonfördelningen
Låt Y_i vara en stokastisk variabel som följer en relativ Poissonfördelning, alltså $Y_i \sim \mathrm{Po}(\omega_i\mu_i)$. Då blir kumulantfunktionen $b(\theta_i) = \exp(\theta_i)$ (se exempel 4.1), vilket ger att $\mu_i = b'(\theta_i) = \exp(\theta_i)$. Dessutom har vi att $b''(\theta_i) = \exp(\theta_i) = \mu_i$, alltså blir $\nu(\mu_i) = \mu_i$ och $\mathrm{Var}(Y_i) = \mu_i\omega_i$, eftersom $\phi = 1$. I fallet då $\omega_i = 1$ fås det kända resultatet att Poissonfördelningen har en varians som är lika med väntevärdet.

Exempel 4.13. Variansfunktionen för Gammafördelningen
Låt $Y_i \sim \Gamma(\omega_i\alpha, \omega_i\beta)$, där vi omparametriserar fördelningen enligt $\mu_i = \alpha/\beta$ och $\phi = 1/\alpha$, se stycket efter (4.1.5). Den kanoniska parametern fås som $\theta_i = -1/\mu_i$ och $b(\theta_i) = -\log(-\theta_i)$. Nu fås att $b'(\theta_i) = -1/\theta_i$ och $b''(\theta_i) = 1/\theta_i^2 = \mu_i^2$, vilket medför att $\nu(\mu) = b''(b'^{-1}(\mu)) = b''(\theta_i) = \mu_i^2$. Till sist får vi $\mathrm{Var}(Y_i) = \phi\mu_i^2/\omega_i$, som alltså är produkten av variansfunktionen och skal- och viktfaktorn.

Vi sammanfattar några variansfunktioner i tabellen nedan:

Fördelning	Normal	Poisson	Gamma	Binomial
$\nu(\mu)$	1	μ	μ^2	$\mu(1-\mu)$

Vi sammanfattar resultaten med följande lemma.

Lemma 4.14. Antag att Y_i tillhör en EDM-klass, med frekvensfunktionen angiven i (4.1.1). Då existerar dess kumulantgenererande funktion för varje t så att $|t| < \delta$ för något $\delta > 0$, och ges av

$$\Psi(t) = \frac{b(\theta_i + t\phi/\omega_i) - b(\theta_i)}{\phi/\omega_i}, \tag{4.1.14}$$

och det gäller att

$$\mu_i = \mathbb{E}(Y_i) = b'(\theta_i);$$

$$\mathrm{Var}(Y_i) = \phi\nu(\mu_i)/\omega_i,$$

där variansfunktionen $\nu(\mu_i)$ ges av $\nu(\mu_i) = b''(b'^{-1}(\mu_i))$.

Sats 4.15. *En familj av sannolikhetsfördelningar som tillhör en EDM-klass är entydigt bestämd av dess variansfunktion.*

Resultatet av denna sats är att om man bestämt sig för att använda en GLM, och därför också en EDM, behöver man bara bestämma variansfunktionen eftersom man då vet sannolikhetsfördelningen inom EDM-klassen. Detta är ett väldigt nyttigt resultat, då modellerandet av endast väntevärdet och variansen är betydligt lättare än att specifiera en hel fördelning.

Kärnan i beviset av denna sats är att notera att eftersom $\nu(\cdot)$ är en funktion av derivatorna till $b(\cdot)$, kan den senare nämnda bestämmas från $\nu(\cdot)$ genom att lösa några differentialekvationer - men $b(\cdot)$ är allt vi behöver för att specifiera EDM-fördelningen i (4.1.1).

Sats 4.16. *EDMs är reproduktiva.*
Antag att vi har två oberoende stokastiska variabler Y_1 och Y_2, med vikter ω_1 respektive ω_2, från samma EDM-familj, alltså med samma kumulantfunktion $b(\cdot)$, väntevärde μ och spridningsparameter ϕ. De tillåts dock ha olika vikter ω_1 och ω_2. Då tillhör deras ω-vägda medelvärde $Y = (\omega_1 Y_1 + \omega_2 Y_2)/(\omega_1 + \omega_2)$ samma EDM-fördelning, men med vikten $\omega_\cdot = \omega_1 + \omega_2$.

Bevis. Vi använder oss av den kumulantgenererande funktionen och lemma 4.9. Det är givet att Y_1 och Y_2 har samma kumulantfunktion $b(\cdot)$, μ och ϕ. I och med att de har samma μ får vi att $\theta_1 = \theta_2 := \theta$, tack vare relationen $\mu = b'(\theta)$. Vi får nu den kumulantgeneranade funktionen för Y enligt

$$\begin{aligned}
\Psi_Y(t) &= \Psi_{\frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1+\omega_2}}(t) = \Psi_{\frac{\omega_1 Y_1}{\omega_1+\omega_2} + \frac{\omega_2 Y_2}{\omega_1+\omega_2}}(t) = \Psi_{Y_1}(\frac{\omega_1}{\omega_1+\omega_2}t) + \Psi_{Y_2}(\frac{\omega_2}{\omega_1+\omega_2}t) \\
&= \frac{b(\theta + t(\frac{\omega_1}{\omega_1+\omega_2})\phi/\omega_1) - b(\theta)}{\phi/\omega_1} + \frac{b(\theta + t(\frac{\omega_2}{\omega_1+\omega_2})\phi/\omega_2) - b(\theta)}{\phi/\omega_2} \\
&= \frac{\omega_1 b(\theta + \frac{t\phi}{\omega_1+\omega_2}) - \omega_1 b(\theta) + \omega_2 b(\theta + \frac{t\phi}{\omega_1+\omega_2}) - \omega_2 b(\theta)}{\phi} \\
&= \frac{(\omega_1+\omega_2) b(\theta + \frac{t\phi}{\omega_1+\omega_2}) - (\omega_1+\omega_2) b(\theta)}{\phi} \\
&= \frac{b(\theta + \frac{t\phi}{\omega_1+\omega_2}) - b(\theta)}{\phi/(\omega_1+\omega_2)},
\end{aligned}$$

vilket visar att Y tillhör samma EDM med vikten $\omega_\cdot = \omega_1 + \omega_2$, enligt lemma 4.14. □

Den praktiska betydelsen av denna sats är att om vi sammanslår två tariffceller, under antagandet att de har samma väntevärde, är vi fortfarande inom samma fördelningsfamilj. Detta är ett viktigt krav inom premieprissättning.

4.1.1 Tweediemodeller

I tillämpningar inom icke-livsförsäkring är det viktigt att arbeta med sannolikhetsfördelningar som är slutna under multiplikation med skalärer, alltså *skalinvarianta*. Låt $c > 0$ och Y vara en stokastisk variabel som tillhör en viss fördelningsfamilj. Om Y är skalinvariant tillhör även cY samma fördelningsfamilj. Denna egenskap är önskvärd om Y mäts i någon monetär enhet: om vi konverterar data från en valuta till en annan vill vi stanna inom samma fördelningsfamilj.

På samma sätt vill vi kunna undersöka t.ex. kravfrekvensen i procent eller promille och fortfarande tillhöra den givna fördelningsfamiljen. Vi kan konstatera att alla nyckeltal i tabell 1 kräver skalinvarians, förutom förhållandet av stora krav, där skala inte är relevant.

Det kan bevisas att de enda EDM:s som är skalinvarianta är de så kallade Tweediemodellerna (döpt efter statistikern och medicinska fysikern Maurice Tweedie, 1919-1996), vilka är definierade genom att ha en variansfunktion

$$\nu(\mu) = \mu^p, \tag{4.1.15}$$

för något p.

	Typ	Namn	Nyckeltal
$p < 0$	Kontinuerlig	-	-
$p = 0$	Kontinuerlig	Normal	-
$0 < p < 1$	Icke-existerande	-	-
$p = 1$	Diskret	Relativ Poisson	Kravfrekvens
$1 < p < 2$	Blandad, icke-negativ	Sammansatt Poisson	Riskpremie
$p = 2$	Kontinuerlig, positiv	Gamma	Kravstorlek
$2 < p < 3$	Kontinuerlig, positiv	-	Kravstorlek
$p = 3$	Kontinuerlig, positiv	Inverterad normal	Kravstorlek
$p > 3$	Kontinuerlig, positiv	-	Kravstorlek

Ovan ses en tabell över Tweediemodeller samt de nyckeltal som de bäst kan användas till. Vi har redan tittat på fallen då $p = 0, 1, 2$. Tweediemodeller med $p \geq 2$ används ofta som fördelningar för kravstorleken, speciellt $p = 2$, men även den inverterade normalfördelingen med $p = 3$.

Klassen av modeller $1 < p < 2$ representerar en så kallad sammansatt Poissonfördelning. Dessa byggs upp med hjälp av summan av N stycken oberoende identiskt fördelade stokastiska variabler $Z_i, i = 1, \ldots, N$, där N är Poissonfördelad. Om man låter N representera antalet krav och Z_i:na kravstorleken, vilket medför att vi kan anta att Z_i:na är gammafördelade, får vi summan av hela kravkostnaden X enligt

$$X = \begin{cases} \sum_{i=1}^{N} Z_i & \text{för } N > 0; \\ 0 & \text{för } N = 0, \end{cases}$$

där $Z_i \sim \text{Gamma}(\alpha, \beta)$, $N \sim \text{Po}(\lambda)$ och X är då sammansatt Poissonfördelad. Denna sammansatta Poissonfördelning är blandad (vilket innebär att den är varken helt diskret eller helt kontinuerlig) och har en kontinuerlig fördelning för de positiva reella talen.

Som tabellen antyder finns det ingen EDM för $0 < p < 1$. Negativa värden för p är dock möjliga och ger då en kontinuerlig fördelning på hela reella axeln, men den praktiska användningen av dessa inom försäkring är obefintlig.

Från och med nu undersöker vi bara fallet då $p \geq 1$, vilket täcker de viktigaste tillämpningarna. Vi börjar med att presentera den motsvarande kumulantfunktionen enligt:

$$b(\theta) = \begin{cases} e^{\theta} & \text{för } p = 1; \\ -\log(-\theta) & \text{för } p = 2; \\ -\frac{1}{p-2}\left[-(p-1)\theta\right]^{(p-2)/(p-1)} & \text{för } 1 < p < 2 \text{ och } p > 2 \end{cases}$$

Parameterrummet M_θ är

$$M_\theta = \begin{cases} -\infty < \theta < \infty & \text{för } p = 1; \\ -\infty < \theta < 0 & \text{för } p > 1; \end{cases}$$

Vi får derivatan $b'(\theta)$ enligt

$$b'(\theta) = \begin{cases} e^{\theta} & \text{för } p = 1; \\ \left[-(p-1)\theta\right]^{-1/(p-1)} & \text{för } p > 1, \end{cases}$$

med inversen

$$h(\mu) = \begin{cases} \log(\mu) & \text{för } p = 1; \\ -\frac{1}{p-1}\mu^{-(p-1)} & \text{för } p > 1. \end{cases}$$

Vi härleder uttrycken ovan. Vi har att variansfunktionen $\mu^p = \nu(\mu) = b''(b'^{-1}(\mu))$. Fallet $p = 1$ har vi redan konstaterat att gäller för Poissonfördelningen, med $b(\theta) = \exp(\theta)$. Vi intresserar oss för fallet $p > 1$. Om vi utgår från att

$$b'(\theta) = \left[-(p-1)\theta\right]^{-1/(p-1)}$$

och att

$$h(\mu) = b'^{-1}(\mu) = -\frac{1}{p-1}\mu^{-(p-1)}$$

fås att

$$b''(\theta) = \frac{-1}{p-1}[-(p-1)\theta]^{-1/(p-1)-1} = [-(p-1)]^{-p/(p-1)}\theta^{-p/(p-1)}.$$

Insättning av $b'^{-1}(\mu)$ på θ:s plats i ekvationen ovan ger

$$\begin{aligned} b''(b'^{-1}(\mu)) &= [-(p-1)]^{-p/(p-1)}\big(-\frac{1}{p-1}\mu^{-(p-1)}\big)^{-p/(p-1)} \\ &= [-(p-1)]^{-p/(p-1)}\big(-\frac{1}{p-1}\big)^{-p/(p-1)}\mu^{p} \\ &= \mu^p = \nu(\mu), \end{aligned}$$

och härledningen är klar.

