Masteroppgave i statistikk

GAMLSS-modeller i bilforsikring

Hallvard Røyrane-Løtvedt
Kandidatnr. 160657

UNIVERSITETET I BERGEN
MATEMATISK INSTITUTT

Veileder: Hans Julius Skaug

1. Juni 2012

GAMLSS-modeller i bilforsikring

1 Sammendrag

I denne oppgaven tester jeg ulike modeller for prediksjon av total skadeutbetaling fra forsikringsselskap til forsikringstaker i et poliseår. Modellene som testes hører til rammeverket *Generalized Additive Models for Location, Shape and Scale* – GAMLSS – introdusert av Rigby og Stasinopoulos (2001). Data brukt i oppgaven er hentet fra et norsk forsikringsselskap, og består av informasjon om poliser og skader i bilforsikring i årene 2000-2005. Ved hjelp av kun 3 forklaringsvariabler; årstall, bilalder og personalder, viser jeg i denne oppgaven at valg av statistisk modell er avgjørende for prediksjonene av skadeutbetalingen (kapittel 9). Videre tester jeg ut hvordan modellprediksjonene kan brukes til å lage en realistisk prismodell, og hvordan prismodellen gir ulike resultater for de ulike prediksjonsmodellene (kapittel 10).

Total skadeutbetaling deles naturlig inn i *skadefrekvens* og *skadepris*. Jeg tester i oppgaven både modeller som modellerer disse separat, og modeller som modellerer total skadeutbetaling direkte. Jeg vil argumentere for at de direkte modellene er å foretrekke. Modellen som anbefales er en *Zero-Adjusted Inverse Gaussian* – ZAIG-modell, der forklaringsvariablenes funksjonelle form er valgt slik at AIC blir så lav som mulig. En ZAIG-fordelt stokastisk variabel tar verdien 0 med sannsynlighet ψ, og følger en Invers-Gaussisk-fordeling med sannsynlighet $(1-\psi)$. Skadepriser er såpass skjevt fordelt at det må en ekstremt skjev sannsynlighetsfordeling, som den Invers-Gaussiske, til, for å beskrive dem. Jeg vil også i oppgaven argumentere for at valg av sannsynlighetsfordeling har stor betydning for kvaliteten på prediksjonene.

Forord

Jeg vil rette en stor takk til min veileder Hans Julius Skaug for god og konstruktiv kritikk under hele skriveprosessen. Videre vil jeg gjerne takke de dyktige foreleserne ved matematisk institutt på UiB for å ha vist meg hvor interessant og faglig utfordrende statistikkfaget, og spesifikt forsikringsmatematikk, kan være. Jeg vil også takke analyseavdelingen i Tryg forsikring, for å ha lært meg utrolig mye om forsikringsfaget. En takk går også til mine foreldre, Knut Løtvedt og Berit Anderssen, for gjennomlesing og konstruktiv kritikk. Sist, men ikke minst, vil jeg takke min kone, Lene Kristin Røyrane-Løtvedt, for gjennomlesing, gode råd og hjelp til å forbedre språket i oppgaven.

Innholdsfortegnelse

Tabeller

Figurer

2 Innledning

2.1 Bakgrunn

Differensiert prising i skadeforsikring er et tema det er blitt skrevet mye om innenfor forsikringsmatematisk litteratur. Den totale utbetalingen fra forsikringsselskap til forsikringstaker i et poliseår, *U*, er det sentrale tallet man ønsker å predikere. Imidlertid er *U* en vanskelig stokastisk variabel å modellere, ettersom den er sammensatt av to svært ulike stokastiske elementer: skadefrekvens[1] og skadepris**.** Går man langt tilbake i tid var datasettene i forsikring ofte av dårlig kvalitet, hvilket gav usikre estimater og prediksjoner (Weisberg og Tomberlin 1982). Imidlertid har man med moderne, sofistikerte IT-verktøy i stor grad overkommet dette problemet (Bortoluzzo et al. 2011). Heller et al. (2006) skriver at mye fokus i aktuarlitteraturen er gitt til ulike sannsynlighetsfordelinger for skadepris. Hogg og Klugman (1984) nevnes som et eksempel på dette. Mange forskere har bygget regresjonsmodeller for skadepris, der skadeprisen predikeres på bakgrunn av forklaringsvariabler. Et eksempel her er Haberman og Renshaw (1996). Disse regresjonsmodellene er imidlertid kun relevante for den gruppen forsikringspoliser som har hatt minst 1 skade i observasjonsperioden. (Heller et al. 2006). Dersom slike regresjonsmodeller brukes til prising av forsikringspoliser, uten samtidig å ta hensyn til skadefrekvensen (eller skadesannsynligheten), gir det ikke risikoriktig[2] pris. Årsaken er at når man ikke tar hensyn til skadefrekvensen, er det ekvivalent med å sette den lik for alle kunder. Jørgensen og de Souza (1994) foreslår å modellere *U* som en Poisson-sum av gammafordelte skadepriser. Dette kan gjøres ved en variant av Tweediefordelingen. (Bortoluzzo et al. 2011). Et problem ved denne fremgangsmåten er at sannsynligheten for 0 skader ikke kan modelleres eksplisitt som en funksjon av forklaringsvariabler (Heller et al 2006). Ved å ta i bruk GAMLSS-modellering, slik jeg gjør, kan man imidlertid la en hvilken som helst fordelingsparameter avhenge direkte og eksplisitt av forklaringsvariabler. Dette gjelder også parametere for nullsannsynlighet.

[1] Jeg vil i denne oppgaven bruke begrepene *skadefrekvens* og *antall skader* om hverandre. Begge skal forstås som antall skader per poliseår, *A*.

[2] *Risiko* skal her forstås i lys av sannsynlighetsfordelingen til *U* for hver enkelt kunde. Dersom er kunde har "Høy risiko", betyr det at sannsynlighetsfordelingen til *U* for denne kunden har negative egenskaper, sett fra forsikringsselskapets ståsted. Disse egenskapene er typisk høy forventningsverdi og høye kvantiler.

Heller et al. (2006) introduserer *Zero Adjusted Inverse Gaussian* – ZAIG-fordelingen for modellering av *U*. Denne modellen bygges opp under GAMLSS-rammeverket (se delkapittel 3.4), som jeg også vil ta i bruk i denne oppgaven. Bortoluzzo et al. (2011) tester ZAIG-fordelingen mot Tweediefordelingen på et datasett for bilforsikring, og konkluderer med at ZAIG-fordelingen gir en modell som bedre beskriver risikoen, og er mer velegnet til prising av forsikringspoliser. GAMLSS-rammeverket, som jeg bruker i denne oppgaven, er relativt nytt. Imidlertid er det publisert en rekke vitenskapelige artikler der GAMLSS anvendes i studiet av kvantitative fenomener. Mens jeg skriver dette er det kun Heller et al. (2006) og Bortoluzzo et al. (2011) som har brukt GAMLSS-metodikk for å modellere *U* i skadeforsikring.[3] Begge har hovedfokus på ZAIG-fordelingen. Jeg finner det derfor interessant å teste potensialet til GAMLSS som rammeverk for å modellere *U*, også ved andre fordelinger.

I denne oppgaven går jeg bredt ut og tester flere mulige modelleringsstrategier. Den klassiske modellen der skadefrekvensen og skadeprisen modelleres separat, er en kandidat, og settes opp mot ZAIG-modellen foreslått av Heller et al. (2006) og Bortoluzzo et al. (2011). I tillegg testes den nært beslektede ZAGA-modellen. For hver modell tester jeg også ut undergrupper med ulik grad av fleksibilitet. All modelltesting gjøres på et stort datasett fra et skadeforsikringsselskap, med data fra årene 2000 til 2005.

2.2 Motivasjon

Jeg jobber selv i forsikringsbransjen og har derfor en viss kjennskap til hvilke problemstillinger det er fokus på i bransjen, og hvordan det tenkes om løsning av problemene. Min hovedmotivasjon for å skrive denne oppgaven er et ønske om å bidra til å utvikle og/eller utprøve statistiske metoder som kan brukes i praksis, i et forsikringsselskap. Riktig prising av forsikringspolisene er essensielt for et forsikringsselskap. Det kan sies å være et gjennomgående tema for oppgaven.

[3] En komplett liste, per 06.05.2012 av alle vitenskapelige artikler publisert, der GAMLSS brukes, finnes på http://gamlss.org/images/stories/bibtex/gamlssrefs.pdf.

2.2.1 *Riktig* prising

Forsikringsbransjen i Norge og internasjonalt er preget av hard konkurranse om kundene. Produktet forsikringsselskapene tilbyr er dekning av store uforutsette utgifter til skader som ikke er selvforskyldt. Det er selvsagt forskjeller mellom forsikringsselskapene, med hensyn til dekningsvilkår, kundeservice, avtaler med leverandører for skadebehandling etc. Imidlertid er dette ofte marginale forskjeller sett fra forsikringstakers ståsted. Når produktene som tilbys er såpass like fra et forsikringsselskap til et annet, vil ofte pris være det primære kriteriet kunden baserer sitt valg av forsikringsselskap på. Dette fører til at forsikringsselskapene er svært opptatt av konkurransedyktig, og ikke minst *riktig* prising.

Riktig prising er et langt mer komplisert begrep i forsikringsbransjen enn i de fleste andre bransjer. Den største delen av forsikringsselskapets utgifter er skadeutbetalinger. Disse er av natur usikre (stokastiske), og kan potensielt ruinere forsikringsselskapet[4], dersom det ikke er nok penger til å dekke skadene. Forsikringsselskapene er pålagt ved lov[5] å sette av nok penger til å dekke forventede økonomiske forpliktelser. Disse pengene må hentes inn som forsikringspremie av kundene. Det er derfor grenser for hvor lav pris man kan sette. En mulig prisingsstrategi er å tilby lik pris for alle kunder. Dette gir enkle og oversiktlige priser, og det kan argumenteres for at det er solidarisk og rettferdig, ettersom skadene vanligvis ikke er selvforskyldte. Imidlertid er det et statistisk veldokumentert faktum at ulike kunder har ulik risiko. Lik pris for alle vil derfor medføre at lavrisikokunder subsidierer høyrisikokunder. Dersom et forsikringsselskap opererer med differensierte, risikoriktige priser, mens et annet opererer med lik pris for alle, vil lavrisikokundene få rimeligere pris hos selskapet som differensierer, og dermed ha et økonomisk insentiv til å bytte forsikringsselskap. Høyrisikokunder i selskapet som prisdifferensierer, vil også ha et økonomisk insentiv til å bytte til selskapet som opererer med lik pris for alle. På sikt vil dette kunne føre til en porteføljeglidning der selskapet som differensierer, sitter igjen med lavrisikokunder, og selskapet som tilbyr lik pris sitter igjen med høyrisikokunder. Selskapet som differensierer prisene vil være langt mer lønnsomt, både fordi skadeutbetalingene vil være færre og mer stabile, og fordi omkostningene til skadebehandling blir redusert. I praksis differensierer alle forsikringsselskapene prisene sine, basert på ulike kriterier.

[4] Se for eksempel Sundt (1999: kapittel 10) for mer om sannsynligheten for ruinering av selskapet.

[5] Se nyeste forskrifter på http://www.lovdata.no/ltavd1/filer/sf-20111221-1480.html

2.2.2 Statistisk modellering

Gitt konkurransesituasjonen, er det klart at forsikringsselskapene må differensiere prisene etter de ulike kunders risikoprofil. Dette fører til et behov for å bygge best mulige statistiske modeller for utbetalingen til kundene. Konkurransesituasjonen i den norske forsikringsbransjen har spisset seg til ytterligere etter at finansportalen.no[6] ble lansert i 2011. Det ble da enklere for kundene å sammenlikne selskapenes priser direkte. Følgelig er behovet for gode statistiske modeller høyere enn noensinne. Differensiert prising gir kun ønsket effekt dersom differensieringen treffer riktig. Det betyr at de statistiske modellene må kunne "spå fremtiden" med best mulig treffsikkerhet. Mer spesifikt kan man si at risikoriktig prising er avhengig av å kunne beskrive sannsynlighetsfordelingen til *U* (total utbetaling per poliseår) mest mulig realistisk. Forsikringsselskapet som klarer dette har et klart konkurransefortrinn.

2.3 Målsetning

Mitt mål med denne oppgaven er å sammenlikne ulike prediksjonsmodeller i bilforsikring. Mer spesifikt ønsker jeg å predikere total utbetaling per poliseår, *U*. Dette er antall kr forsikringsselskapet betaler til forsikringstaker for å dekke skader i løpet av et poliseår.[7] For polise *i*, er størrelsen U_i gitt ved

$$U_i = \sum_{k=0}^{A_i} S_{i,k} \tag{1}$$

der $S_{i,k}$ er skadeprisen på skade *k* for polise *i*, og A_i er antall skader for polise *i*. Ved å innføre konvensjonen $S_{i,0} = 0$, er U_i fullt definert ved (1). Jeg vil ta i bruk GAMLSS-rammeverket (se delkapittel 3.4) til å bygge modellene. Det er et meget fleksibelt modelleringsrammeverk, der responsvariabelens sannsynlighetsfordeling tillates å avhenge av forklaringsvariabler ved en egen formel for hver fordelingsparameter. Først bygger jeg modeller der skadefrekvens[8] og skadepris modelleres hver for seg. Estimatene kobles så sammen for å predikere total utbetaling *U*. Videre bygger jeg modeller der total utbetaling modelleres direkte. Jeg vil teste alle disse modellene parallelt, og drøfte fordeler og ulemper ved dem. Samtidig er det et mål at en modell utpekes som den foretrukne. I drøftingen vil jeg

[6] Finansportalen er opprettet av forbrukerrådet som en tjeneste for sammenlikning av finans- og forsikringsprodukter.

[7] Egenandelen dekker forsikringstaker selv. Den inngår derfor ikke i *U*.

forsøke å tenke praktisk, og konkretisere resultatene i et realistisk forsikringsperspektiv. Tilgjengelige forklaringsvariabler i denne oppgaven er årstall, bilalder og personalder (se delkapittel 4.3). I virkeligheten har forsikringsselskapene vanligvis tilgang til langt flere forklaringsvariabler enn dette. Jeg forsøker imidlertid å få mest mulig forklaringskraft ut av de tilgjengelige forklaringsvariablene. En sekundær målsetning er å drøfte hvorvidt, i hvilken grad og på hvilken måte disse forklaringsvariablene påvirker skadefrekvens, skadepris og total utbetaling.

2.4 Bruk av R

Enhver utregning i denne oppgaven er utført i dataprogrammet R (se r-project.org). Dette er et gratis statistikkprogram som brukes av akademiske fagmiljøer verden rundt. Estimering av modellparametere er i denne oppgaven utført ved bruk av GAMLSS-pakken (se gamlss.org). Denne pakken kjører i R og gir brukeren mulighet til å estimere parameterne i svært fleksible regresjonsmodeller uten å måtte skrive kildekode for alle stegene i algoritmene. R, og i noen tilfeller GAMLSS-pakken, er også brukt til å produsere figurene og grafene i oppgaven.

2.5 Notasjon og konvensjoner

Store latinske bokstaver i kursiv som A, B, X, Y brukes for stokastiske variabler. Små, latinske bokstaver i kursiv som a, b, x, y brukes for observerte verdier, matematiske funksjoner eller realiseringen av stokastiske variabler. Små, fete bokstaver som $\mathbf{a}$, $\mathbf{b}$, $\mathbf{x}$, $\mathbf{y}$ brukes for vektorer. Små, greske bokstaver som α, β, ξ, υ brukes for parametere. Følgende vanlige engelske forkortelser og termer fra statistisk litteratur brukes hyppig:

- PDF: sannsynlighetstetthet
- PMF: punktsannsynlighet
- GLM: Generalized Linear Model
- GAM: Generalized Additive Model
- GAMLSS: Generalized Additive Model for Location, Shape and Scale
- likelihood: sannsynlighet gitt observerte verdier og gjeldende antagelser
- ML: maximum likelihood
- NEGBIN: negativ binomisk fordeling
- ZIP: "Zero-inflated" Poissonfordeling

- ZAGA: “Zero-adjusted”-gammafordeling
- ZAIG: “Zero-adjusted” invers-gaussisk fordeling
- i.i.d.: Independent and identically distributed (uavhengig og identisk fordelt)

Definisjonsmengder for sannsynlighetsfordelinger som er velkjente fra statistisk litteratur sløyfes av plasshensyn. Definisjonsmengder for mer spesielle uttrykk tas med etter behov. Jeg vil for enkelhets skyld bruke *f* som både PMF og PDF, og ikke skille mellom disse der det ikke er behov for det. Her er en liste over andre konvensjoner jeg bruker oppgaven gjennom:

- Indikatorfunksjoner skrives som $I(A)$ der *A* er et kriterium. Dersom *A* er oppfylt, tar indikatorfunksjonen verdien 1, og ellers 0.
- θ brukes som benevnelse på generelle parametere. Dersom jeg omtaler en rekke sannsynlighetsfordelinger, med ulike parametere som en enhet, bruker jeg for eksempel θ som benevnelse på parameterne i alle fordelingene.
- Tegnet Φ brukes kun for fordelingsfunksjoner til standardnormalfordelingen. Det betyr at dersom en stokastisk variabel, *Z*, er standardnormalfordelt, gjelder $\Phi(z) = P(Z \leq z)$.
- *Designvektor* er å forstå som en rad i *designmatrisen* (som inneholder alle observasjoner av forklaringsvariablene, slik de inngår i modellen). Det vil si at en designvektor inneholder alle forklaringsvariablene, i deres gjeldende funksjonelle form, for en enkel observasjon.
- Når jeg skriver $\log(x)$ mener jeg den naturlige logaritmen til *x*, slik at $e^{\log(x)} = x$.
- Jeg bytter på å skrive $\exp(x)$ og e^x for å uttrykke eksponentialfunksjonen av *x*.
- For å spare plass vil jeg noen ganger bruke vektornotasjon når jeg skriver lineære uttrykk. Det betyr for eksempel at $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ vil kunne skrives $\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} \boldsymbol{\beta}$.
- Når jeg navngir en statistisk modell, velger jeg første bokstav i navnet responsvariabelen som første bokstav, og en forkortelse for navnet på sannsynlighetsfordelingen utgjør resten av modellnavnet. Antall skader, *A*, modellert ved Poissonmodell, kalles for eksempel APOI.

---

3 Teori

I dette kapitelet forsøker jeg å presentere det teoretiske grunnlaget for modelleringen i kapitler 6-9. Dette gjøres ved å introdusere det nødvendige begrepsapparatet, samt de anvendte metoder. Jeg starter med en presentasjon av det populære AIC-kriteriet for modellvalg. Videre introduseres GAMLSS-metodikken ved først å presentere dens forløpere, GLM og GAM. Jeg introduserer så alle sannsynlighetsfordelinger som tas i bruk i kapitlene 6-9. Kapittel 3 kan ses som en presentasjon av de nødvendige teoretiske verktøy som tas i bruk i senere kapitler.

3.1 AIC

Når man bygger en statistisk modell, er følgende spørsmål alltid relevant:

- Hvor fleksibel skal modellen være?[9]

Modelltilpasningen blir bedre, jo flere parametere man estimerer. Imidlertid blir også den samlede usikkerheten større, ettersom det introduseres mer usikkerhet for hver parameter som estimeres. Det antas at hver enkelt aktuell parameter har en "sann" verdi som er ukjent. Hver gang det estimeres en størrelse brukes en estimator (vanligvis en ML-estimator). Estimatorer må ses som stokastiske variabler med tilhørende sannsynlighetsfordelinger, der de "sanne" parameterstørrelsene inngår i PDF/PMF. Så lenge disse sannsynlighetsfordelingene tillater variasjon overhodet, må det tas høyde for at estimatene kan, og vil, bomme på de "sanne" parameterverdiene. Om estimatene treffer eller bommer, og eventuelt hvor mye de bommer med, har man i prinsippet ingen mulighet til å finne ut, med mindre det samles inn nye data. Enkelt sagt vil modellen akkumulere estimatorusikkerhet for hver parameter som estimeres. Det er med andre ord et dilemma mellom tilpasning og treffsikkerhet i estimatene. Dilemmaet er meget velkjent og er relevant for all statistisk modellbygging.

Akaike (1974) introduserte størrelsen *Akaike's Information Criterion* – AIC – for å løse dette dilemma. Akaike (1974) viser, ved hjelp av blant annet *Kullback–Leibler divergens* og informasjonsteori, at dersom man har to kandidatmodeller, vil modellen med lavest AIC-verdi

[9] *Fleksibilitet* er et vidt begrep, men kan operasjonaliseres i en modelleringskontekst, ved å la *grad av fleksibilitet* forstås som *antall frie parametere* eller *antall frihetsgrader*. I denne oppgaven vil jeg bruke fleksibilitetsbegrepet i denne betydningen. Jeg vil for eksempel mene *høyt antall frie parametere i modellen* når jeg skriver *svært fleksibel modell*.

være å fortrekke, ettersom den gir relativt mindre forventet *informasjonstap* enn modellen med høyest AIC. *Informasjonstap* skal her forstås relativt mellom den ukjente prosessen som genererer de observerte data, og en statistisk modell som representerer denne prosessen. AIC er definert som

$$\text{AIC} = 2p - 2l$$

der p er antall estimerte parametere, og l er log-likelihooden i modellen. Denne enkle formelen tas i bruk gjennomgående i oppgaven som relativt kriterium for modellvalg. AIC er i prinsippet kun asymptotisk gyldig, når antall observasjoner, n, går mot ∞. Burnham og Anderson (2002) anbefaler å bruke AIC_c i stedet for AIC, for å korrigere for antall observasjoner. AIC_c er definert som

$$\text{AIC}_c = \text{AIC} + \frac{2p(p+1)}{n-p-1}.$$

Det fremgår av uttrykket at $\text{AIC}_c \underset{n\to\infty}{\to} \text{AIC}$, slik at vanlig AIC likevel kan forsvares når antall observasjoner er høyt. Datasettet i denne oppgaven har såpass mange observasjoner (63 165 poliseår og 9 396 skader) at feilkilden ved å bruke AIC i stedet for AIC_c er minimal. Dersom man for eksempel har en modell for skadefrekvens med hele 30 parametere, vil forskjellen mellom AIC_c og AIC være ca. 0,03, hvilket er neglisjerbart. Jeg velger derfor å bruke vanlig AIC i den resterende del av oppgaven.

3.2 GLM

3.2.1 GLM-rammeverket

Nelder og Wedderburn (1972) introduserte *Generalized Linear Models* – GLM. GLM er et sammenhengende rammeverk for statistiske modeller der responsvariabelen Y_i ses som en uavhengig, stokastisk variabel, med fordeling $f(\mu_i, \phi)$, der μ_i avhenger av forklaringsvariabler $\mathbf{x}_i$ gjennom link-funksjonen g, slik at $g(\mu_i) = \eta_i$, der $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ kalles den lineære prediktor. η_i regnes som lineær ettersom den er lineær i koeffisientene $\boldsymbol{\beta} = \{\beta_0, ..., \beta_p\}$. I det opprinnelige GLM-rammeverket må fordelingen $f(\mu_i, \phi)$ tilhøre den eksponentielle familie. Sannsynlighetsfordelingene reparameteriseres slik at forventningsverdien $\text{E}(Y_i)$ tilsvarer en egen parameter μ_i. Det er kun lokasjonsparameteren

μ_i som kobles mot forklaringsvariabler i GLM. Det vil blant annet si at fordelingens varians, skjevhet og kurtose kun indirekte, gjennom μ_i, avhenger av forklaringsvariabler.

Sannsynlighetsfordelingene for *f*, som brukes i GLM-analyse, kan alle skrives på formen

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

der θ_i kalles "den naturlige parameteren" og ϕ er en sekundær parameter som ikke påvirker μ_i. Det kan vises at forventningen og variansen til Y_i kan skrives

$$\mathrm{E}(Y_i) = \frac{\partial a}{\partial \theta_i}\ , \quad \mathrm{Var}(Y_i) = \phi \frac{\partial^2 a}{\partial \theta_i^2}$$

Dette impliserer videre at $\mathrm{Var}(Y_i) = \phi \frac{\partial a}{\partial \theta_i} \mathrm{E}(Y_i) = \phi V(\mu_i)$. Med andre ord vil variansen til Y_i være en funksjon av forventningen til Y_i i GLM. Slik kan også variansen (indirekte) avhenge av forklaringsvariabler. (de Jung og Heller 2008:35-37).

3.2.2 GLM-estimering

Parameterestimeringen i GLM utføres vanligvis ved ML-maksimering. For de fleste fordelinger i den eksponentielle familie er det ikke mulig å uttrykke ML-estimatorene som en kombinasjon av vanlige matematiske funksjoner. Derfor benyttes som hovedregel Newton-Raphson-iterasjon eller Fisher-scoring, slik det er beskrevet for eksempel av Dobson og Barnett (2008:64-66). Man starter med et forslag til parametervektor $\boldsymbol{\beta}^{(0)}$ og finner $\boldsymbol{\beta}^{(m)}, m = 1,2,...$ ved den iterative estimeringslikningen

(3) $$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \left[\Im^{(m-1)} \right]^{-1} \mathbf{u}^{(m-1)}$$

der $\Im$ er Fishers informasjonsmatrise og $\mathbf{u}$ er score-vektoren $\mathbf{u} = \frac{\partial l}{\partial \boldsymbol{\beta}}$. $\boldsymbol{\beta}$ oppdateres helt til $\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)} < \boldsymbol{\varepsilon}$ der $\boldsymbol{\varepsilon}$ er en vektor med konvergensgrenser. Da settes $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m+1)}$ og parametervektoren er ferdig estimert. Som det fremgår av den iterative likningen (3), maksimeres log-likelihooden i GLM kun med hensyn på parametervektor $\boldsymbol{\beta}$. Dispersjonsparameteren ϕ regnes som sekundær, og estimeres først etter at $\boldsymbol{\beta}$ er estimert. Det

er heller ikke mulig innenfor tradisjonell GLM-analyse å la ϕ avhenge av forklaringsvariabler. Verdt å merke seg er også at link-funksjonen *g*, som kobler forventningsverdien til forklaringsvariablene, må være en monoton, deriverbar funksjon ettersom Fisher-scoring krever deriverbar likelihood.

3.3 GAM

3.3.1 GAM-rammeverket

Hastie og Tibshirani (1990) introduserer *Generalized Additive Models* - en utvidelse av GLM - ved å erstatte den lineære prediktoren $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$, med en mer generell additiv prediktor $\eta_i = \beta_0 + \sum_{j=1}^{p} s_j(x_{i,j})$, der s_j er funksjoner av forklaringsvariablene. Det er mange ulike kandidater for funksjonene s_j. Hastie og Tibshirani (1990) foreslår å la s_j være *cubic splines*. I det enkle tilfelle der det kun er en forklaringsvariabel, *x*, la observasjonene være sortert i stigende rekkefølge for *x*, slik at man kan skrive $x_{(i)} = x_i$ for alle *i*. Da kan *cubic spline* defineres slik:

- En *cubic spline*, *s*, er en stykkevis definert funksjon med definisjonsmengde $[x_1, x_n]$. Definisjonsmengden kan deles opp i $n-1$ disjunkte subintervaller $[x_{i-1}, x_i]$ der $x_{\min} = x_1 < x_2 < ... < x_{n-1} < x_n = x_{\max}$. *s* er gitt ved

$$s(x) = \begin{cases} P_1(x), & x_1 \le x < x_2, \\ P_2(x), & x_2 \le x < x_3, \\ ... \\ P_{n-1}(x), & x_{n-1} \le x < x_n \end{cases}$$

  der alle P_i er tredjegrads polynomfunksjoner.[10]

Det spesifikke uttrykket til *s* bestemmes ved å minimere (4), gitt ved

(4) $$\sum_{i=1}^{n} (y_i - s(x_i))^2 + \lambda \int_{x_1}^{x_n} \left(\frac{d^2 s}{dx^2} \right)^2 dx$$

[10] Denne måten å definere *cubic splines* på er i stor grad hentet fra http://en.wikipedia.org/wiki/Spline_(mathematics)

Kvadratsummen $\sum_{i=1}^{n}(y_i - s(x_i))^2$ er det klassiske målet på tilpasning, mens $\lambda\int_{x_1}^{x_n}\left(\frac{d^2s}{sx^2}\right)^2 dx$ kalles en "spline-smoother". Dette leddet er med for å tilføre "glatthet" til *s*. Spline-funksjonen *s* tillates å skifte parametrisk form fra et subintervall til det neste. Glatthetsparameteren λ, som må være positiv, gir en straff for fleksibilitet, ettersom integralet av den andrederiverte øker med koeffisientene til de høyere ledd i polynomene P_i. Dette kan løst beskrives som at λ belønner "glatthet", eller "linearitet".

3.3.2 GAM-estimering

(4) kan generaliseres til *p* forklaringsvariabler $x_1,...,x_p$, ved å bruke den såkalte "backfitting"-teknikken som ble introdusert av Breiman og Friedman (1985). Kort forklart består backfitting i følgende steg:

1. Sett estimat på konstantleddet til $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$ og
2. Sett første estimat på alle spline-funksjoner til $\hat{s}_j^{(0)} = 0$
3. Minimer $\sum_{i=1}^{n}\left(y_i - \sum_{k\neq j}\hat{s}_k(x_{i,k}) - s_j(x_{i,j})\right)^2 + \lambda_j \int_{x_i}^{x_{(n)}} \frac{d^2 s_j}{sx_j^2} dx_j$ for alle s_j. Resultatet er estimater $\hat{s}_j^{(1)}$.
4. Sentrer spline-estimatene ved å sette $\hat{s}_j^{(2)} = \hat{s}_j^{(1)} - \frac{1}{n}\sum_{i=1}^{n}\hat{s}_j^{(1)}(x_{i,j})$.
5. Repeter steg 3 og 4 til alle $\hat{s}_j$ konvergerer mot stabile størrelser.

Backfitting-algoritmen, slik den her er presentert, er en oppskrift på å estimere spline-funksjoner, s_j, som tilpasninger til punkter i et *p*-dimensjonalt plan. Dette gir en enkel illustrasjon av backfitting-teknikken. Når man estimerer spline-funksjonene i en GAM-setting, søker man å maksimere likelihooden til alle $Y_i|\mathbf{x}_i$. Estimeringslikningene kan skrives som IRLS, *iteratively reweighted least squares*. Med andre ord kan maksimering av likelihooden ses som en anvendelse av minste kvadrats metode rundt punkter i *p*-planet. Man kan dermed estimere ved å bytte ut IRLS med backfitting-algoritmen. Detaljert utledning av GAM-estimering finnes i kapittel 6 i Hastie og Tibshirani (1990).

Hastie og Tibshirani (1990) viser at det er mulig å definere en "hyperparameter" $\Lambda(\lambda)$ som avhenger av λ og representerer "effektive parametere" eller "effektive frihetsgrader". Størrelsen Λ kan løst forstås som graden i et polynom, likt definert på hele definisjonsområdet, som nesten kunne erstattet spline-funksjonen. Estimeringen vil da kunne optimaliseres videre, ved også å estimere optimal verdi for hyperparameteren Λ. Tilpasningen til data, og dermed likelihooden, vil øke monotont med Λ. Følgelig kan det ikke brukes vanlig ML-estimering for Λ. I stedet vil minimering av AIC være et naturlig valg for å estimere Λ. GAMLSS-pakken i R har rutiner for å AIC-minimere Λ, slik at man kan få ut et estimat på optimal verdi av Λ.

3.4 GAMLSS

3.4.1 GAMLSS-rammeverket

Rigby og Stasinopoulos (2001) introduserte *Generalized Additive Models for Location, Shape and Scale*, GAMLSS, som en videre generalisering av GLM/GAM-rammeverket.[11] Der man i GLM-modeller kun tillater en parameter, lokasjonsparameteren μ_i, å avhenge av forklaringsvariabler, tillates også andre fordelingsparameterne å avhenge direkte av forklaringsvariabler i GAMLSS-rammeverket. En annen generalisering i GAMLSS er at rammeverket ikke krever at fordelingen til responsvariabelen Y_i skal tilhøre den eksponentielle familie. Forskning på teorien bak og anvendelser av GAMLSS, samt implementering av programvare i *R*, er en pågående prosess, som blant annet utføres av den internasjonale forskergruppen "The GAMLSS team".[12]

[11] Det er også utviklet rammeverk for modellering som ligger "mellom" GAM og GAMLSS. Blant annet ved å tillate opp til 2 parametere å avhenge av forklaringsvariabler. Presentasjonen her skal ikke tas som en komplett "tidslinje", men mer som noe modelleringshistorikk brukt for å presentere nøkkelideer.

[12] Disse er listet som medlemmer av "The GAMLSS team" per 11.04.2012:

- Dr. Tilemahos Efthimiadis (KEPE, Athen, Hellas)
- Prof. Paul Eilers (Erasmus University, Nederland)
- Dr. Nikolaos Georgikopoulos (KEPE, London Metropolitan University og New York University - Stern School of Business)
- Dr. Gillian Heller (Macquarie University, Australia)
- Dr. Vito Muggeo (University of Palermo, Italia)
- Dr. Bob Rigby (London Metropolitan University, Storbritania)
- Prof. Mikis Stasinopoulos (London Metropolitan University, Storbritannia)

GAMLSS-rammeverket, slik det er implementert i R per mai 2012, tillater at opp til 4 fordelingsparametere kan avhenge direkte av forklaringsvariabler. Hver fordelingsparameter kan ha hver sin link-funksjon og hver sin designmatrise. De mindre fleksible modellene innenfor GLM/GAM-rammeverket kan ses som spesialtilfeller av GAMLSS, der kun lokasjonsparameteren μ_i avhenger av forklaringsvariabler. GAMLSS er med andre ord et særdeles fleksibelt rammeverk for univariat statistisk modellering. Det er i hovedsak dette rammeverket jeg vil ta i bruk i denne oppgaven.

Rigby og Stasinopoulos (2009) definerer GAMLSS-modeller ved hjelp av følgende, meget generelle formulering der *i* er observasjonsnummer og *k* er nummeret på fordelingsparameteren:

$$Y_i \big| \boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}_i)$$

$$g_k(\theta_{i,k}) = \eta_{i,k} = \mathbf{x}_{i,k}^T \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{z}_{i,k}^T \boldsymbol{\gamma}_{j,k} \tag{5}$$

der f ikke trenger å tilhøre den eksponentielle familie. Parametervektoren $\boldsymbol{\theta}_i$ kan inneholde opptil 4 parametere, og koblingen mellom en av dem, $\theta_{i,k}$, og forklaringsvariablene, er som vist i (5). $\mathbf{x}_{i,k}$ og $\mathbf{z}_{i,k}$ er designvektorer, hver av dem spesialtilpasset til den spesifikke modellen. $\boldsymbol{\beta}_k$ er en koeffisientvektor, mens $\boldsymbol{\gamma}_{j,k}$ er en vektor kan være enten stokastisk (for å inkorporere "random effects"), eller en deterministisk spline-funksjon av forklaringsvariabler. Jeg vil ikke se på "random effects" i denne oppgaven. For denne oppgavens del tar jeg derfor i bruk følgende (også meget generelle) semi-parametriske modellformulering for GAMLSS der *s* er en *cubic spline*:

$$Y_i \big| \boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}_i)$$

$$g_k(\theta_{i,k}) = \eta_{i,k} = \mathbf{x}_{i,k}^T \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_{j,k}(x_{i,j,k})$$

Jeg vil i denne oppgaven kun bruke spline-funksjonene til tidlig testing av modellene, mens leddet $\mathbf{x}_{i,k}^T \boldsymbol{\beta}_k$ vil brukes gjennomgående.

- Dr. Vlasios Voudouris (LondonMet Business School, Storbritania)

- Dr Ardo van den Hout (Department of Statistical Science, University College London, Storbritania)

Kilde: http://gamlss.org/index.php?option=com_content&view=article&id=19&Itemid=10

3.4.2 GAMLSS-estimering

Estimering i GAMLSS gjøres ved å maksimere den "straffede" log-likelihooden l_p gitt ved

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{j,k}\left(s_{j,k}(\mathbf{x}_{j,k})\right)^T \mathbf{K}_{j,k}\left(s_{j,k}(\mathbf{x}_{j,k})\right)$$

der l er log-likelihood, $\lambda_{j,k}$ er en glatthets-parameter for spline-funksjon j og forklaringsvariabel k, mens $\mathbf{K}_{j,k}$ er en strukturert matrise. (Rigby og Stasinopoulos 2005:509-511). Likelihooden er straffet, i den forstand at noe fratrekkes likelihooden før den maksimeres. Tankegangen bak det å maksimere en straffet likelihood, i stedet for en ren likelihood, er nært beslektet med metodikken fra delkapittel 3.3. Dersom man har et ubestemt antall frihetsgrader i modellen vil maksimering av ren likelihood gi kraftig overparameteriserte modeller. Ribgy og Stasinopoulos (2005) løser dette problemet ved å innføre en straff for overparameterisering, når fleksible modeller som f.eks. inneholder random-effects-ledd eller spline-funksjoner skal estimeres. Det er ikke ukontroversielt å maksimere l_p for å estimere parameterne i GAMLSS. John A. Nelder anbefaler for eksempel å bruke *Restricted Maximum Likelihood* i stedet, da han hevder dette i større grad gir forventningsrette estimater. (Rigby og Stasinopoulos 2005:547).

Det er verdt å merke seg at i en full-parametrisk GAMLSS, der det ikke er noen spline-funksjoner, kollapser l_p til l, og det er i stedet den vanlige log-likelihooden som maksimeres. Ettersom fokus i denne oppgaven stort sett er på full-parametriske GAMLSS-modeller, beholdes metodikken med maksimering av l_p slik den er implementert i R, da dette er tilstrekkelig for oppgavens formål.

3.4.3 Estimeringsalgoritmer for GAMLSS

To algoritmer er implementert i GAMLSS-pakken i R (se gamlss.org) for maksimering av l_p : CG-algoritmen og RS-algoritmen. CG-algoritmen er en generalisert utgave av Cole og Green (1992)-algoritmen. Denne algoritmen bruker de førstederiverte, andrederiverte (Hessianmatrisen) og kryssderiverte av likelihood-funksjonen med hensyn på fordelingsparameterne $\boldsymbol{\theta}$. RS-algoritmen er utviklet av Rigby og Stasinopoulos, og gjør, i motsetning til CG-algoritmen, ikke bruk av de kryss-deriverte av likelihood-funksjonen. Dette

betyr at RS-algoritmen er bedre tilpasset i de tilfeller der parameterne $\boldsymbol{\theta}$ er informasjon-ortogonale på hverandre. Ortogonaliteten forekommer der forventningsverdiene til de kryss-deriverte av likelihood-funksjonen er 0. Rigby og Stasinopoulos (2005) gjennomgår begge algoritmene, og viser at de maksimerer l_p riktig. I modelleringer i denne oppgaven er både RS-algoritmen og CG algoritmen brukt gjennomgående, som en test på at begge algoritmer konvergerer mot de samme estimater.[13]

3.5 Sannsynlighetsfordelinger

Her følger en oversikt over alle sannsynlighetsfordelinger som blir tatt i bruk, eller omtalt, i senere kapitler. Jeg introduserer konvensjoner for hver av disse fordelinger, med hensyn til hvordan de parameteriseres, og hvilke bokstaver som benevner hvilke parametere. Den stokastiske variabel som følger hver enkelt fordeling skrives som *Y*, eller, i realisert form, *y*.

3.5.1 Normalfordelingen

Normalfordelingen er den mest kjente sannsynlighetsfordelingen i statistikkfaget, og er brukt som analyseverktøy i en lang rekke disipliner. Det er en kontinuerlig fordeling med definisjonsmengde $(-\infty,\infty)$. Normalfordelingen har 2 parametere, lokasjonsparameteren μ, og dispersjonsparameteren σ. PDF-kurven er perfekt symmetrisk og letthalet. PDF for denne fordelingen skrives på følgende form

$$f(y;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

PDF for normalfordelingen kan skrives om til formen (2), hvilket gir

$$f(y;\mu,\sigma)=\exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right)=\exp\left(c(y,\phi)+\frac{y\theta-a(\theta)}{\phi}\right).$$

der $c(y,\phi)=-\frac{y^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)$, $\phi=\sigma^2$, $\theta=\mu$ og at $a(\theta)=\theta^2/2$. Dette betyr at normalfordelingen er en del av den eksponentielle familie, og kan passe inn under det tradisjonelle GLM-rammeverket. Dersom man imidlertid ønsker at både μ og σ skal avhenge direkte av forklaringsvariabler, må man benytte det mer generelle GAMLSS-

[13] For alle modeller jeg tester oppnås konvergens mot samme verdier ved RS-, og CG-algoritmene.

rammeverket. Forventning og varians for normalfordelingen er $\mathrm{E}(Y)=\mu$ og $\mathrm{Var}(Y)=\sigma^2$. Dersom Y er normalfordelt med parametere μ og σ, vil jeg i det følgende kun skrive $Y \sim N(\mu,\sigma^2)$ for å indikere dette. Dersom $\mu=0$ og $\sigma=1$, følger Y en *standardnormalfordeling*.

3.5.2 Gammafordelingen

Gammafordelingen er ofte brukt i forsikringssammenheng for å beskrive skadepris (se for eksempel de Jong og Heller (2008: 120-125)). Det er en kontinuerlig fordeling med definisjonsmengde $(0,\infty)$. Gammafordelingen, med standardform på PDF, har 2 parametere, α og β. Samspillet mellom dem avgjør lokasjon og dispersjon. PDF-kurven er moderat skjev og moderat tunghalet. Imidlertid vil graden av skjevhet og kurtose avhenge av størrelsen på parameter α. Den vanlige måten å parameterisere gammafordelingens PDF på er ved uttrykket

$$f(y;\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}y^{\alpha-1}e^{-y/\beta}.$$

Under denne parameteriseringen er forventning og varians gitt ved henholdsvis $\mathrm{E}(Y)=\alpha\beta$ og $\mathrm{Var}(Y)=\alpha\beta^2$. Når gammafordelingen brukes som responsfordeling i modelleringssammenheng er det gunstig å ha en egen lokasjonsparameter som representerer forventningsverdien til Y. Johnson et al. (1994) foreslår derfor å reparameterisere ved å sette $\mu=\alpha\beta$ og $\nu^2=\frac{1}{\alpha}$. Det gir følgende PDF:

$$f(y;\mu,\nu)=\frac{y^{\frac{1}{\nu^2}-1}e^{-\frac{y}{\nu^2\mu}}}{(\nu^2\mu)^{1/\nu^2}\Gamma(1/\nu^2)}.$$

PDF kan nå skrive om til samme form som (2), hvilket gir

$$f(y;\mu,\nu)=\exp\left(-\frac{1}{\nu^2}\left(\frac{y}{\mu}+\log\mu\right)+\left(\frac{1}{\nu^2}-1\right)\log y-\log\Gamma\left(\frac{1}{\nu^2}\right)-\frac{2}{\nu^2}\log\nu\right)=\exp\left(c(y,\phi)+\frac{y\theta-a(\theta)}{\phi}\right)$$

der $c(y,\phi)=\left(\frac{1}{\nu^2}-1\right)\log y-\log\Gamma\left(\frac{1}{\nu^2}\right)-\frac{2}{\nu^2}\log\nu$, $\phi=\nu^2$, $\theta=-\frac{1}{\mu}$, samt $a(\theta)=-\ln(-\theta)$.

Følgelig tilhører gammafordelingen den eksponentielle familie, og kan modelleres under GLM-rammeverket, så lenge kun μ ønskes å avhenge direkte av forklaringsvariabler.

Forventning og varians er $\mathrm{E}(Y) = \mu$ og $\mathrm{Var}(Y) = \nu^2\mu^2$. Dersom Y er gammafordelt med parametere μ og ν, vil jeg i det følgende kun skrive $Y \sim \Gamma(\mu, \nu)$ for å indikere dette.

3.5.3 Kjikvadratfordelingen

Et viktig spesialtilfelle av gammafordelingen er kjikvadratfordelingen. Dersom Y har fordeling $\Gamma\left(\rho, \frac{1}{\sqrt{\rho}}\right)$ i $(\mu, \nu)-$parameteriseringen av gammafordelingen, regnes Y som kjikvadratfordelt med ρ frihetsgrader. Tilsvarende parametere i $(\alpha, \beta)-$parameteriseringen er $\alpha = \frac{\rho}{2}$ og $\beta = 2$. Definisjonsmengden er $(0, \infty)$. ρ er alltid positiv, og vanligvis ett heltall. PDF for kjikvadratfordelingen er

$$f(y; \rho) = \frac{1}{2^{\rho/2} \cdot \Gamma(\rho/2)} y^{\frac{\rho}{2}-1} e^{-\frac{y}{2}}.$$

Forventning og varians er henholdsvis $\mathrm{E}(Y) = \rho$ og $\mathrm{Var}(Y) = 2\rho$. Kjikvadratfordelingen vil ikke brukes i selve modelleringen, men er et nyttig verktøy i enkelte tester, som for eksempel Pearsons kjikvadrattest (se delkapittel 3.8). Dersom Y er kjikvadratfordelt med ρ frihetsgrader vil jeg i det følgende kun skrive $Y \sim \chi^2(\rho)$ for å indikere dette. En svært viktig kobling mellom kjikvadratfordelingen og normalfordelingen er at dersom Z er standardnormalfordelt, gjelder $Z^2 \sim \chi^2(1)$. I denne relasjonen ligger mye av årsaken til at kjikvadratfordelingen er såpass mye brukt i statistikkfaget. For et bevis av denne relasjonen, se Casella og Berger (2002).

3.5.4 Lognormalfordelingen

Dersom $\log(Y)$ er normalfordelt med forventning μ og varians σ^2, er Y *lognormalfordelt* med parametere μ og σ. Lognormalfordelingen er kontinuerlig, og har definisjonsmengde $(0, \infty)$. Fordelingen lar seg ikke skrive på formen (2), og kan følgelig ikke brukes som responsfordeling i GLM-rammeverket, men kan modelleres som GAMLSS. PDF er gitt ved

$$f(y; \mu, \sigma) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu))^2}{2\sigma^2}\right)$$

Parameteren μ regnes som lokasjonsparameter, men er med å bestemme både forventning og varians. Parameteren σ er med å bestemme både forventning, varians, skjevhet og kurtose, og avgjør dermed i stor grad formen på PDF-kurven. Forventning og varians er gitt ved $\mathrm{E}(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ og $\mathrm{Var}(Y) = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2)$. Lognormalfordelingen regnes som en moderat skjev og moderat tunghalet fordeling. Dersom Y er lognormalfordelt med parametere μ og σ, vil jeg i det følgende kun skrive $Y \sim \log N(\mu, \sigma)$ for å indikere dette.

3.5.5 Invers Gaussisk fordeling – IG-fordelingen

IG-fordelingen er en meget skjev sannsynlighetsfordeling, med bratt topp. Den er ofte velegnet til å modellere skadepris (se for eksempel de Jong og Heller 2008: 29-30, 125-127). Det er en kontinuerlig fordeling med definisjonsmengde $(0, \infty)$. Jeg bruker her en variant av parameteriseringen til Johnson et al. (1994), og skriver PDF som

$$f(y; \mu, \nu) = \frac{1}{\sqrt{2\pi\nu^2 y^3}} \exp\left(-\frac{(y-\mu)^2}{2\mu^2\nu^2 y}\right).$$

Denne funksjonen kan skrives om til formen fra (2). Det gir PDF

$$f(y; \mu, \nu) = \exp\left(-\frac{y}{2\mu^2\nu^2} + \frac{1}{\mu\nu^2} - \frac{1}{2}\log(2\pi y^3\nu^2)\right) = \exp\left(c(y, \phi) + \frac{y\theta - a(\theta)}{\phi}\right)$$

der $c(y, \phi) = -\frac{1}{2}\log(2\pi y^3\nu^2)$, $\phi = \nu^2$, $\theta = -\frac{1}{2\mu^2}$ og $a(\theta) = \sqrt{-2\theta}$. Dette demonstrerer at fordelingen tilhører den eksponentielle familie, og at den dermed kan modelleres under GLM-rammeverket. Lokasjonsparameteren μ påvirker også varians, skjevhet og kurtose, mens parameteren ν påvirker varians, skjevhet og kurtose. Forventning og varians er gitt ved $\mathrm{E}(Y) = \mu$ og $\mathrm{Var}(Y) = \mu^3\nu^2$. Dersom Y er en IG-fordelt stokastisk variabel med parametere μ og ν, vil jeg i det følgende kun skrive $Y \sim \mathrm{IG}(\mu, \nu)$ for å indikere dette. En viktig egenskap ved IG-fordelingen er at dersom $Y \sim \mathrm{IG}(\mu, \nu)$, vil skaleringen aY, der a er en konstant, ha fordeling $aY \sim \mathrm{IG}(a\mu, \nu / a)$ (Heller et al. 2006:4).

3.5.6 Weibullfordelingen

Weibullfordelingen er en kontinuerlig sannsynlighetsfordeling med definisjonsmengde $[0,\infty)$. Den er fleksibel, i den forstand at formen på PDF-kurven er svært ulik for ulike verdier av parameterne λ og κ. Jeg velger PDF på formen

$$f(y;\lambda,\kappa)=\frac{\kappa y^{\kappa-1}}{\lambda^{\kappa}}\exp\left(-\left(\frac{y}{\lambda}\right)^{\kappa}\right)$$

der λ er en lokasjons/skalerings-parameter, som også har innvirkning på forventningsverdien, mens κ har størst betydning for formen på PDF-kurven. Weibullfordelingen er en del av den eksponentielle familie og kan passe inn under GLM-rammeverket, men i denne oppgaven modelleres Weibullmodeller ved hjelp av GAMLSS-metodikken. (For reparameteriseringer av Weibullfordelingen, se Johnson et al. (1994)). Forventningen til Weibullfordelingen er gitt ved $\mathrm{E}(Y)=\lambda\Gamma\left(\frac{1}{\kappa}+1\right)$, mens variansen er gitt ved $\mathrm{Var}(Y)=\lambda^2\left(\Gamma\left(\frac{2}{\kappa}+1\right)\right)-(\mathrm{E}(Y))^2$. Dersom Y er Weibullfordelt med parametere λ og κ vil jeg i det følgende kun skrive $Y\sim\mathrm{WEI}(\lambda,\kappa)$ for å indikere dette.

3.5.7 Bernoullifordelingen og binomialfordelingen

Dersom en stokastisk variabel, Y, kan ta 2 mulige verdier (kall dem 0 og 1), og sannsynligheten for at $Y=1$ er ψ, kalles Y Bernoullifordelt. Dette er en elementær diskret sannsynlighetsfordeling, med PMF

$$f(y;\psi)=\begin{cases}1-\psi & \text{dersom } y=0\\ \psi & \text{dersom } y=1\end{cases}$$

Forventning og varians gitt ved $\mathrm{E}(Y)=\psi$ og $\mathrm{Var}(Y)=\psi(1-\psi)$. Dersom Y er Bernoullifordelt med sannsynlighetsparameter ψ skrives det direkte som $Y\sim\mathrm{BER}(\psi)$. Gitt stokastiske variabler $Y_1,...,Y_n$, der alle Y_i har identiske fordelinger, $Y_i\sim\mathrm{BER}(\psi)$, defineres summen $X=\sum_{i=1}^{n}Y_i$ som *binomialfordelt*. Man sier at X representerer summen av n uavhengige *Bernoulli-forsøk*. Binomialfordelingen har følgelig definisjonsmengde (0,1,...,n). Det kan vises (se for eksempel Hogg og Tanis 2010:79) at PMF for binomialfordelingen er

$$f(x;n,\psi)=\binom{n}{x}\psi^{x}(1-\psi)^{n-x}.$$

Forventning og varians er gitt ved $\mathrm{E}(X)=n\psi$ og $\mathrm{Var}(X)=n\psi(1-\psi)$. Dersom X er binomialfordelt med antallsparameter n og sannsynlighetsparameter ψ, vil jeg i det følgende kun skrive $X\sim\mathrm{BIN}(n,\psi)$ for å indikere dette.

3.5.8 Poissonfordelingen

Poissonfordelingen, introdusert av *Simeon Denis Poisson* i 1837[14], er den klassiske sannsynlighetsfordelingen som beskriver "telle-data". For denne oppgavens del er Poissonfordelingen et opplagt valg som responsfordeling for antall skader på en polise, *A*. Poissonfordelingen er av den diskrete type, har definisjonsmengde 0,1,2,..., og har kun 1 parameter, μ, som både bestemmer lokasjon og formen på PMF. PMF skrives på den tradisjonelle måten

$$f(y;\mu)=\frac{\mu^{y}}{y!}e^{-\mu}.$$

Dette uttrykket kan enkelt skrives om til formen (2). Det gir

$$f(y;\mu)=\exp\left(-\log(y!)+y\log\mu-\mu\right)=\exp\left(c(y,\phi)+\frac{y\theta-a(\theta)}{\phi}\right)$$

der $c(y,\phi)=-\log(y!)$, $\phi=1$, $\theta=\log(\mu)$ og $a(\theta)=e^{\theta}$. Følgelig passer Poissonfordelingen inn i GLM-rammeverket. Forventning og varians er identisk gitt ved $\mathrm{E}(Y)=\mathrm{Var}(Y)=\mu$ for denne fordelingen, noe som gjør den lite fleksibel i forhold til mange andre fordelinger. Dersom *Y* er Poissonfordelt med parameter μ vil jeg i det følgende kun skrive $Y\sim\mathrm{PO}(\mu)$ for å indikere dette. En velkjent, og viktig, egenskap ved Poissonfordelingen er at dersom to stokastiske variabler X og Y er uavhengige, og fordelt henholdsvis $X\sim\mathrm{PO}(\mu_X)$ og $Y\sim\mathrm{PO}(\mu_Y)$ vil summen $Z=X+Y$ ha fordeling $Z\sim\mathrm{PO}(\mu_X+\mu_Y)$.

[14] Se artikkel om Simeon Denis Poisson på http://www.britannica.com/EBchecked/topic/466561/Simeon-Denis-Poisson

3.5.9 Negativ binomisk fordeling – NEGBIN-fordelingen

Negativ binomisk fordeling er nært knyttet til Poissonfordelingen[15], men tillater overdispersjon (at variansen er større enn forventningen). Det er en diskret fordeling, med definisjonsmengde 0,1,2,... Jeg bruker en parameterisering som brukes av blant annet de Jong og Heller (2008). I denne formen er PMF gitt ved

$$f(y;\mu,\kappa)=\frac{\Gamma(y+1/\kappa)}{y!\Gamma(1/\kappa)}\left(\frac{1}{1+\kappa\mu}\right)^{1/\kappa}\left(\frac{\kappa\mu}{1+\kappa\mu}\right)^{y}.$$

Omskriving til samme form som (2) gir

$$f(y)=\exp\left(y\log\left(\frac{\mu}{1+\kappa\mu}\right)-\frac{1}{\kappa}\log(1-\kappa\mu)+\log\frac{\kappa^{y}\Gamma(1/\kappa+y)}{y!\Gamma(1/\kappa)}\right)=\exp\left(c(y_i,\phi)+\frac{y_i\theta_i-a(\theta_i)}{\phi}\right)$$

der $\phi=1,\quad \theta=\log\left(\frac{\mu}{1+\kappa\mu}\right)$ og $a(\theta)=-\frac{1}{\kappa}\log(1-\kappa e^{\theta})$. Dette demonstrerer at NEGBIN-fordelingen kan modelleres under GLM-rammeverket. Forventning og varians er $\mathrm{E}(Y)=\mu$ og $\mathrm{Var}(Y)=\mu(1+\kappa\mu)$. Fordelen med denne parameteriseringen er at forventningen kan representeres ved en enkelt lokasjonsparameter, μ, samt at overdispersjon styres av en egen "overdispersjonsparameter" κ. Dersom $\kappa\to 0$, gir det $\mathrm{Var}(Y)\to\mathrm{E}(Y)$. Det kan også vises at fordelingens PMF konvergerer mot Poissonfordelingens PMF når $\kappa\to 0$. Dersom Y er NEGBIN-fordelt, med parametere μ og κ vil jeg i det følgende kun skrive $Y\sim\mathrm{NB}(\mu,\kappa)$ for å indikere dette.

3.6 Finite Mixture – FM

3.6.1 FM-fordelinger

Dersom Y følger en FM-fordeling har den sannsynlighet ψ_i til å følge fordeling f_i, der $i=1,2,...,m.$[16] Det er essensielt at $\sum_{i=1}^{m}\psi_i=1$ for å sikre at Y får en ekte

[15] Det kan vises at i en Poissonprosess $Y(t)$ med parameter λt, dersom frekvensparameteren λ velges stokastisk og fordelt $\Lambda\sim\Gamma(\alpha,\beta)$, får Poissonprosessen fordelingen $Y(t)\sim\mathrm{NB}\left(\alpha,\frac{t}{\beta+t}\right)$.