4.2 GLM: konstruktion, definition och exempel

Vi har nu sett hur man generaliserar normalfördelningen till en EDM-klass och ger oss nu in på att generalisera ordinära linjära modeller.

Som vanligt undersöker vi det enkla fallet med två värderingsfaktorer, denna gång med två (faktor 1) respektive tre klasser (faktor 2). Vi betecknar väntevärdet för nyckeltalet i cell (i, j) med μ_{ij}, där första faktorn är i klass i och andra i klass j. Linjära modeller antar en *additiv modell*-struktur för väntevärdet:

$$\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}. \tag{4.2.1}$$

Som i avsnittet med multiplikativa modeller definierar vi en bascell, som kommer att få värdet 0 eftersom modellen är additiv. Logiskt är att välja $(1, 1)$ som bascell, då låter vi $\gamma_{11} = \gamma_{21} = 0$, så att $\mu_{11} = \gamma_0$. De andra parametrarna mäter då avvikelsen från denna cell. Vi skriver om modellen i listform genom att sortera cellerna $(1, 1); (1, 2); (1, 3); (2, 1); (2, 2); (2, 3)$ och döpa om parametrarna:

$$\beta_1 \doteq \gamma_0,$$

$$\beta_2 \doteq \gamma_{12},$$

$$\beta_3 \doteq \gamma_{22},$$

$$\beta_4 \doteq \gamma_{23}.$$

Observera att vi har endast fyra variabler, då $\gamma_{11} = \gamma_{12} = 0$. Härnäst introducerar vi så kallade *dummyvariabler* genom relationen

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

Tillsammans med de nya parametrarna blir väntevärden för cellerna som i följande tabell:

k	Tariffcell		μ_i
1	1	1	β_1
2	1	2	$\beta_1 + \beta_3$
3	1	3	$\beta_1 + \beta_4$
4	2	1	$\beta_1 + \beta_2$
5	2	2	$\beta_1 + \beta_2 + \beta_3$
6	2	3	$\beta_1 + \beta_2 + \beta_4$

Värden för dummyvariablerna i detta exempel presenteras i följande tabell:

k	Tariffcell		x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	1	1	1	0	0	0
2	1	2	1	0	1	0
3	1	3	1	0	0	1
4	2	1	1	1	0	0
5	2	2	1	1	1	0
6	2	3	1	1	0	1

Förhållandet mellan parametrarna och dummyvariablerna ses enkelt om man jämför tabellerna.

4.2.1 Länkfunktionen

Med variablerna från tabellerna kan den linjära modellen för väntevärdet skrivas som

$$\mu_i = \sum_{j=1}^{4} x_{ij}\beta_j \quad i = 1, 2, ..., 6, \tag{4.2.2}$$

eller på matrisform $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, där $\mathbf{X}$ kallas *designmatrisen*, och

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \\ x_{61} & x_{62} & x_{63} & x_{64} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}. \quad (4.2.3)$$

Vi har här nu presenterat strukturen för en additiv linjär modell. En multiplikativ modell som $\mu_{ij} = \gamma_0\gamma_{1i}\gamma_{2j}$, som vi hellre arbetar med, kan konstrueras om till en "additiv" modell genom att logaritmera:

$$\log(\mu_{ij}) = \log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j}). \quad (4.2.4)$$

Vi väljer igen en bascell, säg $(1, 1)$, där $\gamma_{11} = \gamma_{21} = 1$. Vi byter in nya parametrar enligt

$$\beta_1 \doteq \log(\gamma_0),$$

$$\beta_2 \doteq \log(\gamma_{12}),$$

$$\beta_3 \doteq \log(\gamma_{22}),$$

$$\beta_4 \doteq \log(\gamma_{23}),$$

Tillsammans med dummyvariablerna i tabellen på den föregående sidan får vi ekvationen

$$\log(\mu_i) = \sum_{j=1}^{4} x_{ij}\beta_j \quad i = 1, 2, ..., 6, \quad (4.2.5)$$

vilken har samma linjära struktur som (4.2.2), dock $\log(\mu_i)$ i vänsterled istället för μ_i. Vi har nu fått (en variant av) inledningen till en **generaliserad linjär modell, GLM**, där vänsterledet i ekvationen ovan är *någon monoton funktion* $g(\cdot)$ av μ_i. Denna del av en GLM kallas för *länkfunktionen*, eftersom den länkar väntevärdet till en linjär struktur genom

$$g(\mu_i) = \sum_{j=1}^{r} x_{ij}\beta_j \quad i = 1, 2, ..., n. \quad (4.2.6)$$

Detta är det allmänna fallet i en tariffanalys av en respons Y_i, med r kovariater (oberoende variabler) $X_1, X_2, ..., X_r$. Man brukar definiera den *linjära prediktorn* η_i till en GLM enligt $\eta_i \equiv g(\mu_i)$.

I och med detta har vi nu definierat de hörnstenar som behövs för att bygga upp en generaliserad linjär modell. Härnäst tittar vi på den matematiska definitionen, samt några exempel.

Definition 4.17. En generaliserad linjär modell för en respons Y_i med r kovariater består av en linjär prediktor

$$\eta_i = \sum_{j=1}^{r} x_{ij}\beta_j \quad i = 1, 2, ..., n, \tag{4.2.7}$$

där x_{ij} är värdet av kovariaten X_j för observation i, samt två funktioner:

1. en länkfunktion $g(\cdot)$ som beskriver hur väntevärdet $\mathbb{E}(Y_i) = \mu_i$ beror av den linjära prediktorn

$$g(\mu_i) = \eta_i,$$

2. en variansfunktion $\nu(\cdot)$ som beskriver hur variansen, $\mathrm{Var}(Y_i)$ beror av väntevärdet

$$\mathrm{Var}(Y_i) = \phi\nu(\mu_i),$$

där spridningsparametern ϕ är en konstant, dvs. oberoende av i, se avsnitt 4.1.

Vi har sett att multiplikativa modeller svarar mot en logaritmisk länkfunktion, en s.k. *log-link*:

$$g(\mu_i) = \log(\mu_i),$$

medan den linjära modellen använder *identitetslänken* $\eta_i = \mu_i$, alltså $g(\mu_i) = \mu_i$. Notera att länkfunktionen inte får bero av i.

Om man ska analysera *förhållanden* är det vanligt att använda en *logit-länk*,

$$\eta_i = g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}). \tag{4.2.8}$$

Detta är eftersom väntevärdet hålls mellan ett och noll, vilket behövs i en modell när μ_i är ett förhållande, som i den relativa binomialfördelningen för nyckeltalet *förhållandet av storkrav*. Den motsvarande GLM-analysen går under namnet logistisk regression. Detta är den överlägset mest använda länkfunktionen inom prissättning inom icke-livsförsäkring.

Nedan presenteras en sammanfattning över GLM-generaliseringen av de ordinära linjära modellerna:

Viktad linjär regressionsmodell:

- Y_i följer en normalfördelning med $\mathrm{Var}(Y_i) = \sigma^2/\omega_i$;

- Väntevärdet följer en additiv modell $\mu_i = \sum_j x_{ij}\beta_j$.

Generaliserad linjär modell (GLM):

- Y_i följer en EDM med $\text{Var}(Y_i) = \phi\nu(\mu_i)/\omega_i$;
- Väntevärdet satisfierar $g(\mu_i) = \sum_j x_{ij}\beta_j$, där g är en monoton funktion.

Multiplikativ Tweediemodell, delklass av GLM:s:

- Y_i följer en Tweedie-EDM med $\text{Var}(Y_i) = \phi \cdot \mu_i^p/\omega_i, \quad p \geq 1$;
- Väntevärdet följer en multiplikativ modell $\log(\mu_i) = \sum_j x_{ij}\beta_j$.

Exempel 4.18. Specialfall: Normala generella linjära modellen.
För den generella linjära modellen med $\epsilon \sim N(0, \sigma^2)$ har vi den linjära prediktorn

$$\eta = \beta_0 + \beta_1 x_{1i} + ... + \beta_r x_{ri}.$$

Länkfunktionen är

$$g(\mu_i) = \mu_i,$$

medan variansfunktionen blir

$$\nu(\mu_i) = 1.$$

Nedan tar vi två exempel med modellering av data för Binomial- respektive Poissonfördelade stokastiska variabler:

Exempel 4.19. Antag att $Y_i \sim \text{Bin}(n_i, \mu_i)$ och att vi försöker modellera proportionen Y_i/n_i. Då får vi

$$\mathbb{E}(Y_i/n_i) = \mu_i \text{ och } \text{Var}(Y_i/n_i) = \frac{1}{n_i}\mu_i(1-\mu_i).$$

Ur dessa resultat fås att variansfunktionen blir

$$\nu(\mu_i) = \mu_i(1-\mu_i).$$

Eftersom $\mu_i \in (0,1)$ bör länkfunktionen i detta fall avbilda $(0,1) \to (-\infty, \infty)$. Ett vanligt val är

$$g(\mu_i) = \text{logit}(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$$

Exempel 4.20. Antag att $Y_i \sim Po(\mu_i)$. Då får vi att

$$\mathbb{E}(Y_i) = \mu_i \text{ och } \mathrm{Var}(Y_i) = \mu_i$$

Variansfunktionen blir alltså

$$\nu(\mu_i) = \mu_i$$

Eftersom $\mu_i \in (0, \infty)$ bör länkfunktionen avbilda $(0, \infty) \to (-\infty, \infty)$. Vanligtvis väljer man

$$g(\mu_i) = \log(\mu_i)$$

4.2.2 Kanoniska länken

Det dyker upp många viktiga parametrar och funktioner inom GLM-teorin. Några av dessa kan kopplas ihop enligt följande relation

$$\theta \xrightarrow{b'(\cdot)} \mu \xrightarrow{g(\cdot)} \eta. \tag{4.2.9}$$

Här är $b(\cdot)$, och därför också $b'(\cdot)$, bestämda av strukturen av de slumpmässiga komponenterna, som i sin tur bestäms unikt genom valet av variansfunktion. Länkfunktionen $g(\cdot)$ å andra sidan är en del av modellerandet av väntevärdet, och det finns flera logiska val. Detta leder in oss på en ny definition:

Definition 4.21. Den kanoniska länken definieras som

$$g(\cdot) = (b'(\cdot))^{-1}.$$

Från (4.2.9) får vi att den kanoniska länken medför att $\theta = \eta$ måste gälla, eftersom $\eta = g(\mu) = g'(\theta) = (b')^{-1}(b'(\theta)) = \theta$. Användandet av den kanoniska länken gör vissa uträkningar lättare, men kommer inte att tas upp desto mer under denna avhandling.