[16] Disse fordelingene kalles "Finite mixtures" ettersom *m* er et endelig tall.

sannsynlighetsfordeling. En generell PDF/PMF for en FM-fordelt stokastisk variabel *Y* kan skrives (subskript *M* står for mikstur):

$$f_M(y;\boldsymbol{\psi},\boldsymbol{\theta}) = \sum_{i=1}^{m} \psi_i f_i(y;\boldsymbol{\theta}_i),$$

der parametervektorne $\boldsymbol{\psi}$ og $\boldsymbol{\theta}$ inneholder henholdsvis alle sannsynlighetene, og alle fordelingsparameterne, tilhørende alle subfordelingene. Videre representerer f_i og $\boldsymbol{\theta}_i$ henholdsvis PDF/PMF og parametervektor, tilhørende subfordeling *i*. Jeg vil i denne oppgaven begrense meg til å se på FM-varianter der $m = 2$. Sannsynlighetsfordelingene fra delkapittel 3.5 kan alle i prinsippet regnes som FM-fordelinger der $m = 1$. Fordelingene har varierende grad av fleksibilitet, men alle er unimodale – PDF/PMF-kurven har kun 1 topp. Når det utvides til $m = 2$ kan det gi bimodale sannsynlighetsfordelinger. Det gir en langt større fleksibilitet i modelleringen.

Et spesialtilfelle av FM-fordelinger har ekstra stor relevans i forsikring; nemlig der $f_i(y) = 1$ når $y = 0$ og $f_i(y) = 0$ ellers. Slike FM-fordelinger, der den ene subfordelingen har 100 % av sannsynligheten konsentrert på 0, kalles *nulljusterte* eller *zero adjusted/inflated*. Relevansen til slike FM-varianter er stor innen forsikring ettersom de fleste poliser har 0 skader. Noen utvalgte FM-fordelinger av denne typen er fullt implementert i GAMLSS-pakken i R, og kan således modelleres under det vanlige GAMLSS-rammeverket. Jeg vil se på 3 av disse: *Zero Inflated Poisson* – ZIP, *Zero Adjusted Gamma* – ZAGA og *Zero Adjusted Inverse Gaussian* – ZAIG.

3.6.2 ZIP-fordeling

Dersom en stokastisk variabel *Y* følger en ZIP-fordeling, er PMF, slik den parameteriseres av Lambert (1992) gitt ved

$$f(y;\psi,\mu) = \begin{cases} \psi + (1-\psi)e^{-\mu} & \text{dersom } y = 0 \\ (1-\psi)\dfrac{\mu^y}{y!}e^{-\mu} & \text{dersom } y = 1,2,... \end{cases}$$

Denne sannsynlighetsfordelingen har forventning $\mathrm{E}(Y) = \mu(1-\psi)$ og varians $\mathrm{Var}(Y) = \mu(1-\psi)(1+\mu\psi)$. Fordelingen er av typen FM der $f_1(y) = 1$ når $y = 0$ og $f_1 = 0$

ellers og $f_2(y;\mu)$ er PMF for den Poissonfordeling. Fordelingen har høyere konsentrasjon av sannsynlighet ved 0 enn vanlig Poissonfordeling, og har også overdispersjon, hvilket kan ses ved at $\text{Var}(Y) > \text{E}(Y)$. ZIP-fordelingen er mer fleksibel enn Poissonfordelingen, men det er ikke slik at summen av uavhengige ZIP-fordelte variabler også er ZIP-fordelt. Dersom Y følger en ZIP-fordeling med parametere ψ og μ vil jeg i det følgende kun skrive $Y \sim \text{ZIP}(\psi, \mu)$ for å indikere dette.

3.6.3 ZAGA-fordeling

ZAGA-fordelingen er dels kontinuerlig og dels diskret. Den er av typen FM, og $f_1(y) = 1$ når $y = 0$ og $f_1(y) = 0$ ellers. Den andre delfordelingen f_2, er en gammafordeling med parametere μ og ν. PMF/PDF for ZAGA-fordelingen er

$$f(y;\psi,\mu,\nu) = \begin{cases} \psi & \text{dersom } y = 0 \\ (1-\psi)\dfrac{y^{\frac{1}{\nu^2}-1} e^{-\frac{y}{\nu^2\mu}}}{(\nu^2\mu)^{1/\nu^2}\Gamma(1/\nu^2)} & \text{dersom } y > 0 \end{cases}$$

Forventning og varians er gitt ved $\text{E}(Y) = (1-\psi)\mu$ og $\text{Var}(Y) = (1-\psi)\mu^2(\psi+\nu^2)$. ZAGA-fordelingen er fullt implementert i GAMLSS-pakken i R, og kan modelleres ved hjelp av denne. Alle fordelingsparametere kan avhenge direkte av forklaringsvariablene og estimeringen kan gjøres ved RS- eller CG-algoritmen. Dersom Y følger en ZAGA-fordeling med parametere ψ, μ og ν vil jeg i det følgende kun skrive $Y \sim \text{ZAGA}(\psi, \mu, \nu)$ for å indikere dette.

3.6.4 ZAIG-fordeling

ZAIG-fordelingen er en dels kontinuerlig, dels diskret fordeling. Den er av typen FM der $f_1(y)=1$ når $y=0$ og $f_1(y)=0$ ellers. $f_2(y;\mu,\nu)$ er PDF for en IG-fordeling med parametere μ og ν. Bortoluzzo et al. (2011) bruker ZAIG-fordelingen til å predikere skadepris i bilforsikring, hvilket jeg også vil gjøre i denne oppgaven. Jeg bruker her en variant av deres parameterisering og får følgende PDF/PMF

$$f(y;\psi,\mu,\nu)=\begin{cases}\psi & \text{dersom } y=0\\ (1-\psi)\dfrac{1}{\sqrt{2\pi\nu^2y^3}}\exp\left(-\dfrac{(y-\mu)^2}{2\mu^2\nu^2y}\right) & \text{dersom } y>0\end{cases}$$

Forventning og varians er gitt ved $\mathrm{E}(Y)=(1-\psi)\mu$ og $\mathrm{Var}(Y)=(1-\psi)\mu^2(\psi+\mu\nu^2)$. ZAIG-fordelingen er implementert i GAMLSS-pakken i R, og kan modelleres ved hjelp av denne, slik at alle fordelingsparameterne (inklusiv nullsannsynligheten ψ) kan avhenge direkte av forklaringsvariabler. Dersom *Y* følger en ZAIG-fordeling med parametere ψ, μ og ν vil jeg i det følgende kun skrive $Y \sim \mathrm{ZAIG}(\psi,\mu,\nu)$ for å indikere dette.

3.6.5 Estimering av FM-modeller - EM-algoritmen

Noen få spesielle FM-fordelinger, herunder ZIP, ZAGA og ZAIG, er fullt implementert i GAMLSS-pakken i R som selvstendige GAMLSS-fordelinger. Estimering av disse gjøres ved RS-algoritmen og CG-algoritmen (se delkapittel 3.4.3). Jeg ønsker imidlertid å bygge også andre FM-modeller der ingen av subfordelingene nødvendigvis er konsentrert på 0. Et generelt problem når man skal maksimere likelihooden i slike FM-modeller er at det antas at hver observasjon følger en av subfordelingene. Hvilken av de to subfordelingene hver enkelt observasjon følger kan imidlertid ikke fastslås. Løsningen er å anta en stokastisk vektor **v**, med 1 element for hver observasjon. **v** er en binær vektor, som inneholder koder som avgjør, for hver observasjon, hvilken av de 2 subfordelinger den tilhører. Ved å behandle likelihooden som avhengig av den stokastiske vektoren **v** blir selve likelihooden en stokastisk variabel. Forventningsverdien til den stokastiske likelihooden kan da maksimeres. Dette er ideen bak *Expectation Maximization algorithm* – EM-algoritmen. EM-algoritmen forklares og utdypes i detalj av Gupta og Chen (2010).

3.7 Sentralgrenseteoremet

Det klassiske sentralgrenseteoremet sier at dersom $Y_1,...,Y_n$ er i.i.d. stokastiske variabler, alle med forventningsverdi μ og varians σ^2, vil størrelsen $Z = \frac{\bar{Y} - \mu}{\sqrt{n}\sigma}$, der $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, konvergere i fordeling mot standardnormalfordelingen $N(0,1)$. I denne oppgaven brukes en mer generell variant av sentralgrenseteoremet. La $Y_1,...,Y_n$ være uavhengige stokastiske variabler med hver sine forventninger $\mathrm{E}(Y_i) = \mu_i$ og varianser $\mathrm{Var}(Y_i) = \sigma_i^2$, der ingen av forventningene eller variansene divergerer mot ∞. Definer varianssummen $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$. Under gitte regularitetsforutsetninger (se Le Cam 1986:80) gjelder

$$\frac{1}{s_n}\sum_{i=1}^{n}(Y_i - \mu_i) \xrightarrow{d} N(0,1).$$

Denne versjonen av sentralgrenseteoremet vil tas i bruk i kapittel 9.

3.8 Pearsons kjikvadrattest

Den varianten av Pearsons kjikvadrattest jeg vil bruke i denne oppgaven (se kapittel 4) er varianten der likheten til *m* ulike sannsynligheter testes. Responsvariabelen *Y* kan her kun ta verdiene 0 og 1, og regnes som Bernoullifordelt. Nullhypotesen i Pearsons test er at verdien til den kategoriske forklaringsvariabelen *X* ikke påvirker fordelingen til *Y*. La *X* ha definisjonsmengde 1,2,...,*m*, og la $P(Y_i = 1 \mid X_i = k) = p_k$. Da kan nullhypotesen skrives

$$H_0 : p_1 = p_2 = ... = p_m = p.$$

Etter at man har gjort *n* forsøk (der *n* deles inn etter verdien på forklaringsvariabelen *X*, slik at for eksempel n_k er antall forsøk der $X = k$), summeres antall *Y*-observasjoner av hver type opp, for hver verdi av forklaringsvariabelen *X*. Dette gir for eksempel

$$V_{1,k} = \sum_{i=1}^{n} Y_i \cdot I(X_i = k),$$

som er antall observasjoner der $Y = 1$ og $X = k$. Disse *V*-verdiene antas å være binomialfordelte, ettersom de er summen av Bernoullifordelte variabler med lik sannsynlighetsfordeling.

Testobservator er *Q*, gitt ved

$$Q = \sum_{k=1}^{m} \frac{(V_{0,k} - n_k(1-\bar{Y}))^2}{n_k(1-\bar{Y})} + \sum_{k=1}^{m} \frac{(V_{1,k} - n_k\bar{Y})^2}{n_k\bar{Y}},$$

der snittverdien $\bar{Y}$ er gitt ved $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Hogg og Tanis (2010:417-424) demonstrerer, ved hjelp av sentralgrenseteoremet og definisjonen av kjikvadratfordelte variabler, at dersom H_0 er sann vil *Q* konvergere mot fordeling $\chi^2(m-1)$ når antall observasjoner i hver gruppe, n_k, øker.

3.9 Prising av forsikringspoliser

Anta at den potensielle utbetalingen fra forsikringsselskapet til kunden er en stokastisk variabel, *U*, og at sannsynlighetsfordelingen til *U* er kjent. Premien, Π, kan deles inn slik:

$$\Pi = \mathrm{E}(U) + R + A + M + F$$

der *R* er risikotillegg, *A* er administrasjonskostnader, *M* er markedsjustering og *F* er ønsket fortjeneste. Av disse er det først og fremst $\mathrm{E}(U)$ og *R* som er i søkelyset i denne oppgaven. Sundt (1999:15-23) drøfter noen mulige prinsipper for å fastsette premien på en forsikring. Disse prinsippene er regler (eller formler) for å bestemme verdien på risikotillegget *R*. 3 av prinsippene Sundt (1999:15-23) nevner er

- Forventningsprinsippet: $R = a\,\mathrm{E}(U)$, der *a* er en konstant.
- Standardavviksprinsippet: $R = b\sqrt{\mathrm{Var}(U)}$, der *b* er en konstant.
- Variansprinsippet: $R = c\,\mathrm{Var}(U)$, der *c* er en konstant.

Hvilket av disse prinsippene som er det optimale for god prissetting er ikke entydig besvart. I litteraturen foreslås også andre prinsipper som *eksponentialprinsippet* og *kvantilprinsippet*. For videre drøfting av disse og andre prinsipper, se Young (2004). For denne oppgavens del velger jeg å bruke standardavviksprinsippet til prising (se kapittel 10). Standardavviket er en relativt intuitiv størrelse som sier noe om usikkerheten til forventningsverdien $\mathrm{E}(U)$. Det er naturlig at større standardavvik gir høyere risiko, og dermed høyere risikotillegg.

4 Data

Data for denne oppgaven er hentet fra et skadeforsikringsselskap. Datagrunnlaget er opplysninger om kunder og skader på bilforsikring, spesifikt kaskodekning der glasskader ikke er tatt med. Det er anonymiserte data bestående av en polisetabell og en skadetabell.

4.1 Polisetabellen

Denne tabellen ligger på kunde-bil-årstall-nivå. Det betyr at en rad i tabellen tilsvarer en unik kombinasjon av kunde, bil og årstall. I praksis vil en typisk polise ha et års løpetid og gå over 2 årstall, for eksempel fra 20.3.2008 til 19.3.2009. Det er også vanlig å registrere flere dekninger og eventuelt også flere biler på samme polise, for eksempel kaskodekning og ansvarsdekning på bil 1 og kun ansvarsdekning på bil 2. For enkelhets skyld innfører jeg følgende konvensjon: en *polise* er en unik kombinasjon av kunde, bil og årstall. Slik definerer jeg hver rad i polisetabellen som en polise. Kolonner i polisetabellen er polise-id, kunde-id, årstall, bilalder, personalder, antall aktive dager, antall skader, total skadepris og gjennomsnittlig skadepris. Data fra denne tabellen danner grunnlaget for modellering av skadefrekvens, gjennomsnittlig skadepris gitt skade og total skadepris. Tabellen har 63 165 rader med observasjoner.

polise_id	kunde_id	aar	bilalder	personalder	aktive_dager	skader	pris_tot	pris_snitt
10	12	2000	3	34	185	0		
25	16	2003	1	37	180	0		
390	190	2004	2	38	185	0		
111	159	2005	3	55	365	1	3 570	3 570
123	198	2000	3	40	365	0		

Tabell 4.1 - Utdrag fra polisetabellen

4.2 Skadetabellen

Denne tabellen ligger på skade-nivå. Det betyr at en rad i tabellen tilsvarer en unik skade. Skadetabellen inneholder kunde-id og polise-id for hver enkelt skade. Disse tallene kan brukes som nøkler for å koble de to tabellene sammen. Kolonner i skadetabellen er polise-id, kunde-id, skadedato, bilalder, personalder og skadepris. Data fra denne tabellen danner grunnlaget for modellering av skadepris per skade, *S*. Tabellen har 9 396 rader med observasjoner. Det betyr at det er 9 396 skader i dette datasettet.

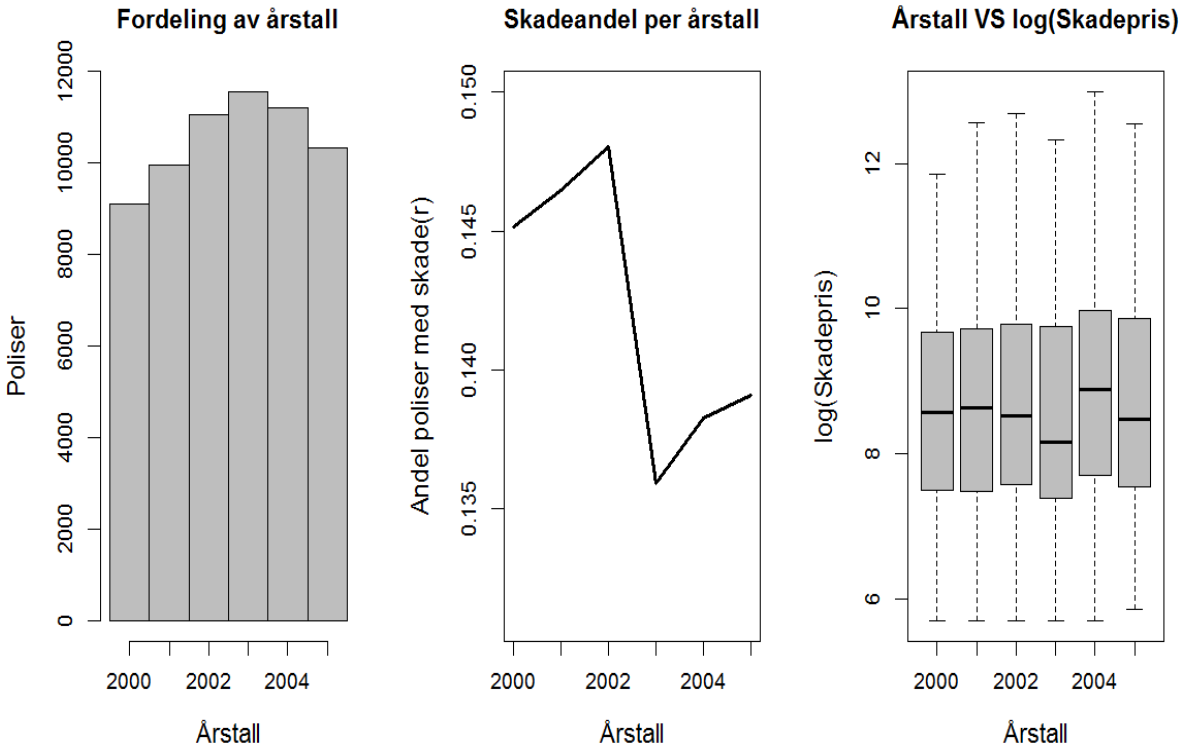
polise_id	kunde_id	skade_id	skadedato	bilalder	personalder	pris
11	13	42150	17.07.2005	3	55	3 570
20	26	58256	02.07.2001	5	50	8 970
27	28	85786	15.09.2000	1	43	11 110
28	28	35578	29.03.2001	2	44	655
30	86	102235	06.11.2003	0	32	14 368

Tabell 4.2 - Utdrag fra skadetabellen

4. 3 Forklaringsvariabler - hypoteser og deskriptiv statistikk

4.3.1 Årstall

Hver polise og hver skade er registrert med årstall. Variabelen strekker seg fra 2000 til 2005 og har naturlig nok kun registreringer som heltall. Det fremgår av figur 4.1 at antall poliser øker frem til 2003 for så å ta noe av fra 2004.



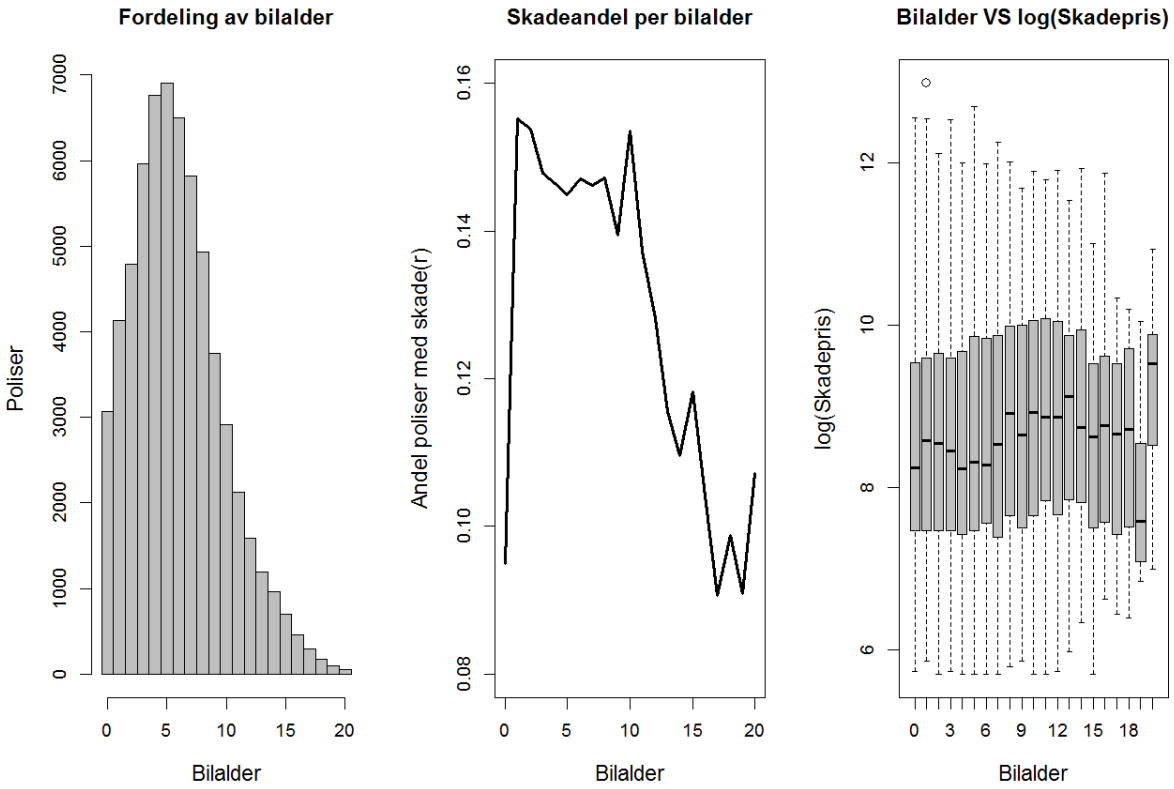
Figur 4.1 - Deskriptiv statistikk for årstall

Det er relevant å se på hvordan skadesannsynligheten utvikler seg over tid. Grafen viser at andelen poliser uten skader er ca. 1 prosentpoeng høyere de tre siste årene enn de tre første årene. Pearsons kjikvadrattest (se delkapittel 3.8) der nullhypotesen er at årstallet ikke har noen betydning for skadesannsynligheten gir $\chi^2 = 11{,}147$ ved 5 frihetsgrader. Dette gir *p*-verdi på 0,048. Denne enkle marginalanalysen gir statistisk grunnlag for å hevde at skadesannsynligheten forandres med årene. Dette styrker troen på at årstall bør være med som forklaringsvariabel for skadefrekvens i den multivariable analysen. Box-plottet viser årstall mot log(skadepris). Det viser relativ stabilitet i perioden 2000 til 2002 og mer variasjon i peridoen 2003 til 2005. En naturlig hypotese er at skadepris vil stige med årene på grunn av inflasjon. Box-plottet gir ikke noe entydig svar på hvorvidt dette stemmer.

4.3.2 Bilalder

Hver enkelt skade og hver enkelt polise er registrert med bilalder. Variabelen strekker seg fra 0 til 20, og er registrert som heltall. En hypotese er at gamle biler er i dårligere stand, og derfor har høyere skadefrekvens enn nye biler. En annen hypotese er at nye biler har nye og

dyre bildeler som gir høyere skadepris enn for gamle biler. I teorien er bilalder en kontinuerlig størrelse, men ettersom den er registrert som et heltall kan den også betraktes som en kategorisk eller ordinal størrelse.

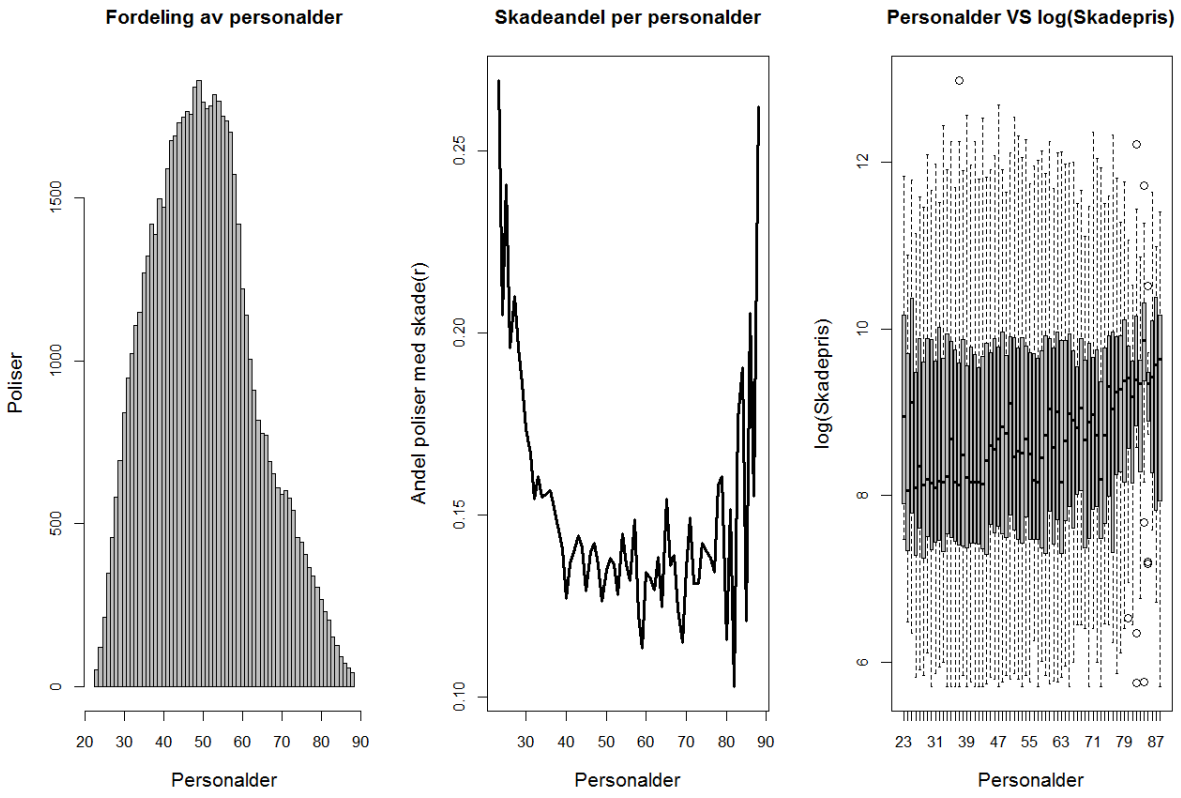


Figur 4.2 - Deskriptiv statistikk for bilalder

Fordelingen av bilalder i polisetabellen er til forveksling lik gammafordelingen. Det ser ut til at skadeandelen forandrer seg relativt mye med bilalder. Den klareste frekvenstrenden er at skadeandelen går ned fra bilen er 11 år, og holder seg så på et lavere nivå enn for nyere biler. Pearsons kjikvadrattest for bilalder mot skadesannsynlighet gir $\chi^2 = 115{,}3$ på 20 frihetsgrader. Dette gir *p*-verdi 0 og følgelig signifikans på alle signifikansnivåer. Det gir grunn til å tro at biladler har betydning for skadesannsynligheten. Skadeprisen har et noe uklart mønster når bilalder blir større. Den klareste trenden i forhold til skadepris er at variasjonen blir mindre når bilalder øker.

4.3.3 Personalder

Hver skade og hver polise er registrert med personalder. Dette er kundens (bilførers) alder i det aktuelle årstall. Personalderen i datasettet strekker seg fra 23 år til 88 år og er registrert som heltall. Personalder er kontinuerlig av natur men kan også inndeles kategorisk eller ordinalt. En hypotese her er at unge bilførere kjører uforsiktig og derfor har høyere skadefrekvens enn andre aldersgrupper.



Figur 4.3 - Deskriptiv statistikk for personalder

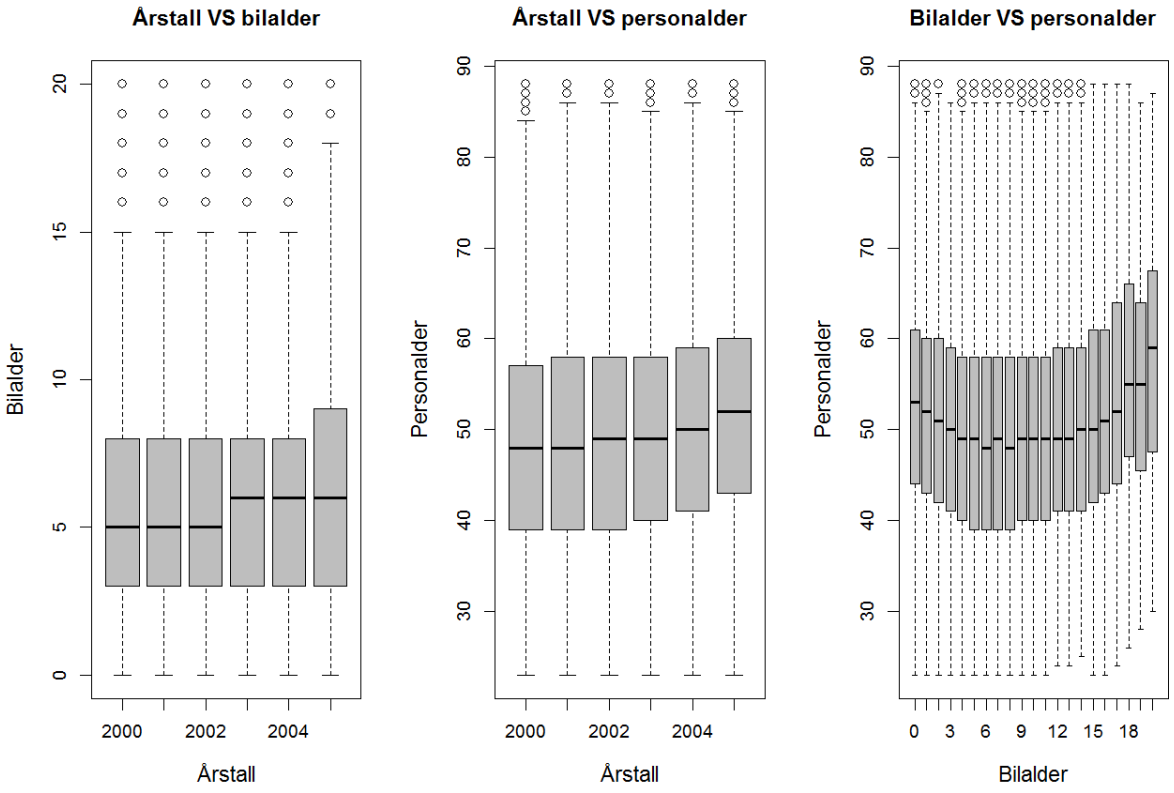
Fordelingen av personaldre ser nesten normalfordelt ut. Som ventet viser figur 4.3 at unge kunder har langt høyere skadeandel enn andre. De eldre skiller seg også ut med høy skadeandel. Pearsons kjikvadrattest for personalder mot skadesannsynlighet gir $\chi^2 = 163,68$ på 65 frihetsgrader. Det gir videre *p*-verdi tilnærmet 0 og svært god signifikans. Det er dermed grunn til å tro at personalder har betydning for skadesannsynligheten. Etter box-plottet å dømme øker skadeprisen med personalder. I delkapitler 6.3, 7.8 og 8.6, foretar jeg en grundigere undersøkelse av forklaringsvariablenes effekter på responsvariablene.

4.3.4 Samvariasjon mellom forklaringsvariablene

Det kan tenkes at noen av forklaringsvariablene er korrelerte. Dersom dette er tilfellet kan det påvirke den multivariable analysen ved å introdusere mer usikkerhet. Pearsons korrelasjonskoeffisient, der årstallene er **a**, bilalderne er **b** og personalderne er **p**, gir følgende høyst signifikante resultater:

$$\text{Corr}(\mathbf{a},\mathbf{b}) = 0,047$$
$$\text{Corr}(\mathbf{a},\mathbf{p}) = 0,080$$
$$\text{Corr}(\mathbf{b},\mathbf{p}) = -0,01$$

Dette betyr at det er en tendens i perioden 2000 til 2005 til at bilforsikringskundene blir noe eldre, og at eldre biler blir forsikret. Det er også en liten tendens til at yngre kunder kjører eldre biler. Dette kan henge sammen med at eldre ofte har bedre økonomi. Figur 4.4 viser box-plot for hver av disse tre mulige samvariasjonene, og man kan se langt mer informasjon enn hva korrelasjonstallene alene gir. Det ser ut som om bilalder har en markant økning fra 2002 til 2003, mens personalder øker noenlunde jevnt med årene. Det ser også ut til at fordelingen av personalder per bilalder er parabelformet.



Figur 4.4 - Box-plot av samvariasjon mellom forklaringsvariablene

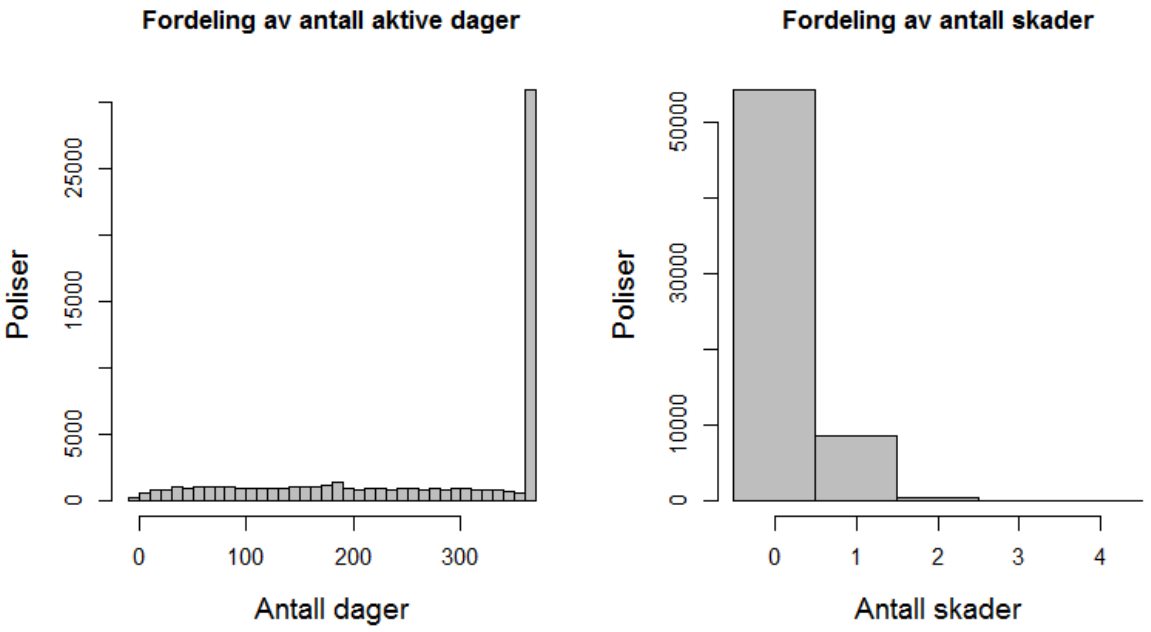
4.4 Responsvariabler – hypoteser og deskriptiv statistikk

4.4.1 Antall skader og antall aktive dager

For hver polise er det registrert et skadeantall. For de aller fleste poliser er antall skader 0. Høyeste antall skader for en polise i datasettet er 4. La antall skader for polise *i* være A_i. Enkel deskriptiv statistikk for antall skader er som følger:

- Gjennomsnitt: $\bar{a} = 0{,}1488$
- Median: $m(\mathbf{a}) = 0$
- Estimert varians: $s^2(\mathbf{a}) = 0{,}1408764$

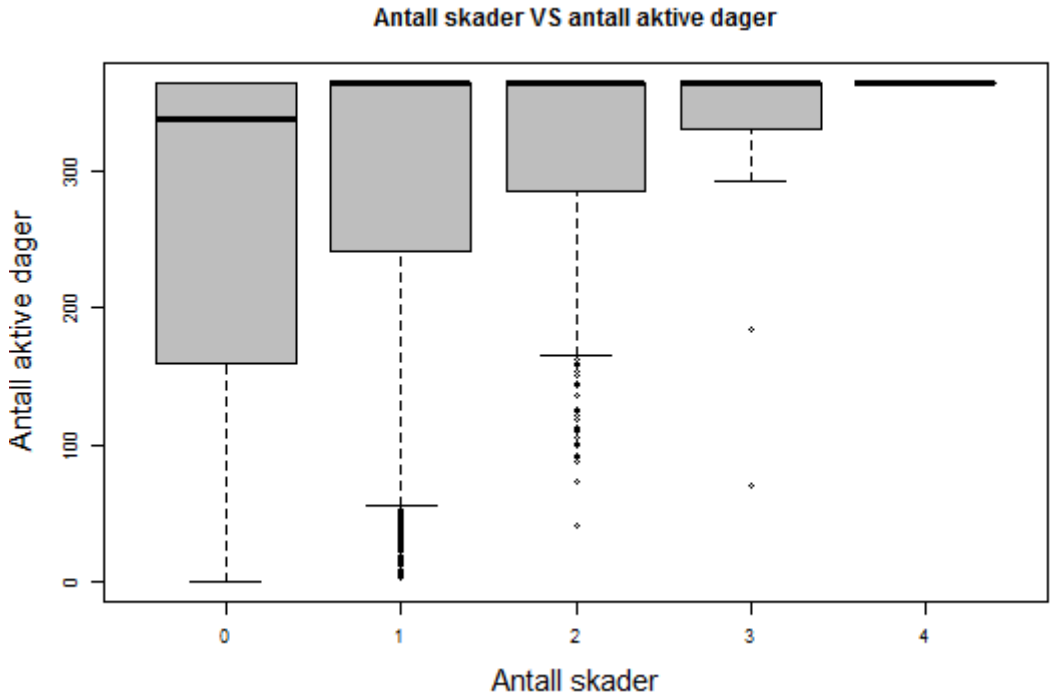
Gjennomsnitt og estimert varians er ikke langt fra hverandre. Dette stemmer godt overens med en Poissonmodell for antall skader. Disse tallene er imidlertid ikke justert for antall aktive dager til hver enkelt polise. Antall aktive dager opptrer som en naturlig eksponeringsvariabel for polisene. Den må derfor ses i sammenheng med antall skader. Det forventes at disse to størrelser er sterkt samvarierende.



Figur 4.5 - Histogrammer for antall aktive dager og antall skader

Det er ikke overraskende at helårspoliser med 365 aktive dager dominerer. Det samme gjør poliser uten skader. Pearsons korrelasjonskoeffisient der antall aktive dager er **d** og antall

skader **a** er $\text{Corr}(\mathbf{d},\mathbf{a}) = 0{,}107$. Dette tilsier at det er en positiv sammenheng mellom antall aktive dager og antall skader, og at sammenhengen går i forventet retning. Box-plottet viser, ikke overraskende, en klar trend der antall aktive dager i snitt er høyere, jo flere skader polisen har.



Figur 4.6 - Box-plot av antall skader vs. antall aktive dager

4.4.2 Skadepris

Skadeprisen er forsikringsselskapets registrerte utbetaling tilknyttet hver enkelt skade. Minste registrerte skadepris er 300 kr mens den største er 434 273 kr. Den totale skadekostnaden for forsikringsselskapet vil ofte være langt høyere enn disse tallene. Skadepris i dette datasettet gjelder kun for kaskodekningen, eksklusiv glass. Skadepris er i teorien kontinuerlig, men noen tall går igjen i datasettet som registrert skadepris. Dette skyldes i hovedsak to forhold:

- Forsikringsselskapet har avtaler med ulike bilverksteder som har fastpris på gitte reparasjoner.
- Skadepris er fordelt mellom ulike dekninger, der kasko er en blant flere. Dette gjøres manuelt og avrundinger eller standardbeløp er vanlig.

En annen ting som er verdt å merke seg, er at skadeprisen går helt ned i 300 kr. Årsaker til dette kan være at skaden er liten, at kunden har valgt høy egenandel, eller at mesteparten av skadeprisen føres på andre dekninger. La s_i være skadeprisen for skade *i*. Enkel deskriptiv statistikk for skadepris er som følger:

- Gjennomsnitt: $\bar{s} = 15\,091$
- Median: $m(\mathbf{s}) = 5\,110$
- Estimert standardavvik: $\sqrt{s^2(\mathbf{s})} = 25\,034$

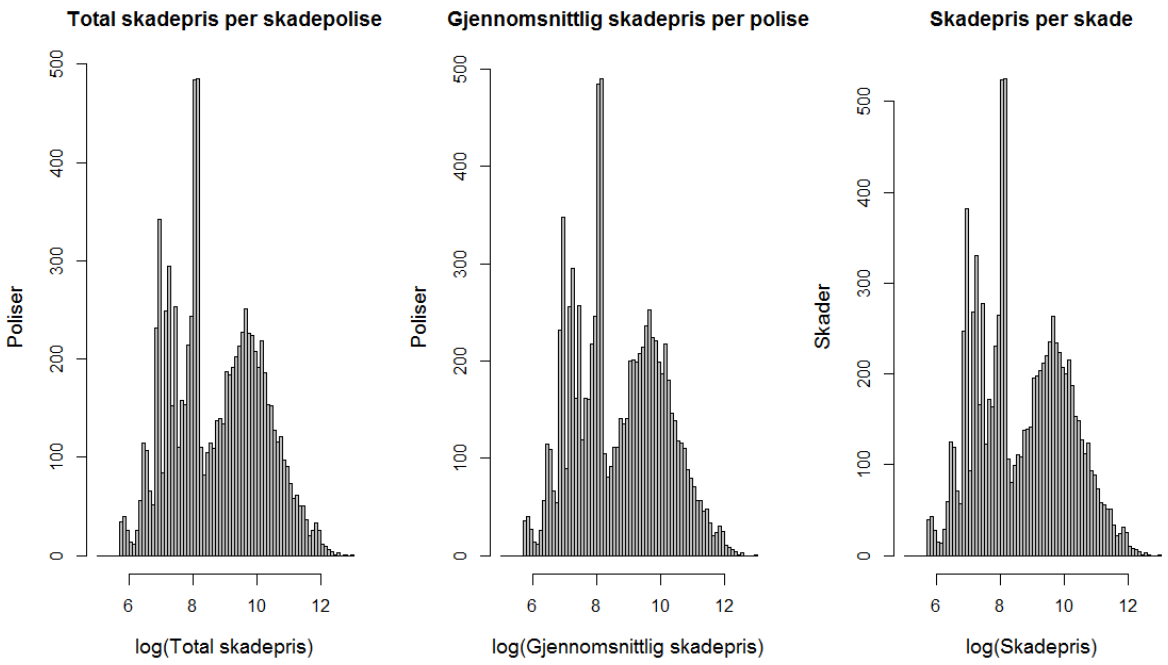
Dette indikerer at de fleste skader har skadepris under gjennomsnittet, men at noen få store skader drar gjennomsnittet opp. Dette tyder videre på at skadeprisen ikke er symmetrisk fordelt.

4.4.3 Aggregering av skadepris

Skadepris, *S*, er naturlig tilknyttet enkeltskader, men det er også meget relevant å se på aggregeringer av skadepris. To relevante aggregeringer er

- *Total skadepris* per skadepolise, U^*.
- *Gjennomsnittlig skadepris* per polise med skade, *G*.

Det kan være modelleringsmessig nyttig å knytte en forventet skadepris til hver polise. Da tjener gjennomsnittlig skadepris per skadepolise, *G*, som observasjoner. Grafisk sett er det mer informativt å se på logaritmen til skadepris enn på ren skadepris.



Figur 4.7 - Histogrammer av log(skadepris) for ulike varianter av skadepris

Det er vanskelig å se noen særlig forskjell på de 3 histogrammene i figur 4.7. Det som er tydelig for alle tre plot er at skadeprisen har minst to topper. Det betyr at mikstur-modeller som tillater flere topper, bimodale FM-modeller, er høyst aktuelle.

5 Metodikk for modellering

5.1 Generelt rammeverk for alle unimodale modeller

De unimodale modellene jeg ser på i denne oppgaven er av typen GAMLSS (se delkapittel 3.4). Hver modell har 1 responsvariabel Y [17]. Observasjonene y_i betraktes som uavhengige realiseringer av stokastiske variabler Y_i, med fordeling $f_i(y_i;\boldsymbol{\theta}_i)$ Et nøkkelpoeng er at parametervektoren $\boldsymbol{\theta}_i$ kan variere med observasjonen *i*. Forklaringsvariablene for observasjon *i* skrives ved hjelp av designvektorene $\mathbf{x}_i^T$, $\mathbf{z}_i^T$ og $\mathbf{w}_i^T$. GAMLSS-rammeverket åpner altså for ulike design for hver av fordelingsparameterne. Generelt benyttes følgende kobling mellom responsfordeling og forklaringsvariabler:

$$
\begin{aligned}
Y_i &\sim f(\theta_{i,1},\theta_{i,2},\theta_{i,3}) \\
g_1(\theta_{i,1}) &= \mathbf{x}_i^T\boldsymbol{\beta} \\
g_2(\theta_{i,2}) &= \mathbf{z}_i^T\boldsymbol{\gamma} \\
g_3(\theta_{i,3}) &= \mathbf{w}_i^T\boldsymbol{\delta}
\end{aligned}
$$

Dette rammeverket gjelder selvsagt for responsfordelinger med 3 fordelingsparametere. Rammeverk for responsfordelinger med 1 eller 2 fordelingsparametere defineres tilsvarende. Det er hensiktsmessig å sentrere, eller referansejustere, alle forklaringsvariabler. Det gjør at konstantleddene β_0, γ_0 og δ_0 får representere underliggende intensitet for link-transformasjonen av fordelingsparameterne $\theta_{i,1}$, $\theta_{i,2}$, og $\theta_{i,3}$. For bilalder og personalder lar jeg gjennomsnittsverdien i porteføljen være referansepunkter, mens for årstall lar jeg 2006 være referansepunkt. La registrerte verdier på polise *i* for årstall, bilalder og personalder være henholdsvis t_i^*, b_i^*, p_i^*. Når disse størrelser behandles som kontinuerlige forklaringsvariabler, velger jeg heller å bruke størrelsene fratrukket referansepunktene. Forklaringsvariablene som faktisk brukes blir da $t_i = t_i^* - 2006$, $b_i = b_i^* - \bar{b}^*$ og $p_i = p_i^* - \bar{p}^*$. Dette betyr at en gjennomsnittlig observasjon vil ha verdier $t_i = b_i = p_i = 0$ for 2006, slik at intensitetene for hver parameter kun er gitt ved konstantleddene ved referansenivåene $g_1^{-1}(\beta_0)$, $g_2^{-1}(\gamma_0)$ og $g_3^{-1}(\delta_0)$. Denne referansejusteringen av forklaringsvariablene er en lineær transformasjon, og vil derfor ikke påvirke parameterne $\beta_j, \gamma_j, \delta_j$, $j \neq 0$. Det går derfor ingen informasjon tapt

[17] Y er betegnelsen på responsvariabelen generelt. Den byttes ut med A for antall skader på en polise, *S* for skadepris per skade, *U* for total skadepris per polise og *G* for gjennomsnittlig skadepris per skadepolise.

ved å trekke fra referansepunkter, og modellene blir lettere å tolke. Når jeg heretter bruker benevnelsene t_i, b_i og p_i, skal de forstås som de referansejusterte størrelsene.

5.2 Generelt rammeverk for alle bimodale FM-modeller

Som det fremgår av figur 4.7, vil det være behov for å lage modeller med responsfordelinger som tillater mer enn en topp. Jeg vil derfor også ta i bruk enkelte bimodale FM-modeller som tillater modellering av 2 ulike topper. Her betraktes observerte verdier, y_i, som realiseringer av de stokastiske variabler Y_i, som antas å ha fordeling $f_{i,M}(\boldsymbol{\theta}_{i,M})$. PDF til fordelingene $f_{i,M}(\boldsymbol{\theta}_{i,M})$ kan generelt skrives

$$f_{i,M}(y_i;\boldsymbol{\theta}_{i,M}) = \psi_i f_{i,1}(y_i;\boldsymbol{\theta}_{i,1}) + (1-\psi_i) f_{i,2}(y_i;\boldsymbol{\theta}_{i,2}).$$

En slik bimodal modell gir mulighet til å modellere to ulike stokastiske prosesser samtidig. En observasjon vil følge fordeling $f_{i,1}(y_i;\boldsymbol{\theta}_{i,1})$ med sannsynlighet ψ_i, og fordeling $f_{i,2}(y_i;\boldsymbol{\theta}_{i,2})$ med sannsynlighet $(1-\psi_i)$. I prinsippet kan alle parameterne i den komplette parametervektoren, $\boldsymbol{\theta}_{i,M}$, avhenge av forklaringsvariabler. I praksis er det mest relevant i å la ψ_i avhenge av forklaringsvariablene. Tankegangen er at kombinasjonen av forklaringsvariabler er avgjørende for hvilken sannsynlighetsfordeling hver observasjon tilhører, mens realiseringen innenfor en av sannsynlighetsfordelingene er stokastisk over en standardfordeling for alle observasjoner.

En annen grunn til at det kan være mest hensiktsmessig å kun la ψ_i avhenge av forklaringsvariabler her, er at det uansett må estimeres mange parametere, hvilket fort kan lede til overparameterisering. Jeg tillater meg likevel å modellere lokasjonsparameterne $\mu_{i,1}$ og $\mu_{i,2}$ (disse inngår i alle aktuelle varianter av $\boldsymbol{\theta}_{i,M}$) som avhengig av årstall t_i. Grunnen til dette illustreres med modellering av skadepriser: La hver enkelt skadepris tilhøre 1 av 2 sannsynlighetsfordelinger (1 for små skader og 1 for store skader). Sannsynligheten for hvorvidt skaden vil tilhøre den ene eller den andre fordelingen er antatt å være betinget av bilalder og personalder. Det er rimelig å tillate de to sannsynlighetsfordelingene å variere i posisjon over tid, ettersom inflasjon sannsynligvis vil forskyve lokasjonsparameteren oppover både for små og store skader. I stedet for å se på dette som modellering ved hjelp av

forklaringsvariabler, ser jeg på dette som korreksjon for årstall. Generell modellformulering blir da

$$
\begin{aligned}
Y_i &\sim f_{i,M}(\boldsymbol{\theta}_{i,M}) \\
g_1(\psi_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \\
g_2(\mu_{i,1}) &= \begin{bmatrix} 1 & t_i \end{bmatrix} \boldsymbol{\gamma} \\
g_3(\mu_{i,2}) &= \begin{bmatrix} 1 & t_i \end{bmatrix} \boldsymbol{\delta}
\end{aligned}
$$

Som for de de unimodale modellene vil det også her kun brukes referansejusterte størrelser t_i, b_i og p_i, samt deres ulike funksjonelle former, som forklaringsvariabler.

5.3 Algoritme for AIC-minimering

Jeg estimerer parametere i en rekke ulike modeller i denne delen av oppgaven. For lettere å holde oversikt, tar jeg i bruk en fast algoritme for AIC-minimering av hver enkelt modell, med hensyn på hvilke forklaringsvariabler som skal tas med, og hvilken funksjonell form de skal ha. Hver enkelt modell er i første rekke karakterisert av hvilken responsfordeling og hvilke link-funksjoner som velges. Etter disse valgene er tatt, søkes et best mulig kompromiss mellom høy grad av tilpasning til data, og lav grad av usikkerhet i estimatene. AIC (se delkapittel 3.1) er mitt foretrukne kriterium for å løse dette dilemmaet. AIC konvergerer mot optimal modell når antall observasjoner går mot ∞. Datasettet er såpass stort at forskjellen mellom AIC og det mer generelle AIC_c (som korrigerer for antall observasjoner) er neglisjerbar. Jeg bruker en stegvis minimerings-algoritme der jeg starter med en grunn-modell og ideelt sett ender opp med en minimert modell i forhold til AIC. For enkelhets skyld definerer jeg "kandidatledd" som *en bestemt funksjonell form av en forklaringsvariabel*. Bilalder som andregradspolynom, $b_i + b_i^2$, er et eksempel på et kandidatledd. AIC-minimerings-algoritmen har følgende steg:

1. Estimer fordelingens parametere uten bruk av forklaringsvariabler. Disse estimater definerer grunnmodellen MOD-1. Regn ut AIC for denne modell.
2. For alle kandidatledd for hovedparameteren (lokasjonsparameteren, μ, for unimodale modeller og sannsynlighetsparameteren, ψ, for bimodale modeller): Regn ut hva AIC vil bli dersom man legger til dette kandidatledd. Ranger så alle kandidatledd etter AIC.
3. Dersom minst 1 av kandidatleddene gir lavere AIC enn grunnmodellen legges dette kandidatleddet til og man har gjeldende modell MOD-2.

4. Repeter steg 2 til 3 helt til det oppnås en optimal kombinasjon av kandidatledd for hovedparameteren.
5. Dersom andre parameterne avhenger av forklaringsvariabler, AIC-minimeres disse steg for steg, på samme måte som for hovedparameteren. Rekkefølgen av parameterne har betydning. Estimer derfor parameterne i prioritert rekkefølge.[18]
6. Når modellen er stegvis AIC-minimert for alle relevante parametere, kan det tenkes at den er overparameterisert. Test derfor hva AIC vil bli ved å fjerne hver enkelt av kandidatleddene.
7. Kandidatleddene som gir AIC-nedgang ved fjerning, rangeres etter potensiell AIC-nedgang. Ta bort kandidatleddet som gir størst AIC-nedgang ved fjerning (dersom noen kandidatledd gir AIC-nedgang).
8. Steg 6 og 7 repeteres helt til ingen kandidatledd gir AIC-nedgang ved fjerning. Da har man den endelige AIC-minimerte modellen MOD-FINAL. Denne modell utgjør algoritmens output.

Denne algoritmen er implementert i R ved funksjonen "stepGAICAll.A" i GAMLSS pakken. For hver enkelt modell vil jeg først kjøre algoritmen for separate forklaringsvariabler, og til slutt for ulike varianter av samspill. Kandidatleddene som testes for hver parameter i hver modell vises i tabell 5.1.

[18] *Prioritert rekkefølge* skal her forstår som rekkefølgen parameterne har i fordelingens kortversjon, slik jeg definerer denne i delkapitler 3.5 og 3.6. Gammafordelingens kortversjon er for eksempel $Y \sim \Gamma(\mu,\nu)$, slik at prioritert rekkefølge på parameterne er μ,ν.

Beskrivelse	**Funksjonell form**
Årstall som lineær funksjon	t_i
Bilalder som polynom av grad 1-6.	b_i, $b_i + b_i^2$, ..., $b_i + b_i^2 + b_i^3 + b_i^4 + b_i^5 + b_i^6$ Jeg skriver $\mathbf{b}_{i,k}$ der k er høyeste eksponent.
Kategorisk bilalder (en kategori per bilalder). Referansekategori er bilalder 5 år.	$I(b_i^* = 0) + ... + I(b_i^* = 4)$ $+I(b_i^* = 6) + ... + I(b_i^* = 20)$
Personalder som polynom av grad 1-6.	p_i, $p_i + p_i^2$, ..., $p_i + p_i^2 + p_i^3 + p_i^4 + p_i^5 + p_i^6$ Jeg skriver $\mathbf{p}_{i,k}$ der k er høyeste eksponent.
Kategorisk personalder (inndelt i tiår). Referansekategori er 50-60 år.	$I(20 \le p_i^* \le 29) + ... + I(40 \le p_i^* \le 49)$ $+I(60 \le p_i \le 69) + ... + I(80 \le p_i \le 89)$

Tabell 5.1 - Kandidatledd for selvstendige forklaringsvariabler i modellene

Grunnen til at årstall kun testes som lineær funksjon er at modellene skal brukes til å spå fremtiden. Strengt tatt er ingen modeller gyldig utenfor datasettet som er brukt til å estimere parameterne. Når man spår fremtiden med en modell, putter man inn en verdi for forklaringsvariabelen *Årstall* som ikke har vært observert i datasettet. Avanserte, fleksible funksjoner vil fort kunne gi ekstreme og urealistiske utslag her. For eksempel kan det tenkes at en tidstrend over observasjonsperioden er formet som del av en parabel. Går man ut av observasjonsperioden er man innom en annen del av parabelen der grafen gjerne tar av. Generelt kan man si at å spå fremtiden er vanskelig, og at en lineær form på tidstrenden gir en nøktern, lettolkelig og relativt realistisk spådom i forhold til fleksible funksjoner med mange parametere.

Når en modell er AIC-minimert for disse kandidatleddene for alle fordelingsparametere, forsøker jeg å innføre samspill mellom forklaringsvariablene, og bruker algoritmen for minimering av AIC på kandidatleddene som er vist i tabell 5.2.