4.3 Parameterestimering

Den viktigaste delen inom GLM-analys är att kunna *estimera parametrarna* (β_j:na) i (4.2.7), vilket ger oss relativiteterna (γ_{ij}) som är hörnstenarna i en tariff. Inledningsvis tittar vi på fallet med endast två värderingsfaktorer i en multiplikativ modell.

En estimeringsmetod som användes förr i tiden var *totalmarginal-metoden* (the method of marginal totals (MMT)). Den bygger på att även om man inte har all info om krav i en viss cell kan man kanske ha det för hela marginalen, alltså när vi summerar över alla värderingsfaktorer förutom en. Därefter sätter man denna marginella summa lika med sitt väntevärde och löser ut parametrarna.

Till exempel om vi inte råkar ha information om en viss bilmodell (i) i en viss region (j), summerar vi först över alla regionsvisa tariffer och åldersvisa tariffer, där i respektive j hålls fixerat. Dessa två väntevärden bör vara lika, och härifrån kan man sen lösa ut parametrarna.

Vi utgår alltså från den multiplikativa modellen i förra avsnittet med två värderingsfaktorer. Den ges av ekvationen (3.3.1), som lyder

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}, \tag{4.3.1}$$

där $i = 1, \ldots, m_1$ och $j = 1 \ldots, m_2$. Vi kallar utfallet för den stokastiska variabeln X för x och får följande ekvationer (fallet med två värderingsfaktorer):

$$\sum_j \mathbb{E}(X_{ij}) = \sum_j x_{ij}; \quad i = 1, ..., m_1;$$

$$\sum_i \mathbb{E}(X_{ij}) = \sum_i x_{ij}; \quad j = 1, ..., m_2;$$

där m_1 och m_2 är antalet klasser för värderingsfaktor i respektive j. Eftersom $X_{ij} = \omega_{ij} Y_{ij}$ och $\mathbb{E}(Y_{ij}) = \mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}$ fås ekvationssystemet

$$\begin{aligned} \sum_j \omega_{ij} \gamma_0 \gamma_{1i} \gamma_{2j} &= \sum_j \omega_{ij} y_{ij}; \quad i = 1, ..., m_1; \\ \sum_i \omega_{ij} \gamma_0 \gamma_{1i} \gamma_{2j} &= \sum_i \omega_{ij} y_{ij}; \quad j = 1, ..., m_2; \end{aligned} \tag{4.3.2}$$

där, märk väl, $\gamma_{11} = \gamma_{21} = 1$. Detta ekvationssystem har ingen lösning i sluten form utan man måste lösa det numeriskt. Efter en hel del omflyttande erhålls ett nytt system:

$$\gamma_0 = \frac{\sum_i \sum_j \omega_{ij} y_{ij}}{\sum_i \sum_j \omega_{ij} \gamma_{1i} \gamma_{2j}} \tag{4.3.3}$$

$$\gamma_{1i} = \frac{\sum_j \omega_{ij} y_{ij}}{\gamma_0 \sum_j \omega_{ij} \gamma_{2j}}; \quad i = 2, ..., m_1; \tag{4.3.4}$$

$$\gamma_{2j} = \frac{\sum_i \omega_{ij} y_{ij}}{\gamma_0 \sum_i \omega_{ij} \gamma_{1i}}; \quad j = 2, ..., m_2. \tag{4.3.5}$$

Ekvationerna för γ_{1i} och γ_{2j} fås direkt ur (4.3.2). För att få γ_0 börjar vi från ekvation (4.3.4) och summerar över $i, i = 2, \ldots m_1$ enligt:

$$\begin{aligned}\sum_i \gamma_{1i} &= \frac{\sum_i \sum_j \omega_{ij} y_{ij}}{\sum_i \gamma_0 \sum_j \omega_{ij} \gamma_{2j}} \\ <=> \gamma_0 &= \frac{\sum_i \sum_j \omega_{ij} y_{ij}}{\sum_i \gamma_{1i} \sum_j \omega_{ij} \gamma_{2j}} \\ <=> \gamma_0 &= \frac{\sum_i \sum_j \omega_{ij} y_{ij}}{\sum_i \sum_j \omega_{ij} \gamma_{1i} \gamma_{2j}}.\end{aligned}$$

Samma resultat fås även genom att börja från ekvation (4.3.5) och summera över $j, j = 2, \ldots, m_2$.

Genom att börja med att ge begynnelsevärden för relativiteterna γ_0, γ_{1i} och γ_{2j} och insätta dessa i (4.3.3) - (4.3.5) fås värden för resten av relativiteterna när man itererar över summorna.

Exempel 4.22. *Fortsättning på mopedförsäkring*
Härnäst gör vi en uppgift i R. Uppgiften bygger på det tidigare exemplet med mopedförsäkringar. Här räknar vi först ut relativiteterna enligt tariffen i tabell 1.2 och sedan med hjälp av GLM:s, för att kunna jämföra skillnaderna.

```
## Laddar tabell1.2
if (!exists("tabell.1.2"))
    load("tabell.1.2.RData")
## Varje kolumn räknas ut individuellt

## Ger kolumnerna vettiga namn
varderingsfaktor <-
    with(tabell.1.2,
         c(rep("Fordonsklass", nlevels(premiekl)),
           rep("Fordonsålder", nlevels(moptva)),
           rep("Zon", nlevels(zon))))

## Kolumnklassen
klass.num <- with(tabell.1.2, c(levels(premiekl), levels(moptva),
    levels(zon)))
```

```
## Varaktigheten är summan av varaktigheterna inom varje klass
## (Note to self:) Tapply-funktionen är användbar om man vill spjälka upp
## en vektor i grupper som definieras av någon klassifierad vektor.
## Därefter räknar man en funktion, t.ex summan, av dessa två vektorer.

varaktighet.total <-
    c(with(tabell.1.2, tapply(dur, premiekl, sum)),
      with(tabell.1.2, tapply(dur, moptva, sum)),
      with(tabell.1.2, tapply(dur, zon, sum)))

## (premiekl är klassen och moptva är ålder)
## Härnäst räknar vi ut relativiteterna i tariffen (helpre i tabell 1.2)
## Nämnaren i bråket är den klass som har högst exponering
## (alltså högsta totala varaktigheten): vi gör det tydligare med
## which.max()-konstruktionen. Nu kommer vi att få denna som bas,
## vilket är användbart för glm()-modellen senare.
klass.bas <- which.max(varaktighet.total[1:2])
alder.bas   <- which.max(varaktighet.total[3:4])
zon.bas  <- which.max(varaktighet.total[5:11])

## 1:2, 3:4 respektive 5:11 motsvarar raderna i sluttabellen

## with() gör något med det som omsluts, först data och sen en
## operation typ. Gör en sub.(någon värderingsfaktor)
## Här återskapas alltså aggregerade summor för 'helpre' i tariffen
sub.klass <- with(tabell.1.2, tapply(helpre, premiekl, sum))
sub.klass <- sub.klass / sub.klass[klass.bas]
sub.alder   <- with(tabell.1.2, tapply(helpre, moptva, sum))
sub.alder   <- sub.alder / sub.alder[alder.bas]
sub.zon  <- with(tabell.1.2, tapply(helpre, zon, sum))
sub.zon  <- sub.zon / sub.zon[zon.bas]

## stackoverflow: "Contrasts are needed when you fit linear models
## with factors (i.e. categorical variables) as explanatory variables.
```

```
## The contrast specifies how the levels of the factors will be coded
## into a family of numeric dummy variables for fitting the model."

## rank() rangordnar efter elementets nummer. Om första elementet i
## en lista är det tredje minsta i hela listan, så ger den ut en 3a.
contrasts(tabell.1.2$premiekl) <-
    contr.treatment(nlevels(tabell.1.2$premiekl))[
rank(-varaktighet.total[1:2], ties.method = "first"), ]
contrasts(tabell.1.2$moptva) <-
    contr.treatment(nlevels(tabell.1.2$moptva))[
rank(-varaktighet.total[3:4], ties.method = "first"), ]
contrasts(tabell.1.2$zon) <-
    contr.treatment(nlevels(tabell.1.2$zon))[
rank(-varaktighet.total[5:11], ties.method = "first"), ]



## Här kommer GLM-biten för att skatta relativiteterna.
## SAS-koden hittas på http://staff.math.su.se/esbj/GLMbook/moppe.sas
## Här har författarna tagit hjälp av en GLM med Poissonfördelning
## och log-länk så gör på liknande sätt
z <- glm(riskpre ~ premiekl + moptva + zon, data = tabell.1.2,
         family = poisson("log"), weights = dur)

## Länkfunktionen som används är log, därför använder vi exp här
## coef() står för koefficienterna framför variabeln
rels <- exp( coef(z)[1] + coef(z)[-1] ) / exp(coef(z)[1])

del.klass <- c(1, rels[1])              # Se del.zon nedan för den
del.alder   <- c(rels[2], 1)              # allmänna ansatsen
del.zon  <- c(1, rels[3:8])[rank(-varaktighet.total[5:11],
ties.method = "first")]

## Skapar och sparar data-tabellen
table.1.4 <-
```

```
    data.frame(Varderingsfaktor = varderingsfaktor, Klass = klass.num,
               Varaktighet = varaktighet.total,
               Rel.tariff = c(sub.klass, sub.alder, sub.zon),
               Rel.GLM    = c(del.klass, del.alder, del.zon))
save(table.1.4, file = "table.1.4.RData")
print(table.1.4, digits = 3)
```

	Varderingsfaktor	Klass	Varaktighet	Rel.tariff	Rel.MMT
1	Fordonsklass	1	9833	1.00	1.000
2	Fordonsklass	2	8825	0.50	0.432
3	Fordonsålder	1	1918	1.67	2.730
4	Fordonsålder	2	16740	1.00	1.000
5	Zon	1	1451	5.17	8.971
6	Zon	2	2486	3.10	4.193
7	Zon	3	2889	1.92	2.523
8	Zon	4	10069	1.00	1.000
9	Zon	5	246	2.50	1.237
10	Zon	6	1369	1.50	0.735
11	Zon	7	148	1.00	1.228

Figur 4.3.1: Tabell 1.4, uträknad med hjälp av R

Man kan fundera kring trovärdigheten i avvikelserna för vissa celler. De som inte har många observationer kan ha ett ganska slumpartat fel. En fördel med GLM framför MMT är att de kan avgöra denna fundering, med hjälp av konfidensintervall och statistiska test.

En viktig anmärkning är att man får samma estimat för relativiteterna om man använder MMT eller denna typ av GLM, alltså en Poisson-GLM med log-länk. Av den orsaken får vi samma resultat i kolumnen ”Rel.MMT”, fast vi använde GLMs. Författarna har också använt sig av en GLM men påpekar att samma resultat fås om man använder MMT. För hänvisning, se sid 31 i [1]. Kolumnen ”Rel.tariff” är relativiteterna för kolumnen ”helpre” i tabell 1.2, som är premien som användes i den ursprungliga tariffen från år 1999. Först summeras premien skilt över respektive fordonsklass, fordonsålder och zon. Därefter undersöks vilken klass, ålder respektive zon som har högst varaktighet och denna väljs som bascell. Därefter divideras de summerade premierna med respektive bascell, vilket ger relativiteten. Detta förklarar varför de rader som har högst varaktighet, dvs. fordonsklass 1, fordonsålder 2 och zon 4, har relativiteten lika

med 1. Det som "Rel.tariff" berättar är hur premierna relativt skiljer från bascellen. Till exempel är premien 67 % högre (relativitet 1.67) för fordonsålder 1 jämfört med fordonsålder 2.

Exempel 4.23. Nedan har vi två tabeller tagna ur sid 13 [1] med data om lastbilar. Tabell 1.5 är antalet kontraktår uppdelat efter årlig körsträcka och tabell 1.6 är samma för kravfrekvens. Datat är ursprungligen från Länsförsäkringar Alliance.

Tabell 1.5:

Antal mil	Fordonets	ålder
per år	Ny	Gammal
Låga	47 039	190 513
Höga	56 455	28 612

Tabell 1.6:

Antal mil	Fordonets	ålder	
per år	Ny	Gammal	Total
Låga	0.033	0.025	0.026
Höga	0.067	0.049	0.061
Total	0.051	0.028	

Vår uppgift är att göra MMT-skattningar för relativiteterna. Vi väljer cellen low/new som referenscell, altså cellen (1, 1). Detta medför alltså att $\gamma_{11} = \gamma_{21} = 1$ och vi vill alltså skatta γ_0, γ_{21} och γ_{22}. Referensceller (eller basceller) togs upp i samband med (3.3.1).

MMT-estimaten visar sig vara följande tal, estimerade från [1], sid 14:
$\gamma_0 = 0.03305, \gamma_{12} = 2.01231$ och $\gamma_{22} = 0.74288$.

```
## Vi sätter in estimatens värden:
rel0  <- 0.03305
rel12 <- 2.01231
rel22 <- 0.74288

##Data från tabeller 1.5 och 1.6:
dim.names <- list(Milage = c("Låg", "Hög"),
                  Age = c("Ny", "Gamla"))
```

```
artot <- matrix(c(47039, 56455, 190513, 28612), nrow = 2,
                dimnames = dim.names)
skada <- matrix(c(0.033, 0.067, 0.025, 0.049), nrow = 2,
                dimnames = dim.names)

## Funktion för att räkna ut estimatens fel
GvalsError <- function (gvals) {
    ## Namnger de tre estimaten
    gamma0  <- gvals[1]
    gamma12 <- gvals[2]
    gamma22 <- gvals[3]

    ## De nuvarande estimaten, presenterade i matrisform
    G  <- matrix(c(1, 1, gamma12, gamma22), nrow = 2)
    Gamma1 <- matrix(c(1, gamma12), nrow = 2, ncol = 2)
    Gamma2 <- matrix(c(1, gamma22), nrow = 2, ncol = 2, byrow = TRUE)

    ## De uträknade värdena, enligt MMT-ekvationerna ovan och
    ## via matrismultiplikation
    Gamma0  <- addmargins(skada * artot)["Sum", "Sum"] /
( sum(artot * Gamma1 * Gamma2) )
    Gamma12 <- addmargins(skada * artot)["Hög", "Sum"] /
          ( gamma0 * addmargins(artot * Gamma2)["Hög", "Sum"] )
    Gamma22 <- addmargins(skada * artot)["Sum", "Gamla"] /
        ( gamma0 * addmargins(artot * Gamma1)["Sum", "Gamla"] )
    ## Summan av kvadrerade felen, alltså minsta kvadrat-metoden
    error <- (gamma0 - Gamma0)^2 + (gamma12 - Gamma12)^2 +
(gamma22 - Gamma22)^2
    return(error)
}

## Vi minimerar felfunktionen ("error") för att få ett estimat
## optim(mängd, felfunktion) optimerar mängden m.a.p. felfunktionen.
opt <- optim(c(rel0, rel12, rel22), GvalsError)
```

```
stopifnot(opt$convergence == 0)

## Erhåller parametrarna ur funktionen 'opt'
opt <- opt$par

varden <- data.frame(legend = c("Uträknat värde", "Bokens värde"),
                     gamma0  = c(opt[1], rel0),
                     gamma12 = c(opt[2], rel12),
                     gamma22 = c(opt[3], rel22),
                     row.names = "legend")
print(varden, digits = 4)
```

```
                gamma0 gamma12 gamma22
Uträknat värde 0.03341   1.995  0.7452
Bokens värde   0.03305   2.012  0.7429
```

Figur 4.3.2: MMT-estimat, jämfört med resultaten i [1]

Svaren skiljer en aning (för hänvisning, se sid 14 i [1]), kan vara på grund av att författarna använt en annan felfunktion (error function). Felfunktionen som användes i denna uppgift bygger på minsta kvadrat-metoden, varefter datat optimeras med avseende på funktionen. Det som felfunktionen gör är att den först tar in tre variabler gamma0, gamma12 och gamma22. Därefter bygger den upp tre 2x2-matriser av dessa variabler, som motsvarar summorna med ω och y i ekvationer (4.3.3) - (4.3.5). Sedan används dessa matriser var för sig för att få MMT-estimat för Gamma0, Gamma12 och Gamma22. Slutligen används minsta kvadrat-metoden på variablerna och MMT-estimaten. Bokens estimat optimeras sedan med avseende på funktionen och ger sedan tre värden som är goda estimat för relativiteterna.

Vi verifierar ännu att dessa resultat stämmer överens med ekvationer (4.3.3) - (4.3.5). Vi har följande värden:
$\omega_{11} = 47039$, $\omega_{12} = 190513$, $\omega_{21} = 56455$, $\omega_{22} = 28612$, $y_{11} = 0.033$, $y_{12} = 0.025$, $y_{21} = 0.067$ och $y_{22} = 0.049$. Dessutom har vi att $\gamma_{11} = \gamma_{21} = 1$. Insättning i ekvationerna ger

$$\gamma_0 = \frac{11499,55...}{344944,29...} = 0,0333...,$$

$$\gamma_{12} = \frac{5184,47...}{2568,32...} = 2,0186...,$$
$$\gamma_{22} = \frac{6164,81...}{8268,81...} = 0,7455...,$$

vilka är goda estimat.

4.3.1 Multiplikativa Poissonmodellen

Vi vill nu göra en analys av kravfrekvensen Y, som vi kan anta att följer en relativ Poissonfördelning, som då har frekvensfunktion som ges av (4.1.3). Vi återgår till fallet med två värderingsfaktorer, dvs $Y = Y_{ij} \sim \text{Po}(\mu_{ij})$. Från antagande 3.3 har vi att cellerna är oberoende, vilket kommer att medföra att likelihooden av hela stickprovet är summan av likelihooden av de individuella cellerna:

$$\begin{aligned} l &= \log \mathcal{L}\big(\exp\{\omega_{ij}[y_{ij}\log(\mu_{ij}) - \mu_{ij}] + c(y_{ij},\omega_{ij})\}\big) \\ &= \log \prod_i \prod_j \exp\{\omega_{ij}[y_{ij}\log(\mu_{ij}) - \mu_{ij}] + c(y_{ij},\omega_{ij})\} \\ &= \sum_i \sum_j \omega_{ij}[y_{ij}(\log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j})) - \gamma_0\gamma_{1i}\gamma_{2j}] + c(y_{ij},\omega_{ij}). \end{aligned}$$

där c som synes inte beror av γ-parametrarna. Genom att derivera l med avseende på varje γ får man ett ekvationssystem för maximum-likelihood-estimaten (MLE), de så kallade ML-ekvationerna.

Som vi konstaterade på sid 44 råkar det sig att en GLM-modell för kravfrekvens med relativ Poissonfördelning och log-länk, har ML-ekvationerna lika med ekvationerna i MMT (totalmarginal-metoden). Detta innebär alltså att de ger samma estimat.

4.3.2 Allmänna resultat

Härnäst vänder vi oss till hur man löses ML-estimaten för β-parametrarna i en GLM. Estimaten bygger på ett stickprov av n observationer. De individuella observationerna följer en EDM-fördelning given i (4.1.1) och, tack vare att de är oberoende, fås log-sannolikheterna som funktioner av θ

$$\begin{aligned} l(\theta;\phi,\mathbf{y}) &= \log \mathcal{L}\big(\exp(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\big) \\ &= \log \prod_i \big(\exp(\frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i,\phi,\omega_i)\big) \\ &= \frac{1}{\phi}\sum_i \omega_i(y_i\theta_i - b(\theta_i)) + \sum_i c(y_i,\phi,\omega_i). \end{aligned} \tag{4.3.6}$$

Spridningsparametern ϕ påverkar inte maximeringen av l med avseende på θ, därför deriverar vi inte med avseende på denna parameter.

Vi kan finna likelihood-skattningen som en funktion av $\boldsymbol{\beta}$, snarare är θ, genom den inversa relationen $\mu_i = b'(\theta)$ och länken $g(\mu_i) = \eta_i = \sum_j x_{ij}\beta_j$. Derivatan av l med avseende på $\beta_j, j = 1,\dots r$ blir, enligt kedjeregeln:

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \sum_i \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_i (\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{1}{\phi}\sum_i (\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \end{aligned} \tag{4.3.7}$$

Vi har att $\mu_i = b'(\theta_i)$, vilket ger att $\partial\mu_i/\partial\theta_i = b''(\theta_i)$. Vi har även att $\partial\theta_i/\partial\mu_i = 1/(\partial\mu_i/\partial\theta_i)$, vilket ger att $\partial\theta_i/\partial\mu_i = 1/b''(\theta_i) = 1/\nu(\mu_i)$, per definition (se kapitel om EDM:s).

Dessutom har vi att $\partial\mu_i/\partial\eta_i = [\partial\eta_i/\partial\mu_i]^{-1} = 1/g'(\mu_i)$. Utöver detta gäller ännu att $\eta_i = \sum_j x_{ij}\beta_j$, som medför att $\partial\eta_i/\partial\beta_i = x_{ij}$.