Beskrivelse	**Funksjonell form**
Årstall multiplisert med bilalder.	$t_i \cdot b_i$
Årstall multiplisert med personalder.	$t_i \cdot p_i$
Bilalder multiplisert med personalder.	$b_i \cdot p_i$
Kombinasjonen gammel bil og eldre kunde.	$I(b_i^* \geq 14, p_i^* \geq 60)$
Kombinasjonen gammel bil og ung kunde	$I(b_i^* \geq 14, p_i^* \leq 30)$
Kombinasjonen ny bil og eldre kunde.	$I(b_i^* \leq 5, p_i^* \geq 60)$
Kombinasjonen ny bil og ung kunde	$I(b_i^* \leq 5, p_i^* \leq 30)$

Tabell 5.2 - Kandidatledd for samspill mellom forklaringsvariablene i modellene

For hver sannsynlighetsfordeling som testes ut, vil jeg først definere en nullmodell (med postfiks 0), der forklaringsvariabler ikke brukes. Så definerer jeg en mellommodell (med postfiks 1) der jeg bruker årstall, bilalder og personalder, alle i referansejustert utgave, som lineære funksjoner. Til slutt definerer jeg en optimal modell (med postfiks 2), resultatet av algoritmen for AIC-minimering. AIC-kriteriet skal i teorien sikre at de AIC-minimerte modellene vil kunne spå fremtiden best. Imidlertid er disse langt mer fleksible enn nullmodellene og mellommodellene, og kan være utsatt for overparameterisering. I AIC-minimerings-algoritmen brukes data fra hele perioden, 2000-2005. Det er ikke gitt at modellen med best tilpasning for hele perioden også vil ha best tilpasning for eksempelvis 2005, gitt at parameterne er estimert på data fra 2000-2004. Jeg vil teste samtlige modeller ved kryss-validering i kapittel 9. Da kan det undersøkes i hvilken grad AIC-minimering fører til overparameterisering.

5.4 GAM-plot

Som et supplement til AIC-minimerings-algoritmen ser jeg også på GAM-plot for modellene. Ideen bak GAM - generaliserte additive modeller – er å erstatte koeffisientene i den lineære prediktor med glatte funksjoner. Vanlig lineær kobling mellom en fordelingsparameter μ_i, og

forklaringsvariablene er av typen $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$. GAM erstatter denne koblingen med $g(\mu_i) = \beta_0 + \sum_{j=1}^{p} s_j(x_{i,j})$, der s_j er glatte funksjoner, for eksempel spline-funksjoner. Det er upraktisk å skrive opp hele uttrykket for spline-funksjonene. GAMLSS-pakken i R gir imidlertid mulighet til å plotte grafene til disse spline-funksjonene. Jeg vil først kjøre AIC-minimerings-algoritmen for hver enkelt modell, og så bruke plottet av spline-funksjonene, GAM-plottet, som en sjekk på hvorvidt de funksjonelle former algoritmen har valgt, ligner på GAM-plottets grafer. Jeg sjekker mot GAM-plot for alle modeller, men i selve oppgaveteksten tar jeg kun med GAM-plot for Poissonmodellen, som et eksempel.

5.5 Korreksjon for eksponering

5.5.1 Generelt om korreksjon for eksponering

Noen av modellene har en observasjon per polise. For disse modellene er det av stor betydning hvor mange dager polisen har vært aktiv i aktuelt år. La $t_{D,.i}$ være antall aktive dager i aktuelt år for polise *i*. Jeg innfører eksponeringsvariabelen r_i gitt ved $r_i = \frac{t_{D,i}}{365 + I(\text{skuddår})}$ slik at r_i er andel av året polisen *i* er i kraft. Det fremgår av figur 4.5 at et stort antall poliser har $r_i < 1$. Det betyr at en eventuell ignorering av eksponeringen, r_i, kan være en stor feilkilde. Å ignorere eksponering er ekvivalent med å sette eksponering $r_i = 1$ for alle observasjoner, hvilket kan gi langt svakere modelltilpasning. Spørsmålet er så på hvilken måte man best tar hensyn til eksponering i modelleringen. Jeg lister opp 3 alternativer.

1. Eksponering kan inngå som "offset".
2. Eksponering kan inngå som forklaringsvariabel.
3. Logaritmen til eksponeringen kan inngå som forklaringsvariabel.

Jeg tester ut hvilket av disse tre alternativer som gir best resultat for den minst komplekse modellen som skal testes i kapitler 6-8, nemlig Poissonmodell for skadefrekvens.

5.5.2 Test av metodikk

I Poissonmodellen for antall skader, A_i, brukes log-link slik at $g(\mu_i) = \log(\mu_i)$. La nå λ_i være Poissonparameter for en fulleksponert polise (slik at λ_i tilsvarer forventet antall skader per år). En rimelig hypotese, H_0, er at forventet antall skader for polise *i* er proporsjonal med eksponeringen r_i. H_0 er ekvivalent med relasjonen

(6) $$\mu_i = r_i \cdot \lambda_i.$$

Når λ_i kobles til forklaringsvariablene, brukes log-link, $\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. Dette uttrykket, sammen med (6), kan skrives om til $\log(\mu_i) = \log(r_i) + \mathbf{x}_i^T \boldsymbol{\beta}$. Størrelsen $\log(r_i)$ kalles da offset. Hypotesen H_0 er derfor ekvivalent med å inkludere eksponeringen som offset. Dersom man bruker offset i modelleringen, estimeres det ingen koeffisient for $\log(r_i)$. Offsettet inngår kun som en korreksjon. Jeg tester de tre ulike alternativene ved å estimere parameterne i tre ulike Poissonmodeller for A_i der ingen forklaringsvariabler tas i bruk. Resultatene av denne testen vises i tabell 5.3.

Alt.	$\log(\mu_i)$	μ_i	**Estimat for** λ	**AIC**
1	$\log(\mu_i) = \log(r_i) + \beta_0$	$\mu_i = r_i \cdot \exp(\beta_0)$	0,2027057	54 770
2	$\log(\mu_i) = \beta_0 + \beta_1 \cdot (r_i - 1)$	$\mu_i = \exp(\beta_0 + \beta_1 \cdot (r_i - 1))$	0,1835075	54 469
3	$\log(\mu_i) = \beta_0 + \beta_1 \cdot \log(r_i)$	$\mu_i = r^{\beta_1} \cdot \exp(\beta_0)$	0,1812823	54 257

Tabell 5.3 - Testing av 3 alternative måter å korrigere for eksponering

Tabell 5.3 viser at alternativ 3 gir klart lavest AIC. Avstanden til alternativ 2 er såpass stor at jeg regner alternativ 3 som udiskutabelt best. Jeg vil derfor bruke korreksjonsmetodikken fra alternativ 3 i selve modelleringen. Med såpass stor forskjell i AIC mellom alternativ 1 og 3, kan det konkluderes med at hypotesen H_0 er feil. Estimatet for β_1 under alternativ 3 er 0,5142. Dette tallet er såpass nært 0,5 at jeg innfører følgende relasjon som et bedre alternativ til (6):

$$\mu_i = \sqrt{r_i} \cdot \lambda_i.$$

Denne relasjonen forteller at forventet antall skader for polise *i* er proporsjonal med kvadratroten av eksponeringen til polise *i*. Det kan være mange årsaker til denne sammenhengen. Her er noen forslag:

- Kunder som nettopp har tegnet en forsikring er mer uforsiktige enn andre kunder.
- Kunder som har opplevd skade skifter forsikringsselskap etter kort tid.

Det kan være meget interessant å gjøre videre undersøkelser rundt dette spørsmålet, men det ligger utenfor denne oppgavens mål. Jeg konkluderer imidlertid med at den beste måten å inkorporere eksponering i modellene på, er å la logaritmen til eksponeringen inngå som forklaringsvariabel.[19] Skadeforsikringskontrakter har vanligvis 1 års gyldighet, og prisene settes derfor i utgangspunktet som helårspriser (se finansportalen.no for eksempler). Når jeg velger å implementere eksponering som forklaringsvariabel, er det kun ment som en korreksjon, ikke som et redskap til prising av bilforsikring med valgfri forsikringsperiode.[20]

5.6 Korreksjon for antall skader

I modellering av skadepris vil jeg bruke gjennomsnittlig skadepris per skadepolise, G, som responsvariabel. I delkapittel 8.1 vises det at gjennomsnittlig skadepris for skadepolise i, G_i, og antall skader for polise i, A_i, er korrelert. I kapittel 6 og 7 vil jeg bygge delmodeller som forutsetter uavhengighet mellom disse størrelsene, og tillater separat modellering. Imidlertid velger jeg å forutsette uavhengighet mellom A_i og $G_i|A_i$, i stedet for mellom A_i og G_i. Det er da mulig å korrigere for antall skader i modellene for G_i.[21] Dette gjøres ved å inkludere referansejustert antall skader som forklaringsvariabel i alle modeller for $G_i|A_i$. Jeg innfører derfor forklaringsvariabelen $a_i^* = a_i - \bar{a}$, der $\bar{a}$ er observert gjennomsnitt av antall skader over alle poliser for hele perioden 2000-2005. Når modellene for skadefrekvens til slutt skal kobles mot modellene for skadepris, settes $a_i^* = \mathrm{E}(A_i) - \bar{a}$ inn som forklaringsvariabel i modellen for $G_i|A_i$.

---

[19] Tilsvarende resultat er også oppnådd ved testing for de andre modellene der eksponering inngår (NEGBIN, ZIP og ZAIG). Også her gir link-funksjon av eksponeringen som forklaringsvariabel den beste tilpasnningen.

[20] Det kunne vært interessant å modellere også for valgfri forsikringsperiode, men datasettet er ikke egent for dette. Tilgjengelige eksponeringsdata er kun "antall dager". Jeg har ikke tilgang til hvilke dager i året polisene er i kraft. Følgelig kan jeg ikke ta høyde for sesongvariasjon ved hjelp av dette datasettet.

[21] Jeg velger i det forestående å skrive G_i i stedet for $G_i|A_i$ for enkelhets skyld.

6 Modellering av skadefrekvens

6.1 Generelt om modellering av skadefrekvens

Dersom det antas at skadefrekvens og skadepris er uavhengige størrelser, er det naturlig å modellere disse separat. Jeg forsøker først å finne frem til best mulig modell for skadefrekvens, deretter best mulig modell for skadepris. Jeg vil så koble disse sammen slik at de endelige modellene kan predikere total utbetaling per polise, U_i. Responsvariabelen i modellering av skadefrekvens er antall skader på polise *i*, A_i. For polise *i* behandles antall skader, A_i, som en tilfeldig variabel med fordeling $f_i(\boldsymbol{\theta}_i)$. Det tillates med andre ord en egen parametervektor for hver polise. Modellene for skadefrekvens faller inn under det unimodale rammeverket, beskrevet i delkapittel 5.1. Generelt benyttes følgende rammeverk for skadefrekvensmodellene:

$$
\begin{aligned}
A_i &\sim f(\boldsymbol{\theta}_i) \\
g_1(\theta_{i,1}) &= \mathbf{x}_i^T \boldsymbol{\beta} \\
g_2(\theta_{i,2}) &= \mathbf{z}_i^T \boldsymbol{\gamma}
\end{aligned}
$$

I modellering av skadefrekvens vil jeg teste 3 ulike responsfordelinger: Poissonfordelingen, NEGBIN-fordelingen og ZIP-fordelingen. Log-transformasjonen av eksponeringen, r_i, er et element i designvektoren $\mathbf{x}_i$, og eventuelt også i designvektoren $\mathbf{z}_i$. $\mathbf{x}_i$ inneholder forklaringsvariabler for hovedparameteren, lokasjonsparameteren $\theta_{i,1}$, som åpenbart vil avhenge av eksponeringen. Hvorvidt $\theta_{i,2}$ også avhenger av eksponeringen, og dermed hvorvidt r_i også bør inngå i $\mathbf{z}_i$, må avgjøres separat for hver enkelt modell. Dette leder til modellformuleringene i tabell 6.1.

Poisson	NEGBIN	ZIP
$A_i \sim \text{Poisson}(\mu_i)$	$A_i \sim \text{NB}(\mu_i, \kappa_i)$	$A_i \sim \text{ZIP}(\psi_i, \mu_i)$
$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ $\log(\kappa_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$ $\log(\mu_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$

Tabell 6.1 - Generell formulering av skadefrekvensmodellene for de ulike fordelinger

Som det fremgår av tabell 6.1 velges log-link i Poissonmodellen. Dette er en velprøvd link-funksjon for Poissonfordelingen og et opplagt valg. NEGBIN-fordelingen er svært lik Poissonfordelingen og jeg velger log-link også her, for begge parametere. ZIP-fordelingen kan ses som en kombinasjon av Bernoullifordeling og Poissonfordeling. Jeg bruker derfor logit-link for ψ_i, som er parameter i Bernoulli-delen av fordelingen, mens jeg bruker log-link for μ_i, som er parameteren i Poisson-delen av fordelingen. Mer generelt er ZIP-fordelingen av typen FM, der den ene parameteren, ψ_i, er en sannsynlighet. Sannsynligheter modelleres naturlig ved logit-link, og jeg velger derfor denne link-funksjonen for ψ_i i alle de aktuelle FM-modellene.

Før jeg inkluderer forklaringsvariabler vil jeg undersøke hvilken av fordelingene som gir best tilpasning til responsvariabelen A, antall skader. Det er selvsagt ikke gitt, men likevel svært sannsynlig, at fordelingen som best beskriver responsvariabelen uten forklaringsvariabler, også vil beskrive responsen best med forklaringsvariabler. Grunnen er at dersom en fordeling beskriver responsobservasjonene godt, er den velegnet til å fange opp variabiliteten i responsvariabelen. Når forklaringsvariabler innføres, minsker variabiliteten, men den vil antagelig ha mye av den samme strukturen. Parameterne estimeres for hver av modellene uten forklaringsvariabler. Jeg tillater meg imidlertid å korrigere for eksponering her. Resultatet er vist i tabell 6.2.

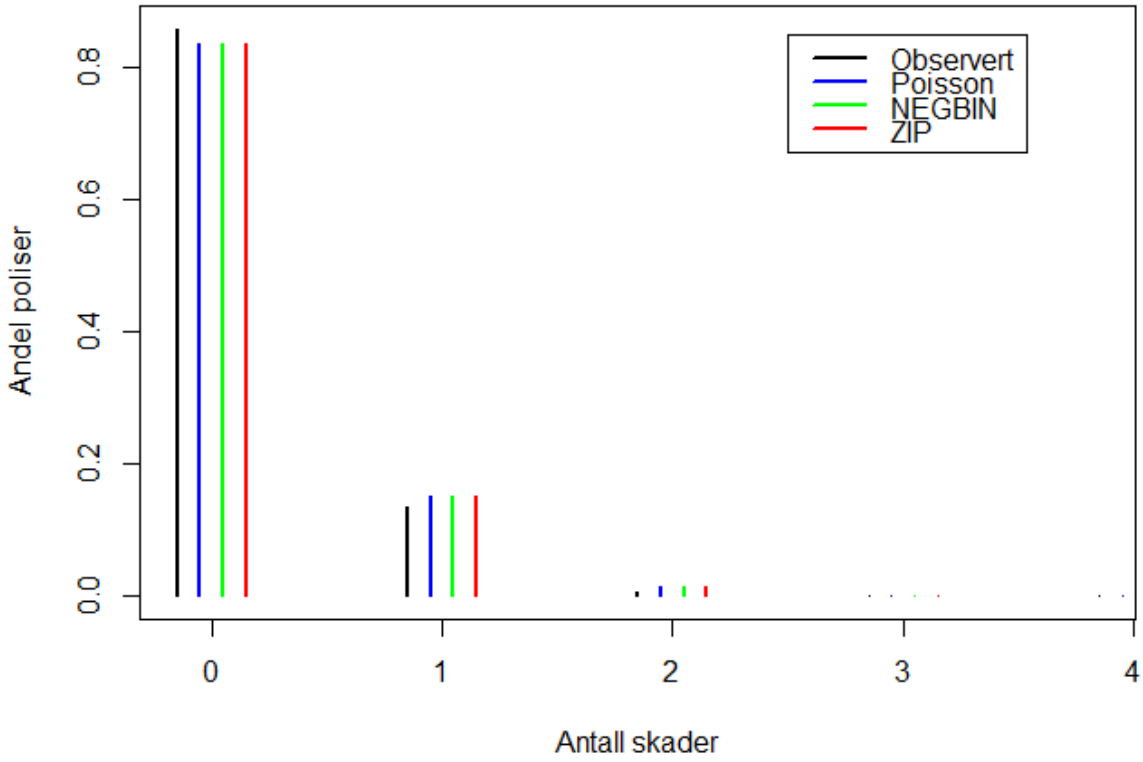
Responsfordeling	Estimat for λ	Estimat for den andre parameteren	AIC
Poisson	0,1812823	NA	54 257
NEGBIN	0,1812823	$\kappa = \exp(-36.04) \approx 0$	54 259
ZIP	0,1812823	$\psi = \text{expit}(-1.447 \cdot 10^{14}) \approx 0$	54 259

Tabell 6.2 – Estimater og AIC for Poissonmodell, NEGBIN-modell og ZIP-modell for skadefrekvens. Ingen forklaringsvariabler er tatt i bruk her (men det er korrigert for eksponering). λ er helårseksponert utgave av μ.

Poissonfordelingen ser ut til å komme best ut her. I NEGBIN-modellen og ZIP-modellen estimeres en ekstra parameter, som bestemmer fordelingens dispersjon. I begge tilfeller konvergerer estimatene for denne ekstra parameteren mot 0. En enkelt marginalanalyse i delkapittel 4.4.1 viste at gjennomsnitt og estimert varians for antall skader var svært nær hverandre. Dette, sammen med resultatene i tabell 6.2, gjør at jeg konkluderer med at det ikke er overdispersjon for antall skader. Det er også verdt å merke seg at estimatene for λ er identiske for alle tre modeller, og at AIC for Poissonfordelingen er nøyaktig 2 mindre enn for de 2 andre fordelingene. Dette betyr i praksis at de 3 modellene er så godt som identiske. Da tilsier Occams barberhøvel[22] at man bør velge den enkleste modellen, nemlig Poissonmodellen.

AIC er gitt ved $\text{AIC} = 2p - 2l$ (se delkapittel 3.1). Når da AIC for NEGBIN-modellen og ZIP-modellen er nøyaktig 2 høyere enn AIC for Poissonmodellen, kommer det av at en ekstra parameter estimeres, uten at den bidrar til å øke likelihooden. Figur 6.1 viser relativ frekvens av observert antall skader mot punktsannsynligheter for hver av de 3 sannsynlighetsfordelingene. Her kan man se grafisk at alle punktestimatene er identiske. Det er derfor ingen tvil om at Poissonmodellen er den foretrukne her. Jeg avskriver derfor NEGBIN-modellen og ZIP-modellen, og konsentrerer meg videre kun om Poissonfordelingen som responsfordeling i modellering av skadefrekvens.

[22] For mer om Occams barberhøvel (Occam's razor), se Dobson og Barnett (2008:36,85). De kaller den også "Law of parsimony". Ideen er at gitt 2 modeller med identisk forklaringskraft, er den enkeleste modellen å foretrekke.



Figur 6.1 - Punktsannsynligheter for 0-4 skader gitt av de ulike fordelingene ved de estimerte parameterne uten forklaringsvariabler. Punktsannsynlighetene er sammenlignet med observert relativ frekvens.

6.2 Poissonmodell for skadefrekvens

6.2.1 Estimering og definisjoner

I delkapittel 6.1 er parameterne i nullmodellen APOI-0, som ikke bruker andre forklaringsvariabler enn log-transformert eksponering, allerede estimert. Videre estimeres mellommodellen APOI-1, som kun bruker lineære forklaringsvariabler uten samspill. Til slutt kjører jeg AIC-minimerings-algoritmen (se delkapittel 5.3), og får den AIC-minimerte modellen APOI-2. Tabell 6.3 definerer disse 3 Poissonmodellene.

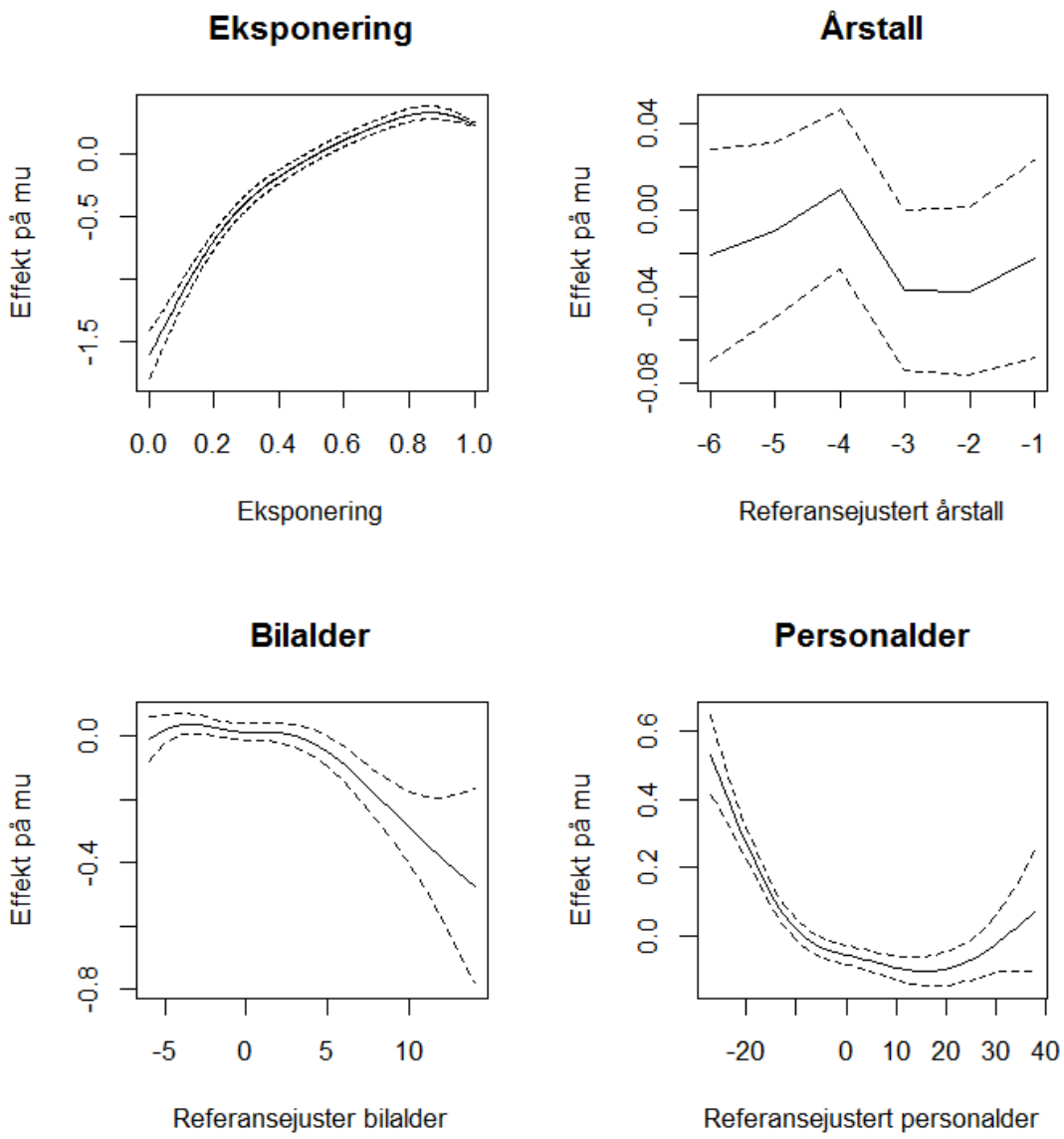
Modell	$A_i \sim$	Kobling til forklaringsvariablene
APOI-0	$\text{PO}(\mu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & \log(r_i) \end{bmatrix} \boldsymbol{\beta}$
APOI-1	$\text{PO}(\mu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & \log(r_i) & t_i & b_i & p_i \end{bmatrix} \boldsymbol{\beta}$
APOI-2	$\text{PO}(\mu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & \log(r_i) & \mathbf{b}_{i,6} & \mathbf{p}_{i,4} & t_i \cdot b_i & I(b_i^* \geq 14, p_i^* \leq 30) \end{bmatrix} \boldsymbol{\beta}$

Tabell 6.3 - Definisjon av APOI-modellene

AIC-verdiene for APOI-modellene er henholdsvis 54 257, 54 140 og 54 043, hvilket tilsier at APOI-2 er den suverent beste tilpasningen etter AIC-kriteriet. Med sine 14 estimerte parametere kan APOI-2 regnes som moderat fleksibel.

6.2.2 GAM-plot

Jeg estimerer spline-funksjoner (se delkapittel 3.3) for eksponeringen, og referansejusterte utgaver av årstall, bilalder og personalder. Disse GAM-plottene sammenliknes så med resultatet av AIC-minimerings-algoritmen. Dette gir en nyttig sjekk på hvorvidt algoritmen har truffet de optimale funksjonelle formene. GAM-plottene vises i figur 6.2. Som det fremgår av denne figuren har eksponeringen en funksjonell form som ikke er veldig ulik en logaritmisk kurve. Det er tvilsomt om årstall har signifikant effekt, ettersom 0 er innenfor standardavviket for hvert av årene. Bilalder ser ut til å modelleres godt ved et polynom av moderat høy grad, mens personalder ser ut til å passe med et andregradspolynom eller et fjerdegradspolynom. Dette stemmer svært godt overens med de funksjonelle formene i modellen APOI-2, som er resultatet av AIC-minimerings-algoritmen. Der er årstall ikke med som forklaringsvariabel, mens bilalder og personalder er med som henholdsvis tredjegrads- og fjærdegradspolynom. Optimalt antall frihetsgrader (hyperparameteren Λ fra delkapittel 3.3) er estimert til 17. I dette estimatet er eksponeringen gitt en funksjonell form som kan minne om et polynom av minst tredje grad. Ved å log-transformere eksponeringen spares det dermed noen frihetsgrader (effektive parametere). Jeg konkluderer med at den ikke-parametriske GAM-analysen og AIC-minimerings-algoritmen gir noenlunde samsvarende resultat.



Figur 6.2 - GAM-plot for selvstendige forklaringsvariabler i Poissonmodell for skadefrekvens. Stiplede linjer er standardavvik.

6.3 Effekter av forklaringsvariablene på skadefrekvens

GAM-plottene i figur 6.2 gir en god grafisk oversikt over effektene av de ulike forklaringsvariablene på forventet skadefrekvens. Et annet godt utgangspunkt for å drøfte effekter, er å studere modellestimatene til APOI-1. Denne modellen har lineære forklaringsvariabler. Det gjør den lettolkelig ettersom det gir en koeffisient per forklaringsvariabel, og denne koeffisientens fortegn viser i hvilken retning forventet skadefrekvens flytter seg ved en økning i verdien på forklaringsvariabelen.

μ – koeffisient for	Estimat	Standardfeil	*p*-verdi	exp(estimat)
1	-1,7170	0,0246	0,0000	0,17960128
$\log(r_i)$	0,5403	0,0196	0,0000	1,7165904
t_i	-0,0044	0,0062	0,4829	0,99562759
b_i	-0,0125	0,0028	0,0000	0,98759656
p_i	-0,0081	0,0008	0,0000	0,99193867

Tabell 6.4 – APOI-1 estimater med standardfeil og p-verdier. exp(estimat) gir faktisk multiplikativ effekt på forventet skadefrekvens.

Som det fremgår av tabell 6.4 er årstallet, t_i, langt fra signifikant. Imidlertid er bilalder, b_i, og personalder, p_i, begge svært signifikante, og begge har negativ effekt på forventet antall skader. At skadefrekvensen synker med personalder, stemmer med hypotesen jeg formulerte i delkapittel 4.3.3. At skadefrekvensen synker med bilalder, strider imidlertid med hypotesen jeg formulerte i delkapittel 4.3.2. Det er rimelig å tro at gamle biler er i dårligere stand enn nye biler. Imidlertid kan andre effekter, som for eksempel kjøremønster, spille inn her slik at skadefrekvensen likevel synker med bilalder. APOI-1 har log-link, $g(\mu_i) = \log(\mu_i)$, og er følgelig en multiplikativ faktor-modell. Ut fra kolonnen exp(estimat) kan man si at forventet skadefrekvens for gjennomsnittlig bilalder (5,9 år) og gjennomsnittlig personalder (50,3 år) er 0,18. Ved å se på tallene i denne kolonnen, kan man også hevde at for hvert år bilalder øker går skadefrekvensen ned med ca. 1 %. For hvert år personalder øker går også skadefrekvensen ned ca. 1 %. Eksponeringen har åpenbart en svært kraftig effekt, men denne regnes ikke som en forklaringsvariabel på samme måte som bilalder og personalder. Logaritmejustert eksponering er kun tatt med som en korreksjon.

7 Modellering av skadepris

7.1 Generelt om modellering av skadepris

Mitt endelige mål er å modellere total utbetaling for alle skader på hver polise *i*, U_i. Når jeg modellerer skadepris for seg, baserer jeg meg på antagelsen om at skadefrekvens og skadepris er uavhengige størrelser. Det er tre mulige valg av responsvariabel når skadepris skal modelleres:

- Skadepris per skade, *S*.
- Gjennomsnittlig skadepris per skadepolise, *G*
- Total utbetaling for alle skader per polise, *U*

Den totale skadeprisen, *U*, modelleres direkte i kapittel 8. Her i kapittel 7 ser jeg først og fremst på *S* og *G* som kandidater. Figur 4.7 viser at histogrammene over log(*S*) og log(*G*) er bortimot identiske. En god regel for statistisk modellering er å ikke aggregere uten grunn, ettersom noe informasjon alltid går tapt når man aggregerer. Denne regelen taler til fordel for modellering av *S*. På den annen side har denne oppgaven som overordnet mål å modellere total utbetaling, *U*. Det vil si at ønsket sluttprodukt ligger naturlig på polisenivå og ikke på skadenivå. Dette taler til fordel av modellering av *G*. Jeg antar på forhånd at modellering av *S* og *G* vil gi svært like modeller. Jeg vil i delkapittel 7.4.2 undersøke hvorvidt dette stemmer. I det følgende menes *G* når uttrykket *skadepris* tas i bruk. Det fremgår av figur 4.7 at det kan være aktuelt å se på bimodale modeller for skadepris, i tillegg til unimodale. I selve modelleringen setter jeg $G = G|A,$ slik at det kan korrigeres for antall skader, som beskrevet i delkapittel 5.6.

7.2 Unimodale modeller for skadepris

Jeg bruker følgende generelle, unimodale modellformulering for gjennomsnittlig skadepris, $G_i|A_i,$ per skadepolise: (Det er underforstått at a_i^* (se delkapittel 5.6) inngår i $\mathbf{x}_i$ og eventuelt også i $\mathbf{z}_i$).

$$G_i|A_i \sim f(\boldsymbol{\theta}_i)$$
$$g_1(\theta_{i,1}) = \mathbf{x}_i^T\boldsymbol{\beta}$$
$$g_2(\theta_{i,2}) = \mathbf{z}_i^T\boldsymbol{\gamma}$$

Skadepris er en positiv, kontinuerlig størrelse som ikke nødvendigvis er symmetrisk. Det finnes en rekke sannsynlighetsfordelinger som oppfyller disse kravene. Jeg tester ut følgende varianter: Normalfordeling, lognormalfordeling, gammafordeling, IG-fordeling og Weibullfordeling. Jeg bruker modellformuleringene som vises i tabell 7.1.

Normal	**Lognormal**	**Gamma**	**IG**	**Weibull**
$G_i \sim N(\mu_i, \sigma_i^2)$	$G_i \sim \log N(\mu_i, \sigma_i)$	$G_i \sim \Gamma(\mu_i, \nu_i)$	$G_i \sim \mathrm{IG}(\mu_i, \nu_i)$	$G_i \sim \mathrm{WEI}(\lambda_i, \kappa_i)$
$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$	$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$
$\log(\sigma_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$	$\log(\sigma_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$	$\log(\nu_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$	$\log(\nu_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$	$\log(\kappa_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$

Tabell 7.1 - Generell formulering av skadeprismodellene for de ulike fordelingene

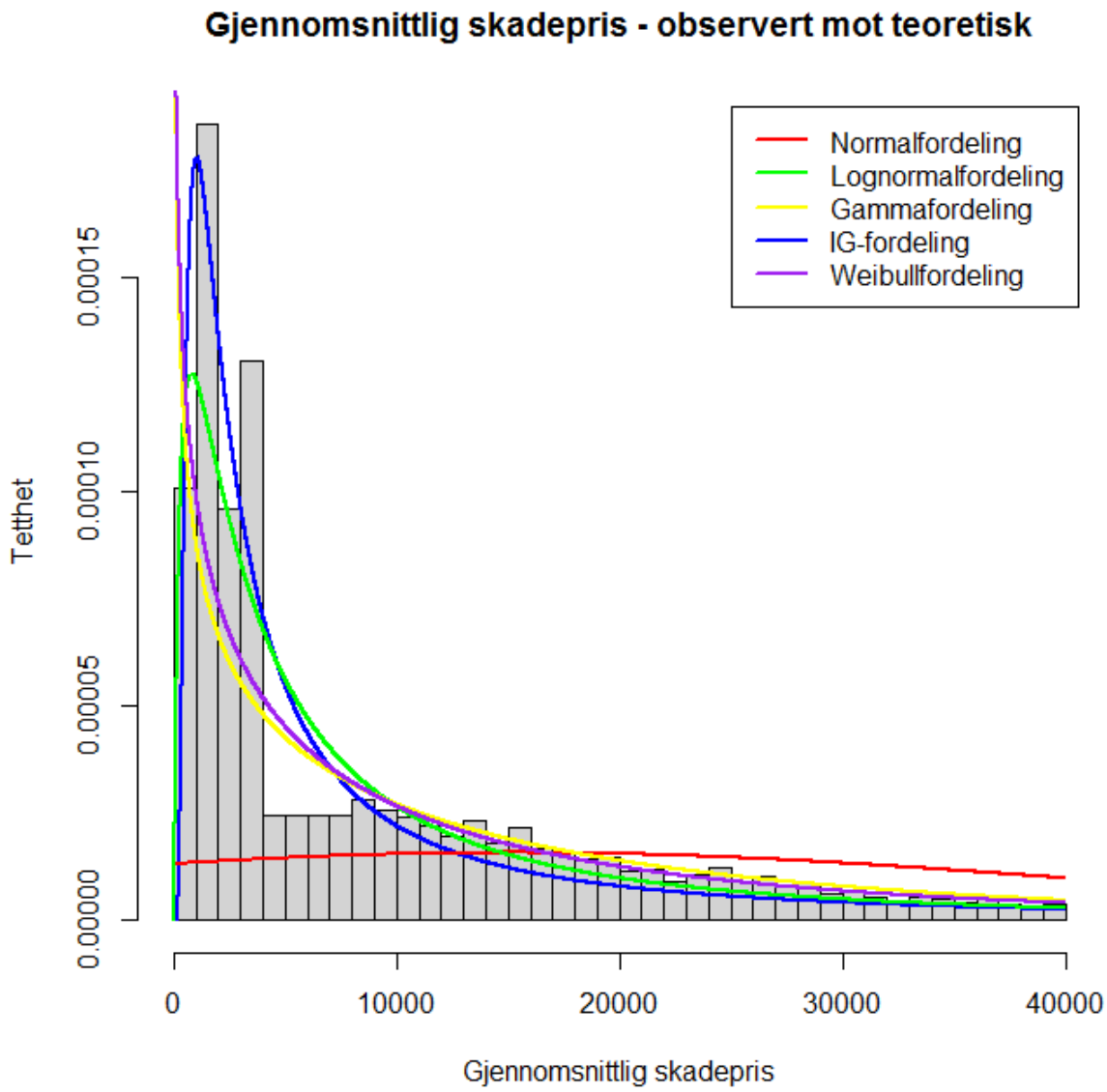
På samme måte som for skadefrekvens, estimeres først parameterne for hver modell uten bruk av forklaringsvariabler (bortsett fra korreksjon for antall skader). Dette gir en pekepinn på hvilke fordelinger som best kan beskrive variabiliteten i skadeprisen. Resultatet vises i tabell 7.2.

Responsfordeling	**Forventet skadepris gitt 1 skade**	**AIC**
Normal	15 269	206 916
Lognormal	15 942	187 445
Gamma	15 257	189 324
IG	15 270	187 062
Weibull	14 750	188 740

Tabell 7.2 - Estimater og AIC for normalmodell, lognormalmodell, gammamodell, IG-modell og Weibullmodell for skadepris. Ingen forklaringsvariabler er tatt i bruk her.

Figur 7.1 viser tetthetskurvene til disse fordelingene i et histogram over de observerte skadepriser. Av dette plottet fremgår det at den observerte fordelingen er meget skeiv. Det er tydelig av denne grafen at normalfordelingens PDF er uegnet til å beskrive de observerte skadepriser. IG-fordelingen ser ut til å passe best, mens lognormalfordelingen passer nest best. Det er rimelig å anta at disse to fordelinger best vil beskrive variabiliteten i skadeprisen, også når forklaringsvariabler er innført. Jeg velger derfor å gå videre med disse fordelinger og forkaste de resterende. Ettersom IG-fordelingen har den aller beste tilpasningen, vil jeg også

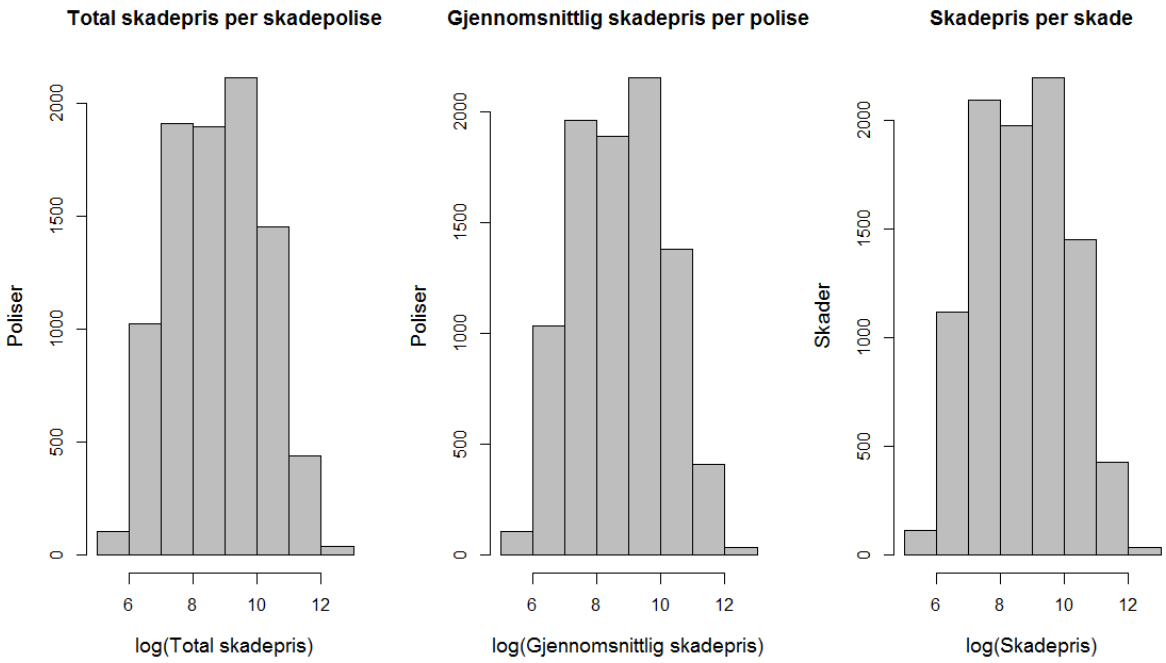
teste den AIC-minimerte IG-modellen for hvorvidt parameterne i modellen forandrer seg når man bytter ut $G_i | A_i$ med S_i som responsvariabel.



Figur 7.1 - Histogram av gjennomsnittlig skadepris per skadepolise, G, sammen med PDF for de ulike fordelinger der parameterne er estimert uten bruk av forklaringsvariabler. Histogrammet tar kun med gjennomsnittlige skadepriser der $G < 40\ 000$. *Det vil si at ca. 91 % av polisene med skader er tatt med. Histogrammet er kuttet av for å gi et tydeligere bilde.*

7.3 Lognormalmodell for skadepris

Lognormalfordelingen er rimelig å bruke dersom logaritmen til skadeprisen er normalfordelt. Plottene i figur 4.7 ser ikke ut normalfordelte ut, men, som nevnt i delkapittel 4.4.2 fører registreringspraksis i forsikringsselskapet til at observert skadepris ikke alltid opptrer som en kontinuerlig variabel. Dersom histogrammene aggregeres til et langt grovere format, gir det histogrammene av log(skadepris) som presenteres i figur 7.2.



Figur 7.2 - Grove histogrammer av log(skadepris) for total skadepris U, gjennomsnittlig skadepris G og skadepris per skade S.

Histogrammene i figur 7.2 ser ikke perfekt normalfordelte ut, men ser heller ikke ut til å være veldig langt unna. Heller enn å modellere logaritmen til skadeprisen ved normalfordelingen velger jeg å modellere skadeprisen direkte ved lognormalfordelingen. Dette gjør modellen lettere å tolke, og gjør AIC-verdien direkte sammenlignbar med AIC-verdier til andre modeller for skadepris. Jeg velger log-link for både μ_i og σ_i.

Etter å ha kjørt AIC-minimerings-algoritmen ender jeg opp med en modell, GLOG-2, som er ekstremt kompleks. Den har hele 24 estimerte parametere, hvorav 6 er koeffisienter for

samspillvariabler. Jeg definerer den AIC-minimerte modellen GLOG-2, sammen med nullmodellen, GLOG-0, og mellommodellen, GLOG-1, i tabell 7.3.

Modell	$G_i \sim$	**Kobling til forklaringsvariablene**
GLOG-0	$\log N(\mu_i, \sigma_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* \end{bmatrix} \boldsymbol{\beta}$ $\log(\sigma_i) = \begin{bmatrix} 1 & a_i^* \end{bmatrix} \boldsymbol{\gamma}$
GLOG-1	$\log N(\mu_i, \sigma_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* & t_i & b_i & p_i \end{bmatrix} \boldsymbol{\beta}$ $\log(\sigma_i) = \begin{bmatrix} 1 & a_i^* & t_i & b_i & p_i \end{bmatrix} \boldsymbol{\gamma}$
GLOG-2	$\log N(\mu_i, \sigma_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* & t_i & \mathbf{b}_{i,5} & \mathbf{p}_{i,4} & t_i \cdot b_i & b_i \cdot p_i \end{bmatrix} \boldsymbol{\beta}$ $\log(\sigma_i) = \begin{bmatrix} 1 & a_i^* & t_i & \mathbf{b}_{i,2} & \mathbf{p}_{i,2} & b_i \cdot p_i \end{bmatrix} \boldsymbol{\gamma}_1$ $+ \begin{bmatrix} I(b_i^* \geq 14, p_i^* \leq 30) & I(b_i^* \geq 14, p_i^* \geq 60) & I(b_i^* \leq 5, p_i^* \geq 60) \end{bmatrix} \boldsymbol{\gamma}_2$

Tabell 7.3 - Definisjon av GLOG-modellene

AIC-verdier for modellene GLOG-0, GLOG-1 og GLOG-2 er henholdsvis 187 446, 187 363 og 187 319. Ikke alle estimerte parametere i GLOG-2-modellen har god statistisk signifikans. Dersom man fjerner for eksempel p_i^2 som forklaringsvariabel for μ_i, vil AIC gå noe ned. Jeg regner det imidlertid ikke som god modelleringspraksis å fjerne for eksempel andregradsleddet i et fjerdegradspolynom. Det kan i enkelte tilfeller gi lavere AIC, men modellen mister mye av fleksibiliteten. Dersom modellen skal estimeres for andre data enn akkurat de modellen er estimert på, vil dette tapet av fleksibilitet merkes og gi en dårligere modell. AIC går markant nedover fra GLOG-0 til GLOG-1, og fra GLOG-1 til GLOG-2. Imidlertid er AIC for IG-modellen uten forklaringsvariabler 187 062. Dette er betraktelig lavere enn AIC-verdien til GLOG-2, til tross for at ingen forklaringsvariabler er tatt i bruk. Dette kan tyde på at mesteparten av variabiliteten i skadepris ikke kan redegjøres for ved hjelp av forklaringsvariablene, men må tilskrives stokastisk variabilitet, eller andre forklaringsvariabler som jeg ikke har tilgang til. Til tross for at IG-modellen gir lavere AIC, har lognormalmodellen den fordel at den er mer lik klassisk lineær modellering, og derfor er lettere å tolke. En annen mulig svakhet ved GLOG-modellene er at de er unimodale. Jeg vil i delkapitler 7.5-7.7 bruke FM-rammeverket til å lage mer fleksible, bimodale modeller for skadepris.

7.4 IG-modell for skadepris

7.4.1 Estimering og definisjon

Tabell 7.2 viser at IG-fordelingen gir den overlegent beste tilpasning til observerte skadepriser av de fem unimodale sannsynlighetsfordelinger som er testet ut. IG-fordelingen er kjennetegnet av dens bratte topp og ekstremt skjeve form. Dette er også en gyldig karakteristikk av observerte skadepriser, og i det ligger nok forklaringen på den gode tilpasningen man får ved IG-fordeling. Faktisk er IG-modellen, selv uten forklaringsvariabler, langt bedre tilpasset data enn lognormalmodeller med en rekke forklaringsvariabler. Følgelig har jeg størst forhåpninger til IG-modellen av de unimodale skadeprismodellene.

Algoritmen for AIC-minimering gir resultatmodellen GIG-2, med 17 parametere, hvorav 5 er koeffisienter for samspillvariabler. Jeg definerer GIG-2, sammen med nullmodell GIG-0 og mellommodell GIG-1, i tabell 7.4.

Modell	$G_i \sim$	**Kobling til forklaringsvariablene**
GIG-0	$\text{IG}(\mu_i, \nu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* \end{bmatrix} \boldsymbol{\beta}$ $\log(\nu_i) = \begin{bmatrix} 1 & a_i^* \end{bmatrix} \boldsymbol{\gamma}$
GIG-1	$\text{IG}(\mu_i, \nu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* & t_i & b_i & p_i \end{bmatrix} \boldsymbol{\beta}$ $\log(\nu_i) = \begin{bmatrix} 1 & a_i^* & t_i & b_i & p_i \end{bmatrix} \boldsymbol{\gamma}$
GIG-2	$\text{IG}(\mu_i, \nu_i)$	$\log(\mu_i) = \begin{bmatrix} 1 & a_i^* & p_i & b_i \cdot p_i & I(b_i^* \le 5, p_i^* \ge 60) \end{bmatrix} \boldsymbol{\beta}$ $\log(\nu_i) = \begin{bmatrix} 1 & a_i^* & \mathbf{b}_{i,5} & \mathbf{p}_{i,2} & t_i \cdot b_i & I(b_i^* \ge 14, p_i^* \le 30) & I(b_i^* \ge 14, p_i^* \ge 60) \end{bmatrix} \boldsymbol{\gamma}$

Tabell 7.4 - Definisjon av GIG-modellene

AIC-verdier for GIG-0, GIG-1 og GIG-2 er henholdsvis 187 062, 187 004 og 186 976. Forbedringen i AIC-verdi er mindre markant her enn for GLOG-modellene. Dette kan være et tegn på at andre forklaringsvariabler jeg ikke har tatt med er av stor betydning. En annen mulighet er at variabiliteten i gjennomsnittlig skadepris er stor, og allerede relativt godt gjort rede for ved valg av responsfordeling. Videre undersøker jeg i hvor stor grad modellen forandrer seg dersom skadepris per skade, S_i, erstatter gjennomsnittlig skadepirs per polise gitt antall skader, $G_i \mid A_i$, som responsvariabel.

7.4.2 Testing av *S* kontra *G* som responsvariabel

Gjennomsnittlig skadepris, *G*, er foreløpig valgt som responsvariabel i skadeprismodellene. Jeg er imidlertid i tvil om hvorvidt dette er et riktig valg. Ettersom skadepris per skade, *S*, ikke er aggregert, vil en modell basert på *S* muligens kunne beskrive data bedre. Jeg ønsker å teste i hvor stor grad disse modellene skiller seg fra hverandre. Jeg definerer derfor SIG-2 som identisk med GIG-2, bortsett fra at denne har *S* som responsvariabel. Etter at parameterne for SIG-2 er estimert, sammenlignes disse med de estimerte parameterne til GIG-2. $\mu-$ estimatene sammenliknes i tabell 7.5, mens $\sigma-$estimatene sammenliknes i tabell 7.6.

μ **-koeffisient for**	**GIG-2-estimat**	**SIG-2-estimat**	**Prosentvis avvik**
1	9,989529315	9,868894	1,2 %
t_i	0,068355899	0,068481	0,2 %
p_i	0,004523711	0,004573	1,1 %
$b_i \cdot p_i$	-0,001378291	-0,00136	1,1 %
$I(b_i^* \leq 5, p_i^* \geq 60)$	-0,139569758	-0,13655	2,2 %
a_i^*	-0,128642	NA	NA

Tabell 7.5 - Sammenlikning av $\mu-$koeffisienter for AIC-minimert modell GIG-2, og tilsvarende modell SIG-2, der prisen på enkeltskader utgjør responsvariabelen.

$\sigma-$**koeffisient for**	**GIG-2-estimat**	**SIG-2-estimat**	**Prosentvis avvik**
1	-3,785114658	-3,992634405	5,5 %
b_i	-0,014143607	-0,014403998	1,8 %
b_i^2	-0,004879906	-0,004307361	11,7 %
b_i^3	0,000106227	9,78164E-05	7,9 %
b_i^4	9,69704E-05	8,27618E-05	14,7 %
b_i^5	-5,93227E-06	-5,04166E-06	15,0 %
p_i	-0,003563736	-0,003520667	1,2 %
p_i^2	-0,000113494	-0,000102706	9,5 %
$t_i \cdot b_i$	-0,003048707	-0,003116169	2,2 %
$I(b_i^* \geq 14, p_i^* \leq 30)$	-0,458965027	-0,466393451	1,6 %
$I(b_i^* \geq 14, p_i^* \geq 60)$	0,13761716	0,133795426	2,8 %
a_i^*	-0,2415	NA	NA

Tabell 7.6 - Sammenlikning av $\sigma-$koeffisienter for AIC-minimert modell GIG-2, og tilsvarende modell SIG-2, der prisen på enkeltskader utgjør responsvariabelen.

Som man enkelt kan regne ut ved hjelp av tabellene 7.5 og 7.6, er gjennomsnittlig avvik henholdsvis 1.1 % og 6,7 % for μ og σ i den lineære prediktor. For noen få av parameterne for σ er avviket relativt stort. Dette gjelder parametere med lav absoluttverdi og muligens lav signifikans. Det betyr at modellens prediksjoner ikke vil bli veldig sterkt påvirket av disse avvikene. Et annet viktig poeng er at alle estimerte parametere har samme fortegn i SIG-2 som i GIG-2. I tillegg er det verdt å merke seg at det ikke ser ut til å være noe mønster i avvikene mellom modellene. For 8 av de 16 parameterne som er felles for modellene, har GIG-2-estimatet høyere verdi enn SIG-2-estimatet. Dette tilsvarer 50-50- fordeling. Dersom det hadde vært en klar tendens til at den ene av modellene gav høyere estimater enn den andre, ville det i større grad tydet på en systematisk forskjell. Ettersom det er ca. 50-50 fordeling i oppgang og nedgang på parameterne fra GIG-2 til SIG-2, er det rimelig å anta at forskjellen er tilfeldig. Alt i alt gir ikke denne testen noe grunnlag for å bytte ut $G|A$ med S i den videre modelleringsprosessen.

7.5 Bimodale modeller for skadepris

Bimodale modeller, slik de defineres i delkapittel 5.2, er svært relevante å se på her. Generelt bruker jeg følgende modellformulering, der det er underforstått at korreksjon for antall skader gjøres ved å inkludere a_i^* som forklaringsvariabel i designvektoeren $\mathbf{x}_i$:

$$
\begin{aligned}
&G_i \sim f_{i,M}(\boldsymbol{\theta}_{i,M}) \\
&f_{i,M}(y_i;\boldsymbol{\theta}_{i,M}) = \psi_i \cdot f_{i,1}(y_i;\boldsymbol{\theta}_{i,1}) + (1-\psi_i) \cdot f_{i,2}(y_i;\boldsymbol{\theta}_{i,2}) \\
&g_1(\psi_i) = \mathbf{x}_i^T \boldsymbol{\beta} \\
&g_2(\mu_{i,1}) = \begin{bmatrix} 1 & t_i \end{bmatrix} \boldsymbol{\gamma} \\
&g_3(\mu_{i,2}) = \begin{bmatrix} 1 & t_i \end{bmatrix} \boldsymbol{\delta}
\end{aligned}
$$

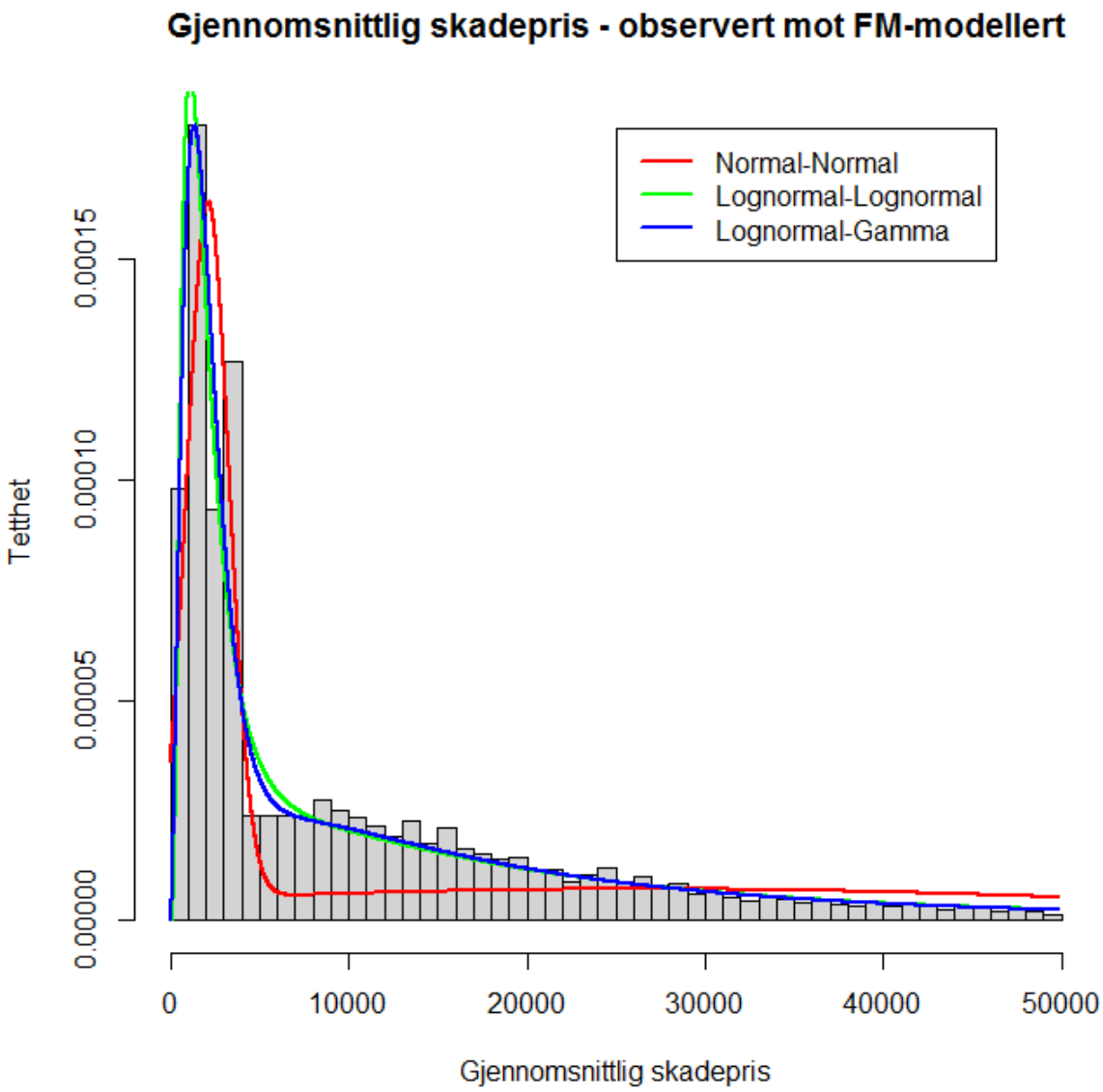
Responsfordelingen $f_{i,M}$ er en mikstur av to delfordelinger ($f_{i,1}$ og $f_{i,2}$), med hver sin estimerte sannsynlighetsparameter (ψ_1 og $\psi_2 = 1-\psi_1$). Parameterne $\mu_{i,1}$ og $\mu_{i,2}$ er lokasjonsparametere tilhørende hver delfordeling. I Weibullfordelingen byttes μ ut med λ. Parameterne som ikke er listet i formuleringen over estimeres uten direkte kobling til forklaringsvariabler. De to delfordelingene må oppfylle tilsvarende krav som de unimodale skadeprisfordelingene. Det er derfor rimelig at jeg tester de samme fordelingene som delfordelinger også her. En første test utføres der jeg ikke lar parameterne avhenge av noen

forklaringsvariabler, bortsett fra at ψ_i korrigeres for antall skader. Resultatet av disse testene presenteres i tabell 7.7.