Genom att kombinera dessa resultat erhålls slutligen

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_i \omega_i \frac{y_i - \mu_i}{\nu(\mu_i)g'(\mu_i)}x_{ij}, \tag{4.3.8}$$

som kallas för *resultatfunktionen* (Eng: score function). Genom att sätta dessa r stycken partiella derivator lika med noll och genom att multiplicera med ϕ, erhåller vi ML-ekvationerna

$$\sum_i \omega_i \frac{y_i - \mu_i}{\nu(\mu_i)g'(\mu_i)}x_{ij} = 0, \quad j = 1,2,...,r. \tag{4.3.9}$$

Man kan tänka sig att lösningen helt enkelt är $\mu_i = y_i$, men då glömmer man att $\mu_i = \mu_i(\boldsymbol{\beta})$ också måste satisfiera relationen som ges av regressionen av x:n, alltså

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\big(\sum_j x_{ij}\beta_j\big). \tag{4.3.10}$$

Det är endast fallet då antalet parametrar (β_j) är lika stort som antalet observationer (i, x_i = värdet för observation i), som tillåter lösningen $\mu_i = y_i$, alltså då $r = n$.

Man kan härmed konstatera att de enda komponenterna i en sannolikhetsfördelning som påverkar ML-ekvationerna i (4.3.9) är **väntevärdet** och **variansen**, genom länkfunktionen g och variansfunktionen ν.

Exempel 4.24. (Tweediemodeller) I en tariffanalys använder man vanligtvis Tweediemodeller, som har $\nu(\mu) = \mu^p$. De har även en multiplikativ modell för väntevärdet, alltså $g(\mu_i) = \log(\mu_i)$, som indikerar att $g'(\mu_i) = 1/\mu_i$. Då blir den allmänna ML-ekvationen i (4.3.9):

$$\sum_i \omega_i \frac{y_i - \mu_i}{\mu_i^{p-1}} x_{ij} = 0 \quad \Longleftrightarrow \quad \sum_i \frac{\omega_i}{\mu_i^{p-1}} y_i x_{ij} = \sum_i \frac{\omega_i}{\mu_i^{p-1}} \mu_i x_{ij}, \tag{4.3.11}$$

där μ:na är kopplade till β:na genom

$$\mu_i = \exp\left\{ \sum_j x_{ij} \beta_j \right\}. \tag{4.3.12}$$

Jämfört med Poissonfallet med $p = 1$, får modeller med $p > 1$ mer tyngd på bägge sidor om den högra ekvationen i (4.3.11) av μ^{p-1}, vilket ger mindre vikt åt celler med ett högt väntevärde. Man kan erhålla relativiteterna γ_j genom relationerna $\gamma_j = \exp\{\beta_j\}$.

4.3.3 Multiplikativ gammamodell för kravstorleken

I avsnittet om den multiplikativa Poissonmodellen såg vi hur man kan använda den för att estimera kravfrekvensen; nästa viktiga fall är kravstorleken. Exponeringen ω_i i cell i är antalet krav och Y_i är kravstorleken, som antas vara relativt gammafördelad med densitet angiven i (4.1.5). Vi börjar med att undersöka relationen mellan väntevärdet μ_i och variansen. Enligt lemma 4.14 och tabell 3 har vi att $\mathbb{E}(Y_i) = \mu_i$ och $\mathrm{Var}(Y_i) = \phi\mu_i^2/\omega_i$. Därav får vi att

$$\frac{\phi}{\omega_i} = \frac{\mathrm{Var}(Y_i)}{\mathbb{E}(Y_i)^2} <=> \sqrt{\frac{\phi}{\omega_i}} = \sqrt{\frac{\mathrm{Var}(Y_i)}{\mathbb{E}(Y_i)^2}} = \frac{\sigma_i}{\mu_i}, \tag{4.3.13}$$

vilket innebär att variationskoefficienten CV_{Y_i} är konstant över celler med samma exponering ω_i, se definition **??**. Detta är ett helt logiskt resultat; om vi har en

tariffcell med väntevärde 50 och standardavvikelse 8 kan vi förvänta oss att i en annan cell med samma exponering och väntevärde 100, att standardavvikelsen är 16 istället för 8.

Från (4.3.11) med $p = 2$ får vi ML-ekvationerna, som i detta fall är

$$\sum_i \omega_i \frac{y_i}{\mu_i} x_{ij} = \sum_i \omega_i x_{ij}. \tag{4.3.14}$$

Vi skriver om dessa ekvationer i tabellform i det simpla fallet med endast två värderingsfaktorer, genom att använda den multiplikativa modellen i (3.3.1)

$$\begin{aligned} \sum_i \sum_j \frac{\omega_{ij} y_{ij}}{\gamma_0 \gamma_{1i} \gamma_{2j}} &= \sum_i \sum_j \omega_{ij}; \\ \sum_j \frac{\omega_{ij} y_{ij}}{\gamma_0 \gamma_{1i} \gamma_{2j}} &= \sum_j \omega_{ij}; \quad i = 2, ..., k_1; \\ \sum_i \frac{\omega_{ij} y_{ij}}{\gamma_0 \gamma_{1i} \gamma_{2j}} &= \sum_i \omega_{ij}; \quad j = 2, ..., k_2. \end{aligned} \tag{4.3.15}$$

Ur detta ekvationssystem kan man relativit enkelt estimera relativiteterna, numeriskt.

4.3.4 Modell för riskpremien samt Case Study

I slutändan är det modellen för riskpremien som ger tariffen. En vanlig rekommendation är att använda en Tweediemodell med $1 < p < 2$ för att direkt analysera riskpremien. Den grundläggande metoden bygger dock på att man först gör separata analyser för kravfrekvensen och kravstorleken, härefter erhålls relativiteterna för riskpremien genom att multiplicera dessa. Orsaken för uppdelning i två GLM:s är:

1. Kravfrekvens är vanligtvis mycket stabilare än kravstorleken och huruvida värderingsfaktorer påverkar analysen starkt eller inte är ofta kopplat till kravfrekvensen. Dessa faktorer kan estimeras bättre i en skild analys;
2. En separat analys ger mer kunskap om hur värderingsfaktorerna påverkar riskpremien.

Exempel 4.25. Vi återgår igen till mopedexemplet med datat från tabell 1.2. I exempel 4.23 använde vi en GLM för att finna riskpremien. Nu gör vi först skilda analyser för kravfrekvensen och kravstorleken för att sedan multiplicera

dessa resultat och på så sätt erhålla riskpremien.

```
## Laddar tabell 1.2
if (!exists("tabell.1.2"))
    load("tabell.1.2.RData")

## Sedan tidigare installerat foreach-packagen
library("foreach")

## Vi återskapar tabell 2.7 (sid 35 i [1]) och lägger till
## kolumnerna en i taget
tabell.2.7 <-
    data.frame(varderings.faktor =
                c(rep("Fordonsklass", nlevels(tabell.1.2$premiekl)),
                  rep("Fordonsålder",   nlevels(tabell.1.2$moptva)),
                  rep("Zon",          nlevels(tabell.1.2$zon))),
                klass =
                c(levels(tabell.1.2$premiekl),
                  levels(tabell.1.2$moptva),
                  levels(tabell.1.2$zon)),
                stringsAsFactors = FALSE)

## Vi räknar ut varaktigheten per värderingsfaktor och bestämmer även
## ''contrastsen''(samma typ av kod som i de föregående uppgifterna)
## Använder foreach för att exekvera loopen då den är användbar för
## contrasts och för att ackumulera summan

new.cols <-
    foreach (varderings.faktor= c("premiekl", "moptva", "zon"),
             .combine = rbind) %do%
{
    antskador <- tapply(tabell.1.2$antskad, tabell.1.2
[[varderings.faktor]], sum)
    summor <- tapply(tabell.1.2$dur, tabell.1.2[[varderings.faktor]], sum)
```

```
    n.levels <- nlevels(tabell.1.2[[varderings.faktor]])
    contrasts(tabell.1.2[[varderings.faktor]]) <-
        contr.treatment(n.levels)[rank(-summor, ties.method = "first"), ]
    data.frame(varaktighet = summor, n.skador = antskador)
}
tabell.2.7 <- cbind(tabell.2.7, new.cols)
rm(new.cols)

## Frekvensmodellen, en GLM med Poissonfördelning
model.frekvens <-
    glm(antskad ~ premiekl + moptva + zon + offset(log(dur)),
        data = tabell.1.2, family = poisson)

rels <- coef( model.frekvens )
rels <- exp( rels[1] + rels[-1] ) / exp( rels[1] )
tabell.2.7$rels.frekvens <-
    c(c(1, rels[1])[rank(-tabell.2.7$varaktighet[1:2], ties.method = "first")],
      c(1, rels[2])[rank(-tabell.2.7$varaktighet[3:4], ties.method = "first")],
      c(1, rels[3:8])[rank(-tabell.2.7$varaktighet[5:11],
ties.method = "first")])

## Kravstorleksmodellen, en GLM med Gammafördelning
model.skadestorlek <-
    glm(medskad ~ premiekl + moptva + zon,
        data = tabell.1.2[tabell.1.2$medskad > 0, ],
        family = Gamma("log"), weights = antskad)

rels <- coef( model.skadestorlek )
rels <- exp( rels[1] + rels[-1] ) / exp( rels[1] )
## (Note to self:) För den kanoniska länkfunktionen använder vi
## rels <- rels[1] / (rels[1] + rels[-1])

tabell.2.7$rels.skadestorlek <-
    c(c(1, rels[1])[rank(-tabell.2.7$varaktighet[1:2], ties.method = "first")],
```

```
      c(1, rels[2])[rank(-tabell.2.7$varaktighet[3:4], ties.method = "first")],
      c(1, rels[3:8])[rank(-tabell.2.7$varaktighet[5:11],
ties.method = "first")])

tabell.2.7$rels.riskpremie <- with(tabell.2.7,
rels.frekvens * rels.skadestorlek )
print(tabell.2.7, digits = 2)
```

	varderings.faktor	klass	varaktighet	n.skador	rels.frekvens	rels.skadestorlek	rels.riskpremie
1	Fordonsklass	1	9833	391	1.00	1.00	1.00
2	Fordonsklass	2	8825	395	0.78	0.55	0.42
11	Fordonsålder	1	1918	141	1.55	1.79	2.78
21	Fordonsålder	2	16740	645	1.00	1.00	1.00
12	Zon	1	1451	206	7.10	1.21	8.62
22	Zon	2	2486	209	4.17	1.07	4.48
3	Zon	3	2889	132	2.23	1.07	2.38
4	Zon	4	10069	207	1.00	1.00	1.00
5	Zon	5	246	6	1.20	1.21	1.46
6	Zon	6	1369	23	0.79	0.98	0.78
7	Zon	7	148	3	1.00	1.20	1.20

Figur 4.3.3: Tabell med relativiteterna för riskpremien

I denna uppgift har vi alltså gjort skilda GLMs för kravfrekvensen och kravstorleken. Först skapas tabell 2.7 som finns på sid 35 i [1], med hjälp av tabell 1.2. Därefter skattas relativiteterna för kravfrekvensen med en Poisson-GLM och för kravstorleken med en gamma-GLM. Slutligen multipliceras relativiteterna radvis för att erhålla relativiteterna för riskpremien. Koden väljer här den värderingsfaktor vars löptid är störst som bascell eller referenscell, och jämför utgående från den de övriga relativiteterna. Detta är förklaringen till att deras relativiteter är lika med 1.