Delfordeling 1	**Delfordeling 2**	**AIC**
Lognormal	Lognormal	186 492
Lognormal	Gamma	186 502
Lognormal	IG	186 519
IG	IG	186 546
Gamma	IG	186 655
IG	Weibull	186 694
Normal	IG	186 738
Lognormal	Weibull	186 848
Gamma	Weibull	186 889
Gamma	Gamma	186 930
Weibull	Weibull	186 945
Normal	Lognormal	187 381
Normal	Weibull	188 551
Normal	Gamma	188 759
Normal	Normal	193 391

Tabell 7.7 - AIC-verdier for bimodale modeller for gjennomsnittlig skadepris, uten forklaringsvariabler.

Det fremgår av tabell 7.7 at delfordelingene som gir lavet AIC-score er lognormalfordelinger for både $f_{i,1}$ og $f_{i,2}$. Modellen med nest lavest AIC-score er lognormalfordeling for $f_{i,1}$ og gammafordeling for $f_{i,2}$. Jeg vil se nærmere på disse to modellene, og forsøke å innføre forklaringsvariabler for ψ_i for disse 2 modeller. Det er verdt å merke seg at de beste av disse bimodale modeller gir langt lavere AIC enn de beste unimodale modeller. Dette kan skyldes at skadepriser fastsettes som to ulike prosesser. Småskader behandles på en mer rigid og strømlinjeformet måte. Her opereres det i større grad med faste priser for standardreparasjoner. Det fører til at man får en topp rundt prisen på disse standardreparasjonene. Større skader vil gjerne behandles mer detaljert og vil derfor i større grad følge et naturlig stokastisk mønster. Skadepriser i denne kategorien kan til en viss grad bestå av oppsummerte priser på standardreparasjoner, men jo mer komplisert skaden er, jo mindre standardisert blir totalprisen.



Figur 7.3 - Histogram over observert gjennomsnittlig skadepris mot tettheten til de to FM-modellene som gir lavest AIC (Lognormal-Lognormal og Lognormal-Gamma), og tettheten til FM-modellen med høyest AIC (Normal-Normal).

Figur 7.3 viser tydelig at normalfordelingen er uegnet til å beskrive skadepriser, selv når den får tilgang til to topper. Denne figuren viser også at de to best tilpassede modellene følger histogrammet godt. Hvis man sammenlikner figur 7.3 med figur 7.1, ser man at de beste bimodale modellene er klart bedre tilpasset enn de beste unimodale. Dette tyder på at den ekstra fleksibiliteten det gir å kombinere to fordelinger, gir langt bedre tilpasning, så lenge man velger riktige delfordelinger.

7.6 FM-log-log-modell for skadepris

FM-modellen med lognormalfordeling for både $f_{i,1}$ og $f_{i,2}$, kaller jeg Log-log-modell for skadepris. Jeg kjører AIC-minimerings-algoritmen og ender opp med modell GLOGLOG-2. Det er en svært kompleks modell med 20 parametere. Jeg definerer den i tabell 7.8, sammen med nullmodellen GLOGLOG-0 og mellommodellen GLOGLOG-1.

Modell	$G \sim$	**Kobling til forklaringsvariablene**
GLOG LOG-0	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \log N(\mu_{2,i}, \sigma_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \left[1 \quad a_i^*\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \gamma_0$ $\log(\mu_{2,i}) = \delta_0$
GLOG LOG-1	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \log N(\mu_{2,i}, \sigma_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \left[1 \quad a_i^* \quad b_i \quad p_i\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \left[1 \quad t_i\right]\boldsymbol{\gamma}$ $\log(\mu_{2,i}) = \left[1 \quad t_i\right]\boldsymbol{\delta}$
GLOG LOG-2	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \log N(\mu_{2,i}, \sigma_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) =$ $\left[1 \quad a_i^* \quad t_i \quad \mathbf{b}_{i,5} \quad \mathbf{p}_{i,4} \quad t_i \cdot b_i \quad b_i \cdot p_i \quad I(b_i^* \geq 14, p_i^* \geq 60)\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \left[1 \quad t_i\right]\boldsymbol{\gamma}$ $\log(\mu_{2,i}) = \left[1 \quad t_i\right]\boldsymbol{\delta}$

Tabell 7.8 - Definisjon av GLOGLOG-modellene. $\sigma-$parameterne estimeres uten direkte kobling mot forklaringsvariabler.

AIC-verdier for modellene GLOGLOG-0, GLOGLOG-1 og GLOGLOG-2 er henholdsvis 186 460, 186 304 og 186 259. Det er med andre ord en klar AIC-gevinst ved å gi modellen større fleksibilitet.

7.7 FM-log-gamma-modell for skadepris

Modellen av typen FM, der $f_{i,1}$ er lognormalfordelt og $f_{i,2}$ er gammafordelt kaller jeg log-gamma-modell for skadepris. Jeg kjører algoritmen for minimering av AIC og ender opp med GLOGGA-2 som er AIC-minimert med hensyn på forklaringsvariablenes funksjonelle form.

Videre definerer jeg nullmodellen GLOGGA-0 og mellommodellen GLOGGA-1. Resultatet presenteres i tabell 7.9.

Modell	$G \sim$	**Kobling til forklaringsvariablene**
GLOG GA-0	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \Gamma(\mu_{2,i}, \nu_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \left[1 \quad a_i^*\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \gamma_0$ $\log(\mu_{2,i}) = \delta_0$
GLOG GA-1	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \Gamma(\mu_{2,i}, \nu_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \left[1 \quad a_i^* \quad b_i \quad p_i\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \left[1 \quad t_i\right]\boldsymbol{\gamma}$ $\log(\mu_{2,i}) = \left[1 \quad t_i\right]\boldsymbol{\delta}$
GLOG GA-2	$\psi_i f_{1,i} + (1-\psi_i) f_{1,i}$ $f_{1,i} = \log N(\mu_{1,i}, \sigma_{1,i})$ $f_{2,i} = \Gamma(\mu_{2,i}, \nu_{2,i})$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \left[1 \quad a_i^* \quad t_i \quad \mathbf{b}_{i,6} \quad \mathbf{p}_{i,6} \quad I(b_i^* \geq 14, p_i^* \geq 60)\right]\boldsymbol{\beta}$ $\log(\mu_{1,i}) = \left[1 \quad t_i\right]\boldsymbol{\gamma}$ $\log(\mu_{2,i}) = \left[1 \quad t_i\right]\boldsymbol{\delta}$

Tabell 7.9 - Definisjon av GLOGGA-modellene. Parameterne ν og σ estimeres uten direkte kobling mot forklaringsvariabler.

AIC-verdier for henholdsvis GLOGGA-0, GLOGGA-1 og GLOGGA-2 er 186 465, 186 310 og 186 277. AIC for log-gamma-modellene er dårligere enn for log-log-modellene, men langt bedre enn for de unimodale modellene.

7.8 Effekter av forklaringsvariablene på skadepris

Foruten modellenes prediksjoner, er det også av interesse å kunne si noe om hvilken effekt de ulike forklaringsvariablene har på skadeprisen. Jeg velger å se på estimatene til modell GIG-1 for å kunne si noe om dette. GIG-1 er en relativt godt tilpasset modell. Den er også unimodal, og har kun lineære forklaringsvariabler. Alt dette gjør parameterestimatene lettolkelige. Jeg vil se på parameterestimatene for både μ og ν. Under GAMLSS, som er modellrammeverket jeg bruker, avhenger begge disse parameterne direkte av forklaringsvariabler, på hver sin måte. Det er essensielt å se hvordan forklaringsvariablene påvirker begge fordelingsparameterne.

μ – koeffisient for	Estimat	Standardfeil	p-verdi	exp(estim)
1	9,8565	0,0559	0,0000	19081,566
a_i^*	-0,1215	0,0761	0,1101	0,8855592
t_i	0,0678	0,0140	0,0000	1,0701577
b_i	0,0014	0,0067	0,8346	1,001407
p_i	0,0024	0,0016	0,1384	1,0024269

Tabell 7.10 - Estimerte μ – koeffisienter for GIG-1. exp(estim) gir faktisk multiplikativ effekt på forventet gjennomsnittlig skadepris, μ.

Tabell 7.10 viser at årstall er den eneste klart signifikante forklaringsvariabelen for μ. Modellen GIG-1 har log-link. Følgelig er det relevant å se på kolonnen exp(estim) for å få den estimerte multiplikative effekten på forventet gjennomsnittlig skadepris for hver enkelt forklaringsvariabel. Estimatene gir grunnlag for å hevde at gjennomsnittlig skadepris øker med 7 % per år.

Personalder, p_i, er, med *p*-verdi 0,1384, ikke en signifikant forklaringsvariabel. Dersom man ser bort fra signifikansen og kun ser på estimert multiplikativ effekt, gir 1 års økning i personalder kun 0,2 % økning i skadeprisen, hvilket er neglisjerbart. Bilalder, b_i, gir enda lavere estimert effekt, og med en *p*-verdi på hele 0,8346 er det ikke noe statistisk grunnlag for å hevde at bilalderen påvirker skadeprisen. Antall skader fratrukket gjennomsnittlig antall skader, a_i^*, er kun med som en korrigering, for å sikre at de andre estimatene blir så riktige som mulig. Dersom bilalder og personalder ligger på porteføljens gjennomsnitt (5,9 år gammel bil og 50,3 år gammel kunde), predikerer GIG-1 at gjennomsnittlig skadepris vil være 19 082 i 2006.

ν – Koeffisient for	Estimat	Standardfeil	p-verdi	exp(estim)
1	-4,0605	0,0173	0,0000	0,0172398
a_i^*	-0,2388	0,0308	0,0000	0,7875952
t_i	-0,0033	0,0045	0,4708	0,9967513
b_i	-0,0031	0,0021	0,1400	0,9968879
p_i	-0,0036	0,0006	0,0000	0,9964125

Tabell 7. 11 - Estimerte ν – koeffisienter for GIG-1. exp(estim) gir faktisk multiplikativ effekt på parameteren ν.

Parameteren ν styrer formen på PDF-kurven til IG-fordelingen på en slik måte at høyere ν gir brattere topp, større skjevhet og "tyngre hale" (mer sannsynlighet for svært høye skadepriser). Generelt kan man si at usikkerheten øker med ν. Det fremgår av tabell 7.11 at personalder har meget signifikant effekt på usikkerhetsparameteren ν. Imidlertid er effekten liten i størrelse. Ved å se på kolonnen exp(estim) kan det, ettersom GIG-1 har log-link for ν, hevdes at verdien til usikkerhetsparameteren synker med 0,36 % når personalderen øker med 1 år. Dette er en relativt beskjeden effekt, selv om den er høyst statistisk signifikant. En mulig tolkning er at eldre bilførere i noe større grad har skader i liknende størrelsesorden, mens det for yngre bilførere er mer spredning i skadestørrelsene. Effekten for bilalder går i samme retning, men er med *p*-verdi 0,14 ikke statistisk signifikant. Effekten for årstall ser også ut til å gå i samme retning, men er meget tvilsom, ettersom absoluttverdien av estimatet er mindre enn standardfeilen, og *p*-verdien er på hele 0,4708. Det er for øvrig ikke overraskende at flere skader gir signifikant mindre variasjon i gjennomsnittlig skadepris.[23]

[23] Dette er et klassisk eksempel på det velkjente fenomenet *Regression To The Mean*. (se Hogg og Tanis 2010:564-565).

8 Modellering av total utbetaling

8.1 Generelt om total utbetaling

Etter å ha funnet det jeg håper er optimale modeller for skadefrekvens og skadepris, er tiden inne til å drøfte hvordan disse modellene kan brukes simultant. La U_i være alle utbetalinger (sum av skadeprisene) forsikringsselskapet har til kunden som innehar polise *i*. La A_i være antall skader for polise *i* og $S_{i,k}$ være skadeprisen for skade *k* på polise *i*. Forventet utbetaling er da gitt ved

$$\mathrm{E}(U_i) = \mathrm{E}\left(\sum_{k=1}^{A_i} S_{ik}\right) = \mathrm{E}\left\{\mathrm{E}\left\{\sum_{k=1}^{A_i} S_{ik} \,|\, A_i\right\}\right\} = \mathrm{E}\left(A_i \cdot (\bar{S}_i \,|\, A_i)\right) = \mathrm{E}(A_i \cdot (G_i \,|\, A_i)).$$

For å kunne regne ut forventningen riktig må det avklares hvorvidt, og eventuelt hvordan, skadefrekvensen og gjennomsnittlig skadepris henger sammen. Dersom disse størrelser er uavhengige, er forventningen gitt ved

$$\mathrm{E}(U_i) = \mathrm{E}\left(A_i\right)\mathrm{E}\left(G_i \,|\, A_i\right).$$

Jeg ønsker å teste hvorvidt skadefrekvens og snittskade er uavhengige størrelser. Et praktisk problem er at de fleste poliser har 0 skader. Disse har selvsagt ingen observasjon for skadepris. Når man ser på kun de poliser som har skader, og regner ut snittskaden for hver av disse, kan det aggregeres til tabell 8.1.

Antall skader på polise	Antall poliser	Gjennomsnittlig skadepris
0	54 198	0
1	8 557	15 240
2	393	13 785
3	15	12 125
4	2	1 697

Tabell 8.1 - Gjennomsnittlig skadepris for ulike antall skader per polise

Disse resultatene kan tyde på at økt antall skader for en polise senker gjennomsnittlig skadepris. Blant polisene med skader er estimert kovarians og korrelasjon gitt ved henholdsvis $\mathrm{Cov}(A,\bar{S}) = -79{,}7858$ og $\mathrm{Corr}(A,\bar{S}) = -0{,}01425725$. Disse resultatene tyder også på at gjennomsnittlig skadepris minker med antall skader. Imidlertid er dette marginaltester som ikke tar hensyn til andre forklaringsvariabler. En bedre måte å teste sammenhengen på, er å ta

vekk a_i^*, referansejustert antall skader, som forklaringsvariabel i en av modellene for skadepris, estimere parameterne, og vurdere forklaringskraften til a_i^* som et multivariabelt avhengighetsmål. Jeg definerer modellene GLOG-3 og GIG-3 som identiske med henholdsvis GLOG-2 og GIG-2, med den forskjell at a_i^* er tatt vekk som forklaringsvariabel for fordelingsparameterne. Etter å ha estimert parameterne til disse nye modellene ser jeg at GLOG-3 får AIC-verdi på 187 351 og at GIG-3 får AIC-verdi på 187 023. Dette er en såpass klar oppgang for begge modeller at det tyder på avhengighet mellom gjennomsnittlig skadepris og antall skader. Grunnen er at GLOG-3 og GIG-3 i prinsippet har G_i som reponsvariabel, i motsetning til $G_i|A_i$ som er responsvariabelen i de andre skadeprismodellene. Dersom G_i og $G_i|A_i$ har ulik sannsynlighetsfordeling, er det et klart tegn på avhengighet mellom G_i og A_i. Modellene GLOG-3 og GIG-3 brukes ikke videre i oppgaven. De er kun estimert for å teste ut uavhengighetshypotesen mellom skadepris og antall skader. Hvorvidt A_i og $G_i|A_i$ er avhengige størrelser, må vurderes ved å sammenlikne prediksjonskraften til modellene som forutsetter uavhengighet mellom disse størrelsene med modellene som predikerer *U* direkte.

8.2 Modeller gitt uavhengighet

Til tross for at testene i kapittel 8.1 gir grunn til å tro at det er avhengighet mellom skadepris og antall skader, vil jeg likevel teste ut modeller som forutsetter uavhengighet mellom disse størrelser. En fordel ved å anta uavhengighet er at man får fleksibilitet til å bruke optimale delmodeller for skadefrekvens og skadepris, og at koblingen er enkel og intuitiv. Fokuset vil være på forventning og varians av U_i. Disse størrelsene gir nok informasjon til å kunne drøfte risikoriktig prising ved hjelp av standardavviksprinsippet (se delkapittel 3.9). Følgende modellformulering er felles for disse uavhengighetsmodellene:

$$U_i = (A_i)(G_i|A_i) \Rightarrow U_i \in [0,\infty)$$
$$\mathrm{E}(U_i) = \mathrm{E}(A_i) \cdot \mathrm{E}(G_i|A_i)$$
$$\mathrm{Var}(U_i) = \mathrm{Var}(A_i)\,\mathrm{Var}(G_i|A_i) + \left(\mathrm{E}(A_i)\right)^2 \mathrm{Var}(G_i|A_i) + \left(\mathrm{E}(G_i|A_i)\right)^2 \mathrm{Var}(A_i)$$

Det kan vises at uttrykket for variansen er riktig, gitt uavhengighet mellom A_i og $G_i|A_i$.[24] Aktuelle modeller under uavhengighetshypotesen er 3 modeller for skadefrekvens og 12 for skadepris. Jeg kombinerer kun modeller med tilsvarende fleksibilitetsnivå. Det vil for eksempel si at APOI-0 kobles mot GLOG-0, men ikke mot for eksempel GLOG-1. Denne regelen gir 12 kombinasjoner, og følgelig 4 AIC-minimerte modeller under uavhengighetshypotesen. Tabell 8.2 gir en skjematisk oversikt over de 12 ulike modellene for utbetaling, gitt uavhengighet.

Delmodell for skadefrekvens	**Delmodell for skadepris**	**Navn på ny modell**	**Frihetsgrader i ny modell**
APOI-0	GLOG-0	UPOILOG-0	6
APOI-1	GLOG-1	UPOILOG-1	15
APOI-2	GLOG-2	UPOILOG-2	39
APOI-0	GIG-0	UPOIIG-0	6
APOI-1	GIG-1	UPOIIG-1	15
APOI-2	GIG-2	UPOIIG-2	31
APOI-0	GLOGLOG-0	UPOILOGLOG-0	8
APOI-1	GLOGLOG-1	UPOILOGLOG-1	15
APOI-2	GLOGLOG-2	UPOILOGLOG-2	35
APOI-0	GLOGGA-0	UPOILOGGA-0	8
APOI-1	GLOGGA-1	UPOILOGGA-1	15
APOI-2	GLOGGA-2	UPOILOGGA-2	36

Tabell 8.2 - Skjematisk oversikt over uavhengighetsmodellen for U

[24] Variansen til et produkt av to uavhengige stokastiske variabler er gitt som over. Her viser jeg en enkel utledning av denne formelen, gjort generelt for uavhengige stokastiske variabler X og Y

$$\mathrm{Var}(XY) = \mathrm{E}\left((XY)^2\right) - \left(\mathrm{E}(X)\,\mathrm{E}(Y)\right)^2 = \mathrm{E}(X^2)\,\mathrm{E}(Y^2) - \left(\mathrm{E}(X)\right)^2\left(\mathrm{E}(Y)\right)^2$$
$$= (\mathrm{Var}(X) + \left(\mathrm{E}(X)\right)^2)(\mathrm{Var}(Y) + \left(\mathrm{E}(Y)\right)^2 - \left(\mathrm{E}(X)\right)^2\left(\mathrm{E}(Y)\right)^2$$
$$= \mathrm{Var}(X)\,\mathrm{Var}(Y) + \left(\mathrm{E}(Y)\right)^2\,\mathrm{Var}(X) + \left(\mathrm{E}(X)\right)^2\,\mathrm{Var}(Y)$$

8.3 Modellering av utbetaling direkte ved ZAIG og ZAGA

8.3.1 Generelt om ZAIG/ZAGA-modellene

Modellene under punkt 8.2 er lette å tolke og delmodellene er skreddersydd for henholdsvis skadefrekvens og skadestørrelse. Uavhengighetsmodellene har imidlertid minst 2 svakheter:

- De er basert på en uavhengighetsantagelse som er svært tvilsom.
- Mange av modellene har et høyt antall parametere. Dette kommer av at de er skreddersydd gjennom delmodellene. Det kan føre til overparameterisering, altså at modellene blir for komplekse. Dette øker risikoen for at parameterne beskriver stokastisk støy i stedet for reelle effekter, hvilket går ut over prediksjonskraften.

Jeg vil i dette delkapitlet tilpasse 2 modeller som unngår disse problemene ved å modellere total skadepris per polise direkte. Responsvariabel er da U_i. Det er rimelig at disse modeller må være av typen FM, og at de ideelt sett bør være svært skeive. ZAIG-fordelingen, og i noe mindre grad ZAGA-fordelingen, oppfyller disse kravene. Jeg velger derfor disse som responsfordelinger for de direkte modellene.

Enten den stokastiske variabelen U_i ses som ZAIG-fordelt eller ZAGA-fordelt, har den, ved realisering, sannsynlighet ψ_i for å bli 0, og sannsynlighet $(1-\psi_i)$ for å få følge henholdsvis fordelingen $\text{IG}(\mu_i,\nu_i)$, eller $\Gamma(\mu_i,\nu_i)$. Følgelig bør man se på ψ_i som sannsynligheten for at polise *i* blir skadefri ved realisering. Videre bør man se på μ_i og ν_i som de tilsvarende parameterne i IG-fordelingen eller gammafordelingen. Jeg velger derfor log-link for disse parameterne, som jeg også gjorde i skadeprismodellene der IG- og gammafordelingen ble definert.

Modellene i kapittel 7 brukte $G_i|A_i$ som responsvariabel, slik at antall skader kunne brukes som forklaringsvariabel i modellene. Imidlertid kan den totale skadeutbetalingen U_i ikke justeres for antall skader ved å bruke a_i^* som forklaringsvariabel, ettersom denne ikke estimeres separat i ZAIG/ZAGA-modellene. Dersom man ønsker en modell som skulle beskrive data best mulig, vil man helt sikkert få en bedre tilpasset modell ved å inkorporere

antall skader som forklaringsvariabel. Imidlertid bygger jeg først og fremst modeller for å predikere fremover i tid, hvilket betyr at a_i^* ikke kan brukes som forklaringsvariabel her.

Observasjonene, U_i, ligger på polisenivå i denne modellen. Følgelig er det behov for å korrigere for eksponering. Jeg velger kun å korrigere sannsynlighetsparameteren ψ_i for eksponering, ettersom det er denne parameteren eksponeringen vil ha overveldende betydning for. Størrelsen $(1-\psi_i)$ er sannsynligheten for at polise *i* har minst 1 skade. En variant av hypotesen H_0 fra delkapittel 5.5.2, er å anta at denne sannsynligheten er proporsjonal med eksponeringen. La nå λ_i være helårseksponert utgave av ψ_i og definer den nye nullhypotesen H_0^* ved $H_0^*:(1-\psi_i)=r_i\cdot(1-\lambda_i)$. Parameteren ψ_i er en sannsynlighet, og kobles derfor til forklaringsvariabler ved logit-link. I fulleksponert utgave (der man bruker λ_i i stedet for ψ_i), skrives koblingen som $\log\left(\frac{\lambda_i}{1-\lambda_i}\right)=\mathbf{w}_i^T\boldsymbol{\delta}$. Under H_0^* er dette ekvivalent med

$$\log\left(\frac{\psi_i+r_i-1}{1-\psi_i}\right)=\mathbf{w}_i^T\boldsymbol{\delta}. \tag{7}$$

Heller et al. (2006:4) bruker tilsvarende metodikk for å korrigere for eksponering i sin ZAIG-modell. Linkfunksjonen gitt ved (7) har jeg selv programmert i R. Jeg velger å kalle denne for *Eksponeringsjustert Alternativ Logit-Link* eller bare EAL-link.[25] En enkelt test for hvorvidt H_0^* er sann, utføres ved å estimere parameterne i ZAIG-modellen, først ved å bruke (7) (jeg kaller dette alternativ 1), så ved å la log-transformasjonen av eksponeringen inngå som forklaringsvariabel (jeg kaller dette alternativ 2). Resultatet av testen er gitt i tabell 8.3.

[25] *Alternativ* refererer til at det er $(1-\psi_i)$ som korrigeres for eksponering, heller enn ψ_i.

Alt.	Link-funksjon	Invers link-funksjon	Estimat for ψ ved $r_i = 1$	AIC
1	$\log\left(\frac{\psi_i + r_i - 1}{1 - \psi_i}\right) = \beta_0$	$\psi_i = 1 - \frac{r_i}{1 + e^{\beta_0}}$	0,7585182	239 777
2	$\log\left(\frac{\psi_i}{1 - \psi_i}\right) = \beta_0 + \beta_1 \log(r_i)$	$\psi_i = \frac{r_i^{\beta_1} e^{\beta_0}}{1 + r_i^{\beta_1} e^{\beta_0}}$	0,8283104	238 514

Tabell 8.3 - Testing av 2 alternative måter å inkorporere eksponering på, i ZAIG-modellene

Tabell 8.3 viser at alternativ 2, der log-transformert eksponering inkluderes som forklaringsvariabel, er å foretrekke, ettersom AIC-nedgangen er stor fra alternativ 1. Jeg konkluderer med at nullhypotesen H_0^* er feil, og bruker derfor metodikken fra alternativ 2 i den videre modelleringen, både for ZAGA-modeller og ZAIG-modeller.

8.4 ZAIG-modell for total skadepris

Jeg definerer nullmodellen UZAIG-0, der jeg bruker log-transformert eksponering men ingen andre forklaringsvariabler, og mellommodellen UZAIG-1, der jeg bruker lineære forklaringsvariabler. Etter å ha kjørt AIC-minimerings-algoritmen, kan jeg også definere UZAIG-2, som er den AIC-minimerte utgaven av ZAIG-modellen.

Modell	$U_i \sim$	**Kobling til forklaringsvariablene**
UZAIG-0	$\text{ZAIG}(\psi_i, \mu_i, \nu_i)$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i)\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \gamma_0$ $\log(\nu_i) = \delta_0$
UZAIG-1	$\text{ZAIG}(\psi_i, \mu_i, \nu_i)$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i) & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \begin{bmatrix}1 & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\gamma}$ $\log(\nu_i) = \begin{bmatrix}1 & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\delta}$
UZAIG-2	$\text{ZAIG}(\psi_i, \mu_i, \nu_i)$	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i) & \mathbf{b}_{i,6} & \mathbf{p}_{i,4} & t_i \cdot b_i & I(b_i^* \geq 14, p_i^* \leq 30)\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \begin{bmatrix}1 & t_i & t_i \cdot b_i & b_i \cdot p_i & I(b_i^* \leq 5, p_i^* \geq 60)\end{bmatrix}\boldsymbol{\gamma}$ $\log(\nu_i) = \begin{bmatrix}1 & \mathbf{b}_{i,5} & \mathbf{p}_{i,2} & t_i \cdot b_i & I(b_i^* \geq 14, p_i^* \leq 30)\end{bmatrix}\boldsymbol{\delta}$

Tabell 8.4 - Definisjon av UZAIG-modellene

AIC-verdier for UZAIG-0, UZAIG-1 og UZAIG-2 er henholdsvis 238 514, 238 342 og 238 211. Disse AIC-verdier kan ikke sammenliknes med AIC-verdier for modeller for skadefrekvens eller skadepris, ettersom responsvariabelen er *U*, og ikke *A* eller *G*. Modellen UZAIG-2 har hele 29 estimerte parametere, og er således ekstremt kompleks. Fleksibiliteten i UZAIG-2 gir utslag i veldig klar AIC-nedgang fra UZAIG-0 og UZAIG-1. Det gjenstår imidlertid å se om denne fleksibiliteten bidrar til bedre prediksjoner.

8.5 ZAGA-modell for total skadepris

Jeg definerer nullmodellen UZAGA-0, der kun log-transformert eksponering brukes som forklaringsvariabel, mellommodellen UZAGA-1 der kun selvstendige, lineære forklaringsvariabler brukes, og UZAGA-2 som er den AIC-minimerte modellen, i tabell 8.5.