Variablerna *Fordonsklass* och *Fordonsålder* påverkar kravfrekvensen och kravstorleken åt samma håll, vilket betyder att nyare och starkare fordon är dyrare att ersätta när de blir stulna, samtidigt som de oftare *blir* stulna. Den geografiska zonen påverkar inte riskpremien märkbart, med undantag för storstäderna i zon 1.

Härnäst tar vi itu med en omfattande "Case study", från [1].

Exempel 4.26. Försäkringsdatat är från Wasa och handlar om motorcyklar. Datat innehåller krav från 1994-1998. Orsaken till att vi använder äldre data är eftersom nyare data är hemligstämplat, då de kan användas av försäkringsbolag i dagens läge. Datamängden **mccase.txt** finns tillgänglig på www.math.su.se/GLMbook och innehåller följande variabler:

- **AGARALD**: Ägarens ålder, mellan 0 och 99.
- **KON**: Ägarens kön, M för man och K för kvinna.
- **ZON**: Geografisk zon numrerad från 1 till 7. Numret för respektive zon är samma som i mopedexemplet
- **MCKLASS**: Motorcykelklass som är bestämd av det så kallade EV-förhållandet, definierad som (Motorns effekt i kW $\times$ 100) / (Fordonets vikt i kg + 75), avrundat till det närmaste nedre heltalet. 75 kg representerar förarens genomsnittliga vikt. EV-förhållandet delas upp i sju klasser som är representerade nedan
- **FORDALD**: Fordonets ålder, mellan 0 och 99.
- **BONUSKL**: En bonusklass som tar värden mellan 1 och 7. Samtliga förare börjar med bonusklass 1 och för varje kravfritt år ökar klassen med 1. Om ett krav uppstår sänks klassen med 2, detta innebär att en förare inte kan nå klass 7 utan att ha 6 kravfria år i rad.
- **DURATION**: Antalet försäkringsår.
- **ANTSKAD**: Antalet krav.
- **SKADKOST**: Kravkostnaden.

Denna case study är uppbyggd av tre uppgifter. Den första är att aggregera datat och att räkna ut den empiriska kravfrekvensen och kravstorleken. Vi börjar dock med att läsa in datat och förbereda inför utförandet av uppgifterna.

```
## install.packages(c("data.table", "foreach", "ggplot2"),
## dependencies = TRUE)
```

```
## Början: läser in datat
## 2L tar 'de första två från vänster', nästa 1L betyder därpå
## följande, osv
kolumner <- c(agarald = 2L, kon = 1L, zon = 1L, mcklass = 1L,
       fordald = 2L, bonuskl = 1L, duration = 8L,
antskad = 4L, skadkost = 8L)
## Ger varje faktor en egen 'typ'. Ålder är heltal, följande
## tre är strängar, alltså factor osv
klass.kolumner <- c("integer", "factor", "factor",
  "factor","integer", "factor",
  "numeric", "integer", "integer")
stopifnot(length(kolumner) == length(klass.kolumner))
## Länken där datat finns
lank <- url("http://www2.math.su.se/~esbj/GLMbook/mccase.txt")
motorc <- read.fwf(lank, widths = kolumner, header = FALSE,
                   col.names = names(kolumner),
                   colClasses = klass.kolumner,
                   na.strings = NULL, comment.char = "")
try(close(lank), silent = TRUE)
rm(kolumner, klass.kolumner, lank)
motorc$mcklass <- ordered(motorc$mcklass)
motorc$bonuskl <- ordered(motorc$bonuskl)

## Här konstrueras värderingsfaktorerna 'vard.1-4'.
motorc$vard.1 <- motorc$zon
motorc$vard.2 <- motorc$mcklass
motorc$vard.3 <-
    cut(motorc$fordald, breaks = c(0, 1, 4, 99),
        labels = as.character(1:3), include.lowest = TRUE,
        ordered_result = TRUE)
motorc$vard.4 <- ordered(motorc$bonuskl)
levels(motorc$vard.4) <-
    c("1", "1", "2", "2", rep("3", 3))
```

```
## Spara datat
save(motorc, file = "motorc.RData")
```

Nu är grunderna gjorda för att kunna börja på med uppgifterna. Vi har alltså kombinerat motorcykeldatat med informationen i tabell 2.8, sid 36 i [1], utan relativiteterna.

Uppgift 1: Aggregera cellerna till den nuvarande tariffen och räkna ut kravfrekvensen och kravstorleken.

```
## install.packages("data.table")
## install.packages("ggplot2")

if (!exists("motorc"))
    load("motorc.RData")

## Hämtar paketet data.table, som kan användas för att smidigt bygga
## upp matriser
library("data.table")
motorc <- data.table(motorc, key = paste("vard", 1:4, sep = "."))

## Skapar motorc.andra som är en aggregering av motorc med avseende
## på summorna för värderingsfaktorerna
motorc.andra <-
    motorc[,
           list(duration = sum(duration),
                antskad = sum(antskad),
                skadkost = sum(skadkost),
                num.policies = .N),
           by = key(motorc)]

## Kravfrekvens- och storlek, ändrar NaN till NA.
motorc.andra$frekv <-
    with(motorc.andra, ifelse(duration != 0, antskad / duration, NA_real_))
```

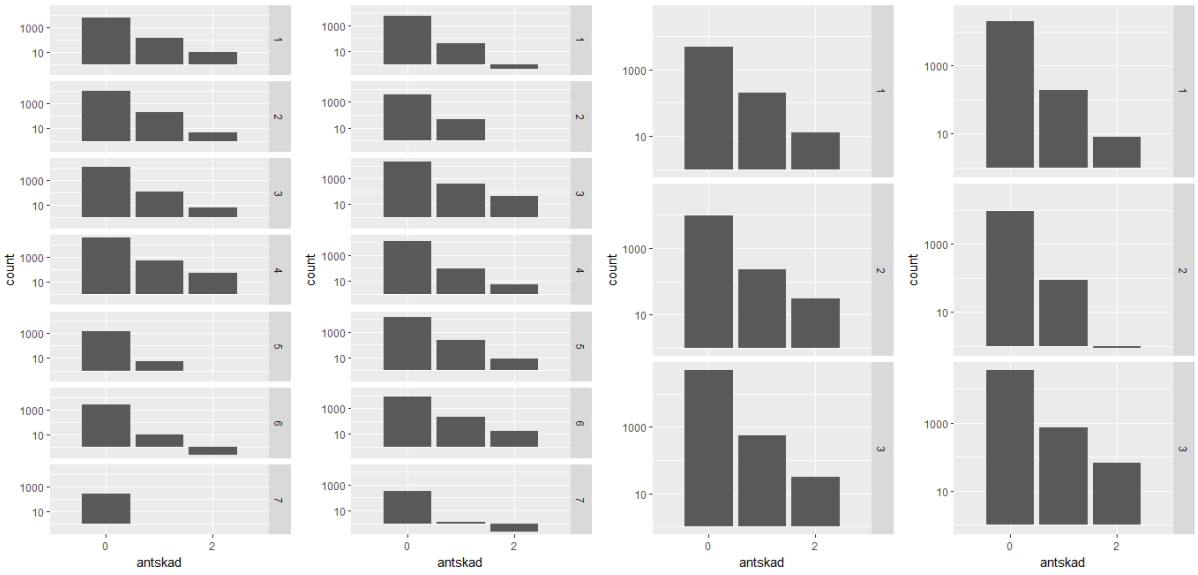
```
motorc.andra$storl <-
    with(motorc.andra, ifelse(antskad != 0, skadkost / antskad, NA_real_))

## Spara igen
save(motorc.andra, file = "motorc.andra.RData")
```

Det var uppgift 1. Vi skriver inte ut några resultat då dessa aggregeringar i sig inte ser så spännande ut, men är mycket användbara för resten av uppgiften. Vad vi hittills har gjort är att skapa datat ”motorc.andra” som är en aggregering av ”motorc” med avseende på summorna för värderingsfaktorerna. Kravfrekvensen och kravstorleken är sparade i datamängderna ”motorc.andra$frekv” respektive ”motorc.andra$storl”.

Uppgift 2 A: Bestäm hur antalet krav är fördelade.

```
## Laddar datat
library("data.table")
if (!exists("motorc"))
    load("motorc.RData")
if (!exists("motorc.andra"))
    load("motorc.andra.RData")
if (!is(motorc, "data.table"))
    motorc <- data.table(motorc, key = paste("vard", 1:4, sep = "."))

library("grid")
library("ggplot2")

## Vi börjar med att åskådliggöra värderingsfaktorerna i en plot. Det visar
## sig att värdemängden för antalet krav är {0, 1, 2} i detta data.
plot.titles <- c("Geografisk zon", "MC-klass", "Fordonsålder", "Bonusklass")
plots <-
   lapply(1:4,
          function(i)
          ggplot(motorc, aes(antskad))
          + geom_histogram(aes(weight = duration), stat = "count")
```

```
          + scale_x_discrete(limits = c(0, 2))
          + scale_y_log10()
          + facet_grid(paste("vard.", i, " ~ .", sep = ""),
                       scales = "fixed")
## Lämnar bort axeln för att få mer plats för datat
+ labs(axis.title.x = element_blank(), axis.title.y = element_blank(),
           axis.text.x = element_blank(),  axis.text.y = element_blank(),
           axis.title.y = element_blank(), axis.ticks = element_blank(),
           title = plot.titles[i])
           )

grid.newpage()
pushViewport(viewport(layout = grid.layout(nrow = 1, ncol = 4)))

for (i in 1:4) print(plots[[i]], vp = viewport(layout.pos.col = i))
```



Figur 4.3.4: Plot över antalet krav för de olika värderingsfaktorerna

Graferna är inte så vackra men ger en ungefärlig bild av hur antalet krav är fördelat över värderingsfaktorerna. Vi presenterar mer tydliga tabeller senare. De fyra i ordning från vänster är Geografisk zon (1-7), MC-klass (1-7), Fordonsålder (1-3) samt Bonusklass (1-3).

```
## Härnäst skapar vi en Poisson-GLM över antalet krav, då vi från
## teorin kan anta att antalet krav är Poissonfördelat.
data <- motorc[order(antskad), list(N = .N,
w = sum(duration)), by = antskad]
M <- glm(N ~ antskad, family = poisson(), weights = w, data = data)
data$uppskattad <- round(predict(M, data
[, list(antskad)], type = "response"))
print(data[, list(antskad, N, uppskattad)], digits = 1)
```

	antskad	N	uppskattad
1:	0	63878	63878
2:	1	643	646
3:	2	27	7

Figur 4.3.5: Antalet kontrakt med motsvarande antal krav, samt det skattade antalet krav givet Poissonfördelningen

Vi har här alltså presenterat det riktiga antalet krav/skador (N) och vad antalet krav skulle ha för värde, givet att de följer en Poissonfördelning ('uppskattad'). Resultaten är väldigt nära varandra tack vare den mycket tunga vikten kring $N = 0$, alltså inga krav.

Vi spjälker upp detta resultat på värderingsfaktornivå för att få en bättre inblick i datat.

```
## vard.1 representerar Geografisk zon.

data <-
    motorc[order(vard.1, antskad),
           list(N = .N, w = sum(duration)),
           by = list(vard.1, antskad)]

## Skapar en funktion 'modelleraPoisson', som gör GLM-analysen i
## ett svep, så att den lätt kan återanvändas.
modelleraPoisson <- function(data) {
    M <- glm(N ~ antskad, family = poisson(), weights = w, data = data)
```

```
    return(M)
}

## Vi skapar även en funktion 'skattaPoisson' som använder sig av
## modelleraPoisson och 'uppskattar' värden givet Poissonfördelningen,
## alltså på liknande sätt som tabell 4.3.5.
skattaPoisson <- function (data) {
    M <- modelleraPoisson(data)
    p <- predict(M, data[, list(antskad)], type = "response")
    return(p)
}
data$uppskattad <-
    unlist(lapply(levels(data$vard.1),
                  function (l) round(skattaPoisson(data[vard.1 == l]))))
print(data[, list(vard.1, antskad, N, uppskattad)], digits = 1)
```

	vard.1	antskad	N	uppskattad
1:	1	0	8409	8409
2:	1	1	163	164
3:	1	2	10	3
4:	2	0	11632	11632
5:	2	1	157	157
6:	2	2	5	2
7:	3	0	12604	12604
8:	3	1	113	113
9:	3	2	5	1
10:	4	0	24626	24626
11:	4	1	184	185
12:	4	2	6	1
13:	5	0	2368	2368
14:	5	1	9	9
15:	6	0	3867	3867
16:	6	1	16	16
17:	6	2	1	0
18:	7	0	372	372
19:	7	1	1	1

Figur 4.3.6: Antalet krav, geografisk zon

På ett liknande sätt har vi tung vikt vid N = 0, som gör att de riktiga värdena

och uppskattningen blir nästintill identiska när man antar en Poissonfördelning. Detta resultat är alltså för värderingsfaktor 1, geografisk zon.

Vi gör en lite annorlunda analys för vard.3, alltså fordonsålder. Vi märker att resultaten som vi jämför är ganska självklara, när vi alltid har en så tung vikt vid noll. Det finns alltså inte så mycket att analysera. Om vi istället aggregerar antalet krav med avseende på de fyra varaktigheterna får vi mer intressanta siffror. Detta innebär att vi summerar över antalet krav för alla likadana celler, alltså (1,1,1,1) och (1,1,1,1), (2,3,2,1) och (2,3,2,1) och så vidare. Datamängden som vi utgår från är alltså ”motorc.andra” istället för ”motorc”, se ovan.

```
maxant <- 7L

data <-
    motorc.andra[order(vard.3, antskad),
                 list(N = .N, w = sum(duration)),
                 by = list(vard.3, antskad)]
data$uppskattad <-
    unlist(lapply(levels(data$vard.3),
                  function (l) round(skattaPoisson(data[vard.3 == l]))))

print(data[antskad <= maxant,
           list(vard.3, antskad, N, uppskattad)], digits = 1)
```

	vard.3	antskad	N	uppskattad
1:	1	0	86	76
2:	1	1	13	37
3:	1	2	10	18
4:	1	3	9	9
5:	1	4	4	4
6:	1	5	5	2
7:	1	6	1	1
8:	1	7	1	0
9:	2	0	81	66
10:	2	1	26	34
11:	2	2	8	18
12:	2	3	8	9
13:	2	4	4	5
14:	2	5	4	2
15:	2	6	1	1
16:	2	7	3	1
17:	3	0	64	34
18:	3	1	19	27
19:	3	2	5	21
20:	3	3	10	17
21:	3	4	8	13
22:	3	5	10	10
23:	3	6	6	8
24:	3	7	8	6

Figur 4.3.7: Aggregerad summa för antalet krav, fordonsålder

Här blev skattningarna mindre självklara. Man ser dock fortfarande att Poissonfördelningens antagande uppfylls ganska bra. Vi har här filtrerat bort fall där antalet krav (”antskad”) är större än 7, då de bara förvränger resultatet. Det finns till exempel något enstaka fall med 10-15 krav, samt ett med över 30.

På samma sätt som för fordonsålder visar vi ännu bonusklassen.

```
## (vard.4)
data <-
    motorc.andra[order(vard.4, antskad),
                 list(N = .N, w = sum(duration)),
                 by = list(vard.4, antskad)]
data$uppskattad <-
    unlist(lapply(levels(data$vard.4),
```

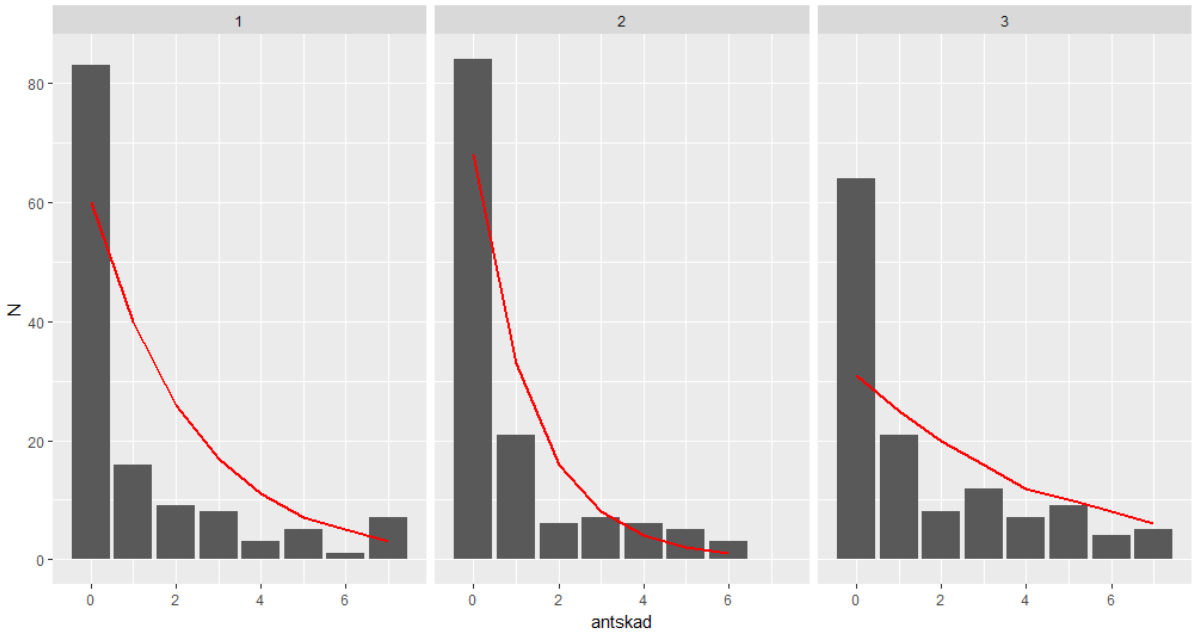
```
                function (l) round(skattaPoisson(data[vard.4 == l])))

print(data[antskad <= maxant,
           list(vard.4, antskad, N, uppskattad)], digits = 1)
```

	vard.4	antskad	N	uppskattad
1:	1	0	83	60
2:	1	1	16	40
3:	1	2	9	26
4:	1	3	8	17
5:	1	4	3	11
6:	1	5	5	7
7:	1	6	1	5
8:	1	7	7	3
9:	2	0	84	68
10:	2	1	21	33
11:	2	2	6	16
12:	2	3	7	8
13:	2	4	6	4
14:	2	5	5	2
15:	2	6	3	1
16:	3	0	64	31
17:	3	1	21	25
18:	3	2	8	20
19:	3	3	12	16
20:	3	4	7	12
21:	3	5	9	10
22:	3	6	4	8
23:	3	7	5	6

Figur 4.3.8: Aggregerad summa för antalet krav, bonusklass

Vi tar ännu och åskådliggör grafen för fallet ovan med bonusklassen. Den röda linjen representerar approximationen för Poissonfördelningen (’uppskattad’).

```
ggplot(data, aes(x = antskad, y = N)) +
    geom_bar(breaks = 0L:maxant, stat = "identity") +
    facet_wrap( ~ vard.4) +
    geom_line(aes(y = uppskattad), data = data, colour = "red", size = 1) +
    xlim(-0.5, maxant + 0.5)
```



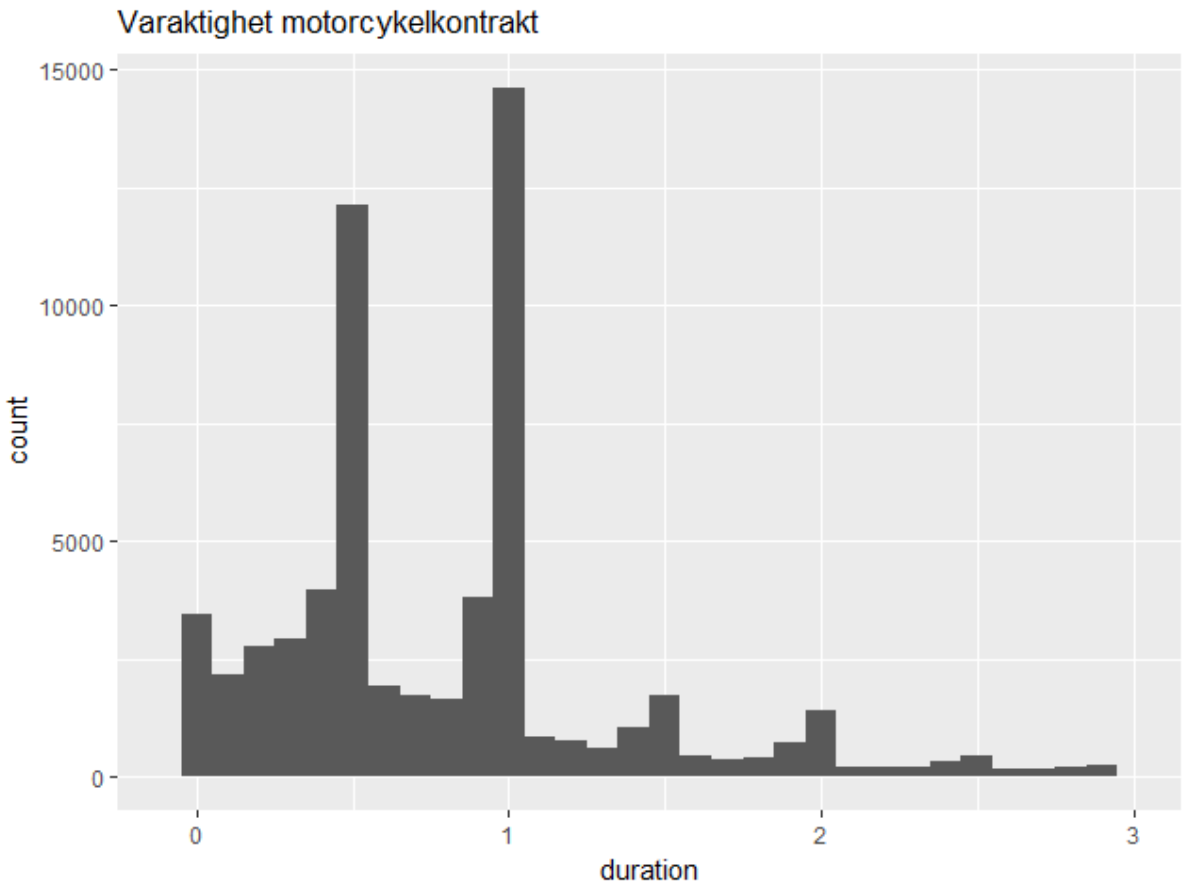
Figur 4.3.9: Graf med aggregerade antalet krav som staplar och Poissonfördelad approximation för bonusklass som röd linje

Vi kan alltså dra slutsatsen att antalet krav/skador är Poissonfördelat.

Uppgift 2 B: Bestäm hur varaktigheten är fördelad.

Löptiden/varaktigheten är mindre självklar, vad gäller fördelning. I och med att vi har kodat allt som behövs kan vi direkt åskådliggöra den aggregerade löptiden i grafen nedan.

```
ggplot(motorc, aes(duration)) +
    geom_histogram(binwidth = 0.1) +
    xlim(-0.1, 3) +
    labs(title = "Varaktighet motorcykelkontrakt") +
    annotate("text", 3.0, 1e4, label = "",
             size = 3, hjust = 1, vjust = 1)
```



Figur 4.3.10: Fördelningen för varaktigheten/löptiden

Vid 1/2 år och 1 år ses tydliga toppar, vilket är väntat eftersom försäkringskontrakt brukar gälla i ett år. Förvånansvärt många kontrakt med en varaktighet på 1/2 år, men det kanske var vanligt år 1995, då denna tariff är gjord. Det tycks dock inte finnas någon klar fördelning som i fallet med antalet krav.

Uppgift 3: Bestäm relativiteterna för kravfrekvensen och kravstorleken via GLM:s och använd dessa för att få relativiteterna för riskpremien.

Vi börjar med att återskapa tabell 2.8 i [1], sid 36. Nu tar vi även med relativiteterna för varje par av värderingsfaktor och klass.

```
## Hämtar datat
library("data.table")
if (!exists("motorc.andra"))
    load("motorc.andra.RData")
```

```
mangd24 <-
    data.frame(vard.faktor =
               c(rep("Zon",         nlevels(motorc.andra$vard.1)),
                 rep("MC-klass",    nlevels(motorc.andra$vard.2)),
                 rep("Åldersklass", nlevels(motorc.andra$vard.3)),
                 rep("Bonusklass", nlevels(motorc.andra$vard.4))),
               klass =
               with(motorc.andra,
                    c(levels(vard.1), levels(vard.2),
                      levels(vard.3), levels(vard.4))),
               ## Här insätts relativiteternas värden från
    ## tabell 2.8 i boken
               relativiteter =
               c(7.678, 4.227, 1.336, 1.000, 1.734, 1.402, 1.402,
                 0.625, 0.769, 1.000, 1.406, 1.875, 4.062, 6.873,
                 2.000, 1.200, 1.000,
                 1.250, 1.125, 1.000),
               stringsAsFactors = FALSE)
print(mangd24, digits = 3)
```

	vard.faktor	klass	relativiteter
1	Zon	1	7.678
2	Zon	2	4.227
3	Zon	3	1.336
4	Zon	4	1.000
5	Zon	5	1.734
6	Zon	6	1.402
7	Zon	7	1.402
8	MC-klass	1	0.625
9	MC-klass	2	0.769
10	MC-klass	3	1.000
11	MC-klass	4	1.406
12	MC-klass	5	1.875
13	MC-klass	6	4.062
14	MC-klass	7	6.873
15	Fordonsålder	1	2.000
16	Fordonsålder	2	1.200
17	Fordonsålder	3	1.000
18	Bonusklass	1	1.250
19	Bonusklass	2	1.125
20	Bonusklass	3	1.000

Figur 4.3.11: Tabell 2.8, sid 36 i [1]

Nu tar vi och gör GLM-analyser för kravfrekvensen och kravstorleken. Precis som i exempel 4.24 använder vi oss av en Poisson-GLM för kravfrekvensen och en gamma-GLM för kravstorleken, med log-länk för bägge. Sist och slutligen multipliceras dessa för att erhålla riskpremien.

```
## Lägger upp contrasten, som är relevanta senare
library("foreach")
new.cols <-
    foreach (vard.faktor = paste("vard", 1:4, sep = "."),
             .combine = rbind) %do%
{
    totals <- motorc.andra[, list(D = sum(duration),
                                  N = sum(antskad),
                                  C = sum(skadkost)),
                              by = vard.faktor]
    n.levels <- nlevels(motorc.andra[[vard.faktor]])
    contrasts(motorc.andra[[vard.faktor]]) <-
```

```
        contr.treatment(n.levels)[rank(-totals[["D"]],
 ties.method = "first"), ]
    data.frame(duration = totals[["D"]],
               n.claims = totals[["N"]],
               skadkost = totals[["C"]])
}
## Kopplar ihop kolumnvis
mangd24 <- cbind(mangd24, new.cols)
rm(new.cols)

## Skapar modellen för frekvensen
model.skadefrekvens <-
    glm(antskad ~
        vard.1 + vard.2 + vard.3 + vard.4 + offset(log(duration)),
        data = motorc.andra[duration > 0], family = poisson)

## coef och rels-biten som i tidigare uppgifter:
rels <- coef( model.skadefrekvens )
rels <- exp( rels[1] + rels[-1] ) / exp( rels[1] )
mangd24$rels.skadefrekvens <-
    c(c(1, rels[1:6])[rank(-mangd24$duration[1:7],
ties.method = "first")],
      c(1, rels[7:12])[rank(-mangd24$duration[8:14],
ties.method = "first")],
      c(1, rels[13:14])[rank(-mangd24$duration[15:17],
ties.method = "first")],
      c(1, rels[15:16])[rank(-mangd24$duration[18:20],
ties.method = "first")])

## Skapar modellen för kravstorleken
model.skadestorlek <-
    glm(skadkost ~ vard.1 + vard.2 + vard.3 + vard.4,
        data = motorc.andra[skadkost > 0,],
        family = Gamma("log"), weights = antskad)
```

```
rels <- coef( model.skadestorlek )
rels <- exp( rels[1] + rels[-1] ) / exp( rels[1] )
mangd24$rels.skadestorlek <-
    c(c(1, rels[1:6])[rank(-mangd24$duration[1:7],
ties.method = "first")],
      c(1, rels[7:12])[rank(-mangd24$duration[8:14],
ties.method = "first")],
      c(1, rels[13:14])[rank(-mangd24$duration[15:17],
 ties.method = "first")],
      c(1, rels[15:16])[rank(-mangd24$duration[18:20],
ties.method = "first")])

## Slutligen räknar vi ut riskpremien som produkten av
## kravfrekvensen och kravstorleken:
mangd24$riskpremie <- with(mangd24,
rels.skadefrekvens * rels.skadestorlek)

## Konverterar till data.table
library("data.table")
mangd24 <- data.table(mangd24)

## Sparar datat
save(mangd24, file = "mangd24.RData")

## Skriver ut resultatet
print(mangd24[,
               list(vard.faktor, klass, duration, n.claims,
                    skadkostK = round(skadkost / 1000),
                    relativiteter, riskpremie)],
      digits = 3)
```

	vard.faktor	klass	duration	n.skador	skadkostK	relativiteter	riskpremie
1:	Zon	1	6205	183	5540	7.678	6.23938
2:	Zon	2	10103	167	4811	4.227	3.17895
3:	Zon	3	11677	123	2523	1.336	0.99315
4:	Zon	4	32628	196	3775	1.000	1.00000
5:	Zon	5	1582	9	105	1.734	0.15542
6:	Zon	6	2800	18	288	1.402	0.22335
7:	Zon	7	241	1	1	1.402	0.00176
8:	MC-klass	1	5190	46	993	0.625	0.39465
9:	MC-klass	2	3990	57	883	0.769	0.53593
10:	MC-klass	3	21666	166	5372	1.000	1.00000
11:	MC-klass	4	11740	98	2192	1.406	0.57397
12:	MC-klass	5	13440	149	3297	1.875	1.41279
13:	MC-klass	6	8880	175	4161	4.062	4.87620
14:	MC-klass	7	331	6	145	6.873	0.60186
15:	Fordonsålder	1	4955	126	4964	2.000	4.71500
16:	Fordonsålder	2	9754	145	5507	1.200	2.02124
17:	Fordonsålder	3	50528	426	6570	1.000	1.00000
18:	Bonusklass	1	19893	207	4558	1.250	0.79746
19:	Bonusklass	2	9616	121	3627	1.125	0.60356
20:	Bonusklass	3	35728	369	8857	1.000	1.00000

Figur 4.3.12: Ny tariff jämfört med den ursprungliga tariffen från 1995

Kolumnen längst till höger (”riskpremie”) är de uträknade relativiteterna för riskpremien, medan (”relativiteter”) är motsvarande för tariffen från 1995. Utöver dessa resultat har vi med totala varaktigheten (”duration”), totala antalet krav (”n.skador”) samt totala kravkostnaden i tusental (”skadkostK”).

Resultatet verkar rimligt. Vid höga varaktigheter fås som väntat ganska liknande estimat för relativiteterna, medan låga varaktigheter ger en stor spridning. Till exempel, för zon 7 har vi endast ett krav, vilket inte går att analysera. Ett alternativ kan vara att sammanslå zon 7 tillsammans med en annan tillräckligt homogen zon.

Litteraturförteckning

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

[2] http://statmath.wu-wien.ac.at/courses/heather_turner/glmCourse_001.pdf (Hämtad 04.01.2018)

[3] http://web.abo.fi/fak/mnf/mate/kurser/sannolik/anteckningar_2010.pdf (Hämtad 04.01.2018)

[4] http://staff.math.su.se/esbj/GLMbook/moppe.sas (Hämtad 04.01.2018)

[5] https://en.wikipedia.org/wiki/Cumulant#Some_properties_of_the_cumulant_generating_function (Hämtad 04.01.2018)

[6] http://www.cybaea.net/Journal/2012/03/01/R-code-for-Chapter-1-of-Non_Life-Insurance-Pricing-with-GLM/ (Hämtad 04.01.2018)

[7] http://www.cybaea.net/Journal/2012/03/13/R-code-for-Chapter-2-of-Non_Life-Insurance-Pricing-with-GLM/ (Hämtad 04.01.2018)

[8] https://www.ma.utexas.edu/users/gordanz/notes/introduction_to_stochastic_processes.pdf (Hämtad 04.01.2018)

[9] https://www.r-bloggers.com/r-function-of-the-day-tapply-2/ (Hämtad 04.01.2018)

[10] https://magesblog.com/post/2013-03-12-how-to-use-optim-in-r/ (Hämtad 04.01.2018)

[11] https://en.wikipedia.org/wiki/Coefficient_of_variation (Hämtad 04.01.2018)