Modell	$U_i \sim$	**Kobling til forklaringsvariablene**
UZAGA-0	ZAGA(ψ_i, μ_i, ν_i)	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i)\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \gamma_0$ $\log(\nu_i) = \delta_0$
UZAGA-1	ZAGA(ψ_i, μ_i, ν_i)	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i) & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \begin{bmatrix}1 & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\gamma}$ $\log(\nu_i) = \begin{bmatrix}1 & t_i & b_i & p_i\end{bmatrix}\boldsymbol{\delta}$
UZAGA-2	ZAGA(ψ_i, μ_i, ν_i)	$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \begin{bmatrix}1 & \log(r_i) & \mathbf{b}_{i,6} & \mathbf{p}_{i,4} & t_i \cdot b_i & I(b_i^* \geq 14, p_i^* \leq 30)\end{bmatrix}\boldsymbol{\beta}$ $\log(\mu_i) = \begin{bmatrix}1 & t_i & \mathbf{b}_{i,3} & \mathbf{p}_{i,4} & t_i \cdot b_i & b_i \cdot p_i & I(b_i^* \leq 5, p_i^* \leq 30)\end{bmatrix}\boldsymbol{\gamma}$ $\log(\nu_i) = \begin{bmatrix}1 & t_i & b_i & \mathbf{p}_{i,2} & b_i \cdot p_i\end{bmatrix}\boldsymbol{\delta}_1$ $+\begin{bmatrix}I(b_i^* \geq 14, p_i^* \leq 30) & I(b_i^* \geq 14, p_i^* \geq 60)\end{bmatrix}\boldsymbol{\delta}_2$

Tabell 8.5 - Definisjon av UZAGA-modellene

AIC-verdier for UZAGA-0, UZAGA-1 og UZAGA-2 er henholdsvis 240 678, 240 432 og 240 282. Det er en klar nedgang ved innføring av mer fleksibilitet i modellen. Imidlertid er AIC-verdiene for UZAIG-modellene betraktelig lavere enn for UZAGA-modellene. Det tilsier at ZAIG-fordelingen beskriver observasjonene bedre.

8.6 Effekter av forklaringsvariablene på total skadepris

Jeg vil se på de estimerte effektene av forklaringsvariablene på total skadepris, modellert direkte. Det er interessant å sammenlikne disse estimerte effektene med de tilsvarende estimerte effektene på skadefrekvens og skadepris, modellert separat. Jeg bruker UZAIG-1 til å se på effektene, ettersom UZAIG-modellene har lavere AIC-verdier enn UZAGA-modellene, og ettersom UZAIG-1 har lineære forklaringsvariabler. Jeg har i delkapitlene 6.3 og 7.8 sett på effektene av forklaringsvariablene på skadefrekvens og skadepris, ved hjelp av modellene APOI-1 og GIG-1. Den teoretiske forskjellen mellom kombinasjonen (APOI-1, GIG-1) og UZAIG-1, er ikke veldig stor. APOI-1 estimerer *skadefrekvens*, mens parameteren ψ i UZAIG-1 estimerer sannsynligheten for at en polise er skadefri. Følgelig kan man enkelt finne estimert *skadesannsynlighet* ved å regne ut $(1-\psi)$. Ettersom kun 4,6 % av polisene med skader har flere enn 1 skade, vil forskjellen mellom frekvens og sannsynlighet neppe være veldig stor. Den andre sentrale teoretiske forskjellen er at (APOI-1, GIG-1)-kombinasjonen estimerer parameterne for skadefrekvens og skadepris hver for seg, mens de tilsvarende parameterne i UZAIG-1 er estimert simultant.

ψ **–koeffisient for**	**Estimat**	**Standardfeil**	***p*-verdi**
1	1,5835	0,0274	0,0000
$\log(r_i)$	-0,5902	0,0209	0,0000
t_i	0,0051	0,0069	0,4580
b_i	0,0147	0,0031	0,0000
p_i	0,0090	0,0009	0,0000

Tabell 8.6 - Estimerte ψ –koeffisienter for UZAIG-1. Merk at ψ har logit-link, og sannsynligheten for skade er $(1-\psi)$. Det betyr at positivt estimat minsker skadesannsynligheten ved økning av verdien av forklaringsvariabelen.

Tabell 8.6 viser estimerte effekter for sannsynlighetsparameteren ψ i UZAIG-1. Bilalder og personalder er begge høyst signifikante som forklaringsvariabler her. Estimatet er positivt for begge disse forklaringsvariablene, og en god del større for bilalder enn for personalder. Det betyr at høyere bilalder og høyere personalder gir høyere verdi av ψ, og dermed lavere skadesannsynlighet. Dette stemmer godt overens med effektene som ble observert i delkapittel 6.3, ettersom høyere bilalder og høyere personalder, ifølge modell APOI-1, gir lavere skadefrekvens.

Størrelsen på effektene, slik de fremstår i tabellene 6.4 og 8.6 er ikke direkte sammenlignbare, ettersom ψ i UZAIG-1, sannsynligheten for 0 skader, er modellert med logit-link, og μ i APOI-1, forventet skadefrekvens, er modellert med log-link. Det er imidlertid felles for begge modeller at bilalder gir større effekt enn personalder, og at årstall ikke gir signifikant effekt.

μ – **koeffisient for**	**Estimat**	**Standardfeil**	***p*-verdi**	**exp(estim)**
1	9,8995	0,0573	0,0000	19920,667
t_i	0,0681	0,0144	0,0000	1,070498
b_i	0,0016	0,0069	0,8214	1,0015632
p_i	0,0022	0,0017	0,1957	1,0021653

Tabell 8.7 - Estimerte μ – koeffisienter for UZAIG-1. Modellen har log-link for μ. Derfor representerer exp(estim) den faktiske multiplikative effekten hver forklaringsvariabel har på μ.

Av tabell 8.7 fremgår det at årstall har høyst signifikant effekt på forventet skadepris gitt minst 1 skade, μ. Den multiplikative effekten er estimert til ca. 7 % årlig økning per år. Dette stemmer meget godt overens med tilsvarende estimert koeffisient i GIG-1 (se delkapittel 7.8). Forventet totalt skadebeløp, gitt minst 1 skade, i 2006, for en kunde med gjennomsnittlig bilalder og personalder (5,9 år gammel bil og 50,3 år gammel kunde), er 19 921. Dette er noe høyere enn tilsvarende estimat i GIG-1. Det skyldes at GIG-1 predikerer *gjennomsnittlig* skadebeløp, gitt minst 1 skade, mens μ i UZAIG-1 predikerer *totalt* skadebeløp, gitt minst 1 skade. Sagt på en annen måte er ikke UZAIG-1 i stand til å korrigere for multiple skader på en polise. Dette skyldes at ZAIG-fordelingens ψ – parameter er en sannsynlighet, ikke en frekvens som μ i Poissonfordelingen. Bilalder og personalder har ikke signifikant effekt på μ, hverken i estimatene til GIG-1 eller UZAIG-1.

ν – **koeffisient for**	**Estimat**	**Standardfeil**	***p*-verdi**	**exp(estim)**
1	-4,0645	0,0173	0,0000	0,0171722
t_i	-0,0038	0,0045	0,3921	0,9961604
b_i	-0,0035	0,0021	0,0980	0,996528
p_i	-0,0034	0,0006	0,0000	0,9966167

Tabell 8.8 - Estimerte ν – koeffisienter for UZAIG-1. Modellen har log-link for ν. Derfor representerer exp(estim) den faktiske multiplikative effekten hver forklaringsvariabel har på ν.

Som det vises i delkapittel 7.8, i estimatene til GIG-1, har kun personalder signifikant effekt på skjevhets-parameteren ν. Effekten er svært liten, men tilsier at for hvert år eldre en bilfører blir, vil forventet verdi av ν synke med 0,34 %. De andre forklaringsvariablene er ikke statistisk signifikante. Imidlertid tilsier punktestimatene for både årstall og bilalder at en økning i en av disse gir marginalt lavere verdi av ν.

Generelt ser man at estimerte effekter ved UZAIG-1 er svært like tilsvarende estimerte effekter i modellene APOI-1 og GIG-1. Dette styrker troen på at effektene er reelle, samtidig som det kan antyde at forskjellen mellom simultan estimering av skadesannsynlighet og skadepris, og separat estimering av skadefrekvens og skadepris, ikke er så stor.

9 Testing av modellene for *U*

9.1 Testmetodikk

Den overordnede målsetningen i denne oppgaven er å finne en modell som best mulig beskriver total skadeutbetaling, *U*, på hver enkelt polise. Spesifikt ønsker jeg å kunne ta i bruk forventning og varians i prisfastsettelsen av bilforsikringen for kunde *i*, i år *j*. Fokus er derfor på at størrelsene $\mathrm{E}(U_{i,j})$ og $\mathrm{Var}(U_{i,j})$ skal være estimert best mulig. Jeg vil bruke kryssvalidering for å teste de 18 modellene for $U_{i,j}$ opp mot hverandre. Kryssvalidering skal her forstås på følgende måte: Parameterne estimeres for hver enkelt modell ved å bruke alle data fra alle årene i datagrunnlaget, utenom år *j*. De estimerte parameterne brukes så til å beregne $\mathrm{E}(U_{i,j})$ og $\mathrm{Var}(U_{i,j})$ for nettopp år *j*. Disse størrelsene er dermed å regne for genuine *prediksjoner*. Dette gir totalt 63 165 sett med estimater av forventning og varians, som tilsvarer 1 sett med estimater for hver observert polise i datasettet. Jeg forkaster 65 tilfeldig valgte observasjoner, og deler de resterende 63 100 inn i 631 grupper, med 100 observasjoner i hver, etter et helt tilfeldig mønster[26]. La observasjon *k* i gruppe *l* være $U_{k,l}$. For hver gruppe, *l*, regner jeg ut verdien Z_l, for hver enkelt modell. Den er gitt ved

$$Z_l = \frac{\sum_{k=1}^{100} U_{k,l} - \sum_{k=1}^{100} \mathrm{E}(U_{k,l})}{\sqrt{\sum_{k=1}^{100} \mathrm{Var}(U_{k,l})}}.$$

Under sentralgrenseteoremet (se delkapittel 3.7) vil Z_l konvergere mot fordeling $N(0,1)$, forutsatt at modellen er sann. Dette gir mulighet til å utføre en statistisk test der modellene kan settes opp mot hverandre. Z_l – verdiene må forstås som standardiserte aggregerte residualer. Dersom alle modellene hadde hatt normalfordelingen som responsfordeling, ville en bedre test vært å se direkte på de standardiserte residualer $\frac{U_i - \mathrm{E}(U_i)}{\sqrt{\mathrm{Var}(U_i)}}$, og sammenlikne disse. Imidlertid er det ingen grunn til å tro at de enkelte residualene fra modellene i denne oppgaven er normalfordelt. Ved å aggregere residualene derimot, får man størrelser som kan

[26] Jeg bruker "pseudo-random-number-generator" til å generere et trekk $z^*_{j,k}$ av standardnormalfordelingen for hver observasjon. Jeg sorterer så observasjonene etter størrelsen på $z^*_{j,k}$ og plasserer de 100 første observasjonene i den sorterte listen i gruppe 1, de 100 neste i gruppe 2, osv.

testes mot normalfordelingen. Det er 631 Z_l – observasjoner for hver modell. Dersom en modell er sann, vil dens Z_l tendere mot en standardnormalfordeling. Dette undersøker jeg ved å lage QQ-plot.

9.1.1 QQ-plot for Z-verdiene

For hver modell sorteres *Z*-verdiene i stigende rekkefølge. La *Z*-verdi nr i, i den sorterte listen være $Z_{(i)}$. Jeg definerer videre en enkel empirisk fordelingsfunksjon, som gir empiriske "sannsynligheter" ψ_i som $\psi_i(i) = \frac{i}{631}$. Punktene i QQ-plottet er da gitt ved koordinater $\left(\Phi^{-1}(\psi_i)\ ,\ Z_{(i)}\right)$. Jo nærmere disse punktene er linjen $y = x$ i koordinatsystemet, jo større grunn har man til å tro at Z er standardnormalfordelt.

9.1.2 Årstabeller

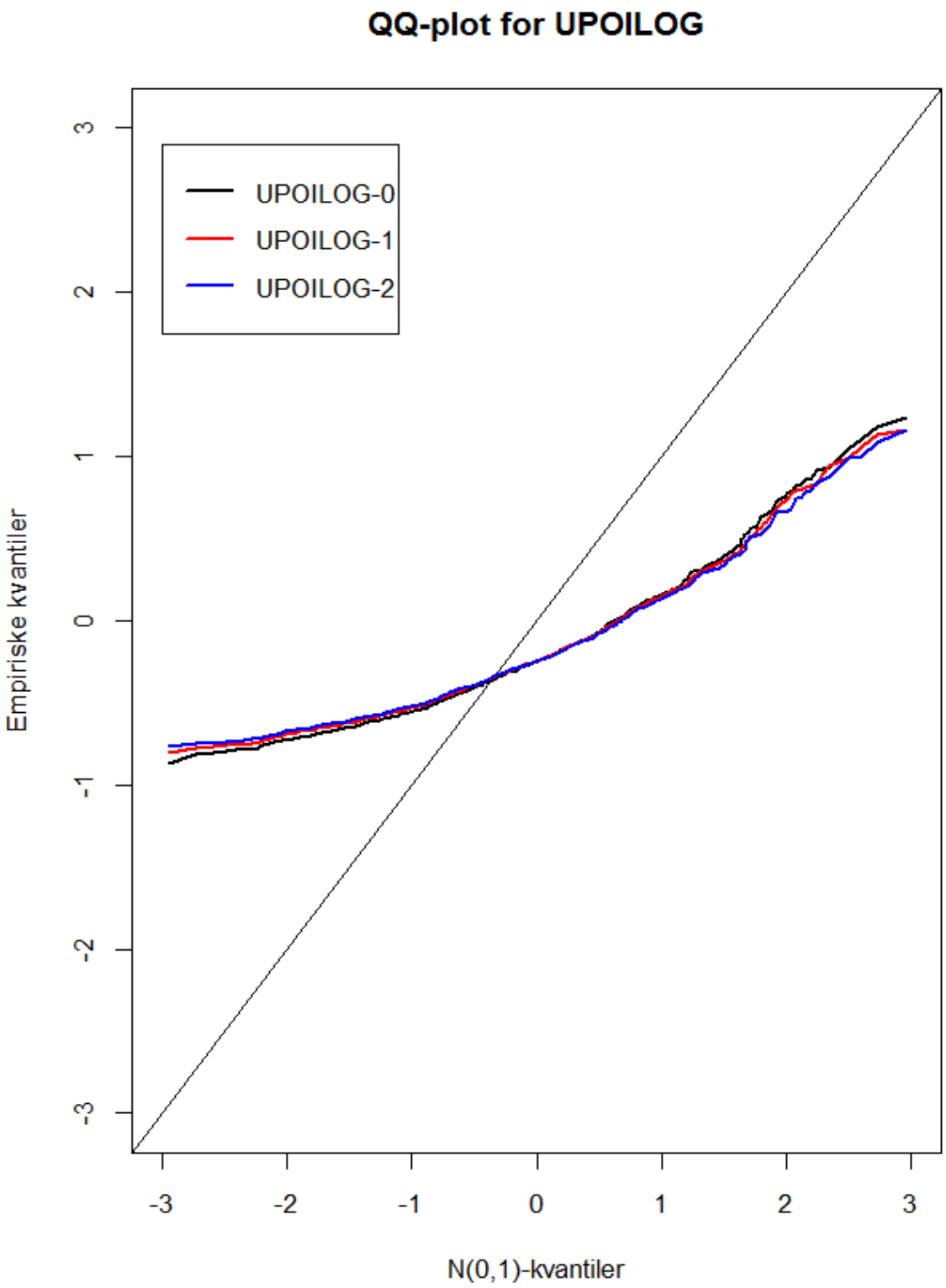
Kryssvalideringen, med sine predikerte verdier $\mathrm{E}(U_{i,j})$ og $\mathrm{Var}(U_{i,j})$, brukes primært til å produsere *Z*-verdier som analyseres ved QQ-plot, som beskrevet over. En sekundær, mer pragmatisk metodikk, er å bruke prediksjonene til å regne ut noen nøkkeltall. Disse regnes ut separat for hvert år, og presenteres i *årstabeller*. Årstabellene kan ses som "regnskap" for hvor godt prediksjonene traff i hver årgang. Alle størrelser som inngår i tabellene er *gjennomsnittstall*, i den forstand at de er en sum dividert med ρ_j som representerer total eksponering i år *j*. Det vil si at $\rho_j = \sum_{i=1}^{n_j} r_{i,j}$. Altså er ρ_j eksponeringsjustert antall observasjoner. Fordelen med å dividere med ρ_j heller enn antall rene observasjoner n_j, er at størrelsene blir gjennomsnittlige for hele poliseår. Følgende størrelser inngår i årstabellene (*j* er årstall):

- **Kvadrert avvik** er definert som $\frac{1}{\rho_j}\sum_{i=1}^{n_j}\left(U_{i,j} - \mathrm{E}(U_{i,j})\right)^2$. Jo lavere dette tallet er, jo bedre treffer prediksjonene for $E(U_{i,j})$.

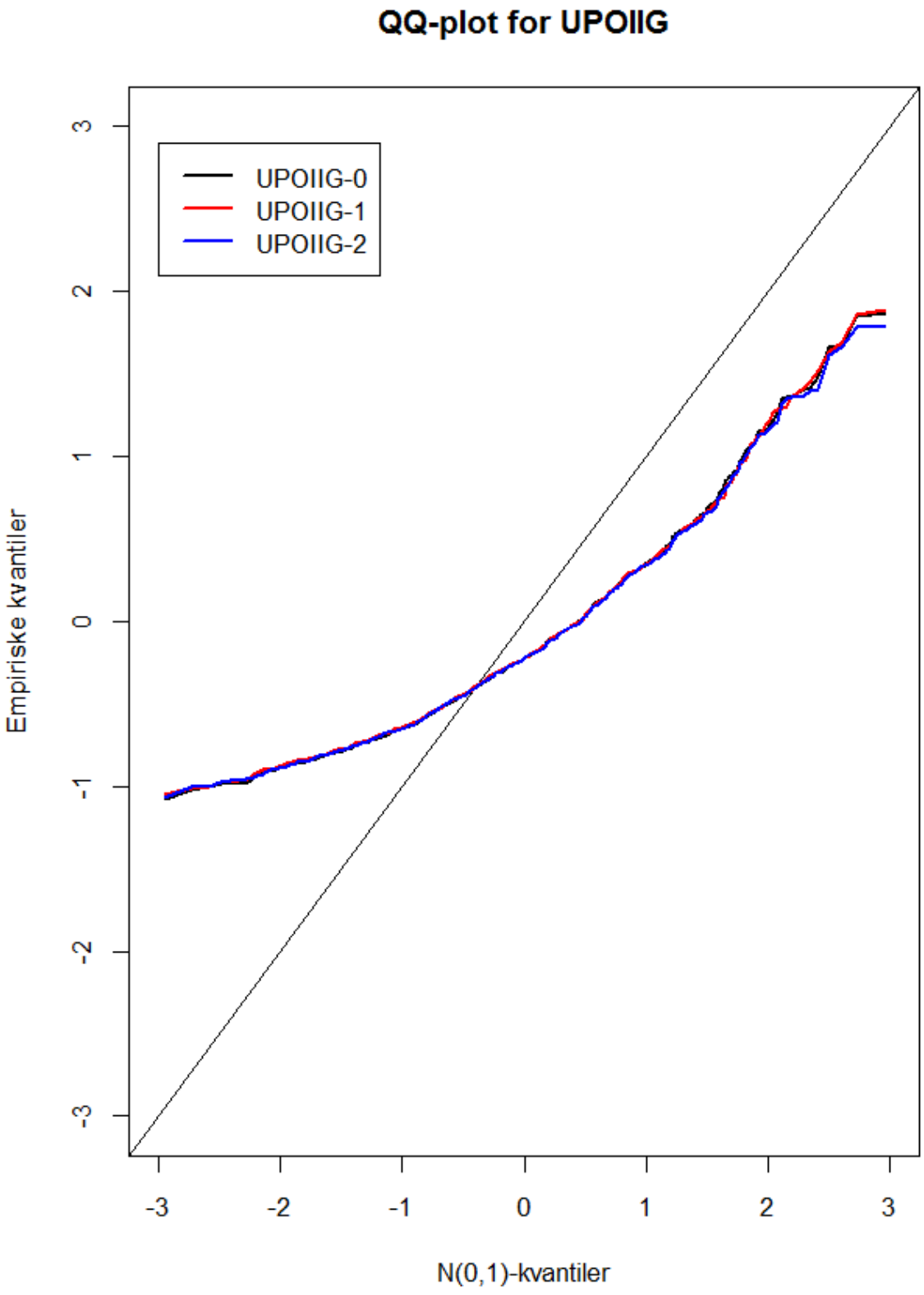
- **Absolutt avvik** er definert som $\frac{1}{\rho_j}\sum_{i=1}^{n_j}\left|U_{i,j} - \mathrm{E}(U_{i,j})\right|$. Denne størrelsen er nært beslektet med kvadrert avvik, men er mer robust, i betydningen at den ikke lar seg dominere like lett av ekstreme observasjoner.
- **Aggregert avvik** er gitt ved $\frac{1}{\rho_j}\sum_{i=1}^{n_j}\left(\mathrm{E}(U_{i,j}) - U_{i,k}\right)$, og kan ses som et regnskap for hvor mange kr for mye eller for lite modellen predikerte i snittutbetalinger per poliseår.
- **Rot av forventet varians** er gitt ved $\frac{1}{\rho_j}\sqrt{\sum_{i=1}^{n_j}\mathrm{Var}(U_{i,j})}$.

Jeg vil gjøre en totalvurdering av årstabeller opp mot QQ-plottene for å avgjøre hvilken eller hvilke modeller som anbefales.

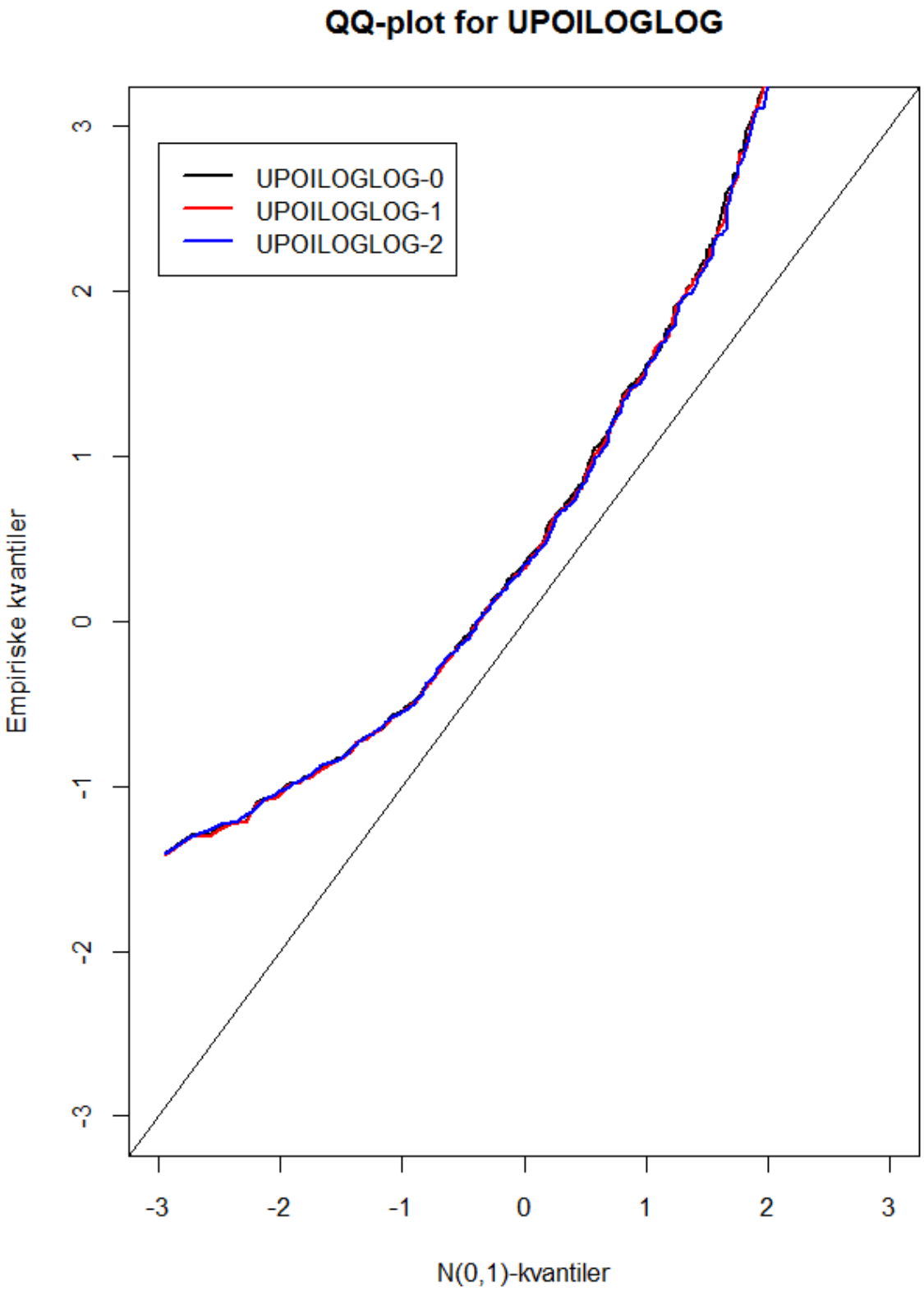
9.2 Resultater



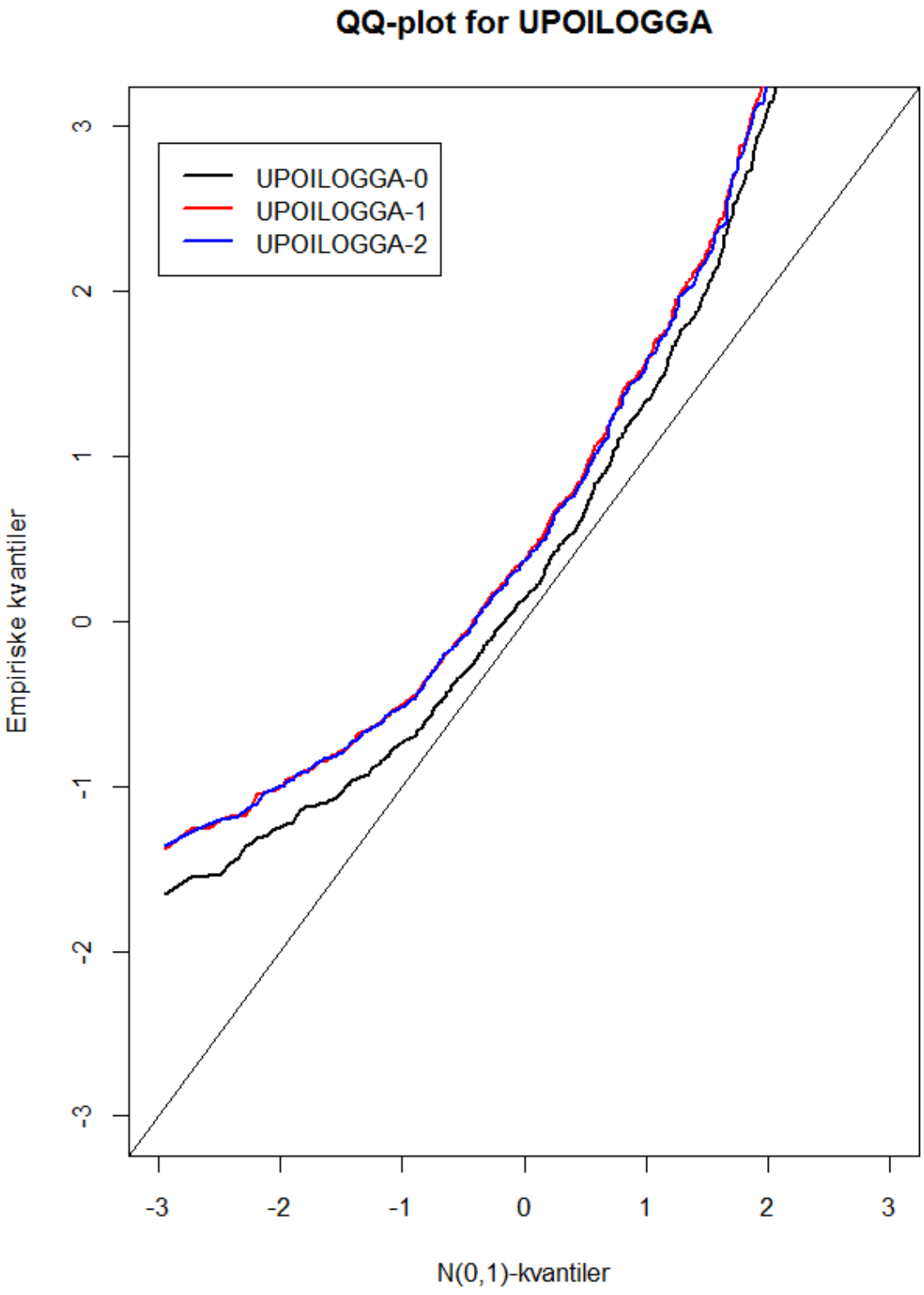
Figur 9.1 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UPOILOG-modellene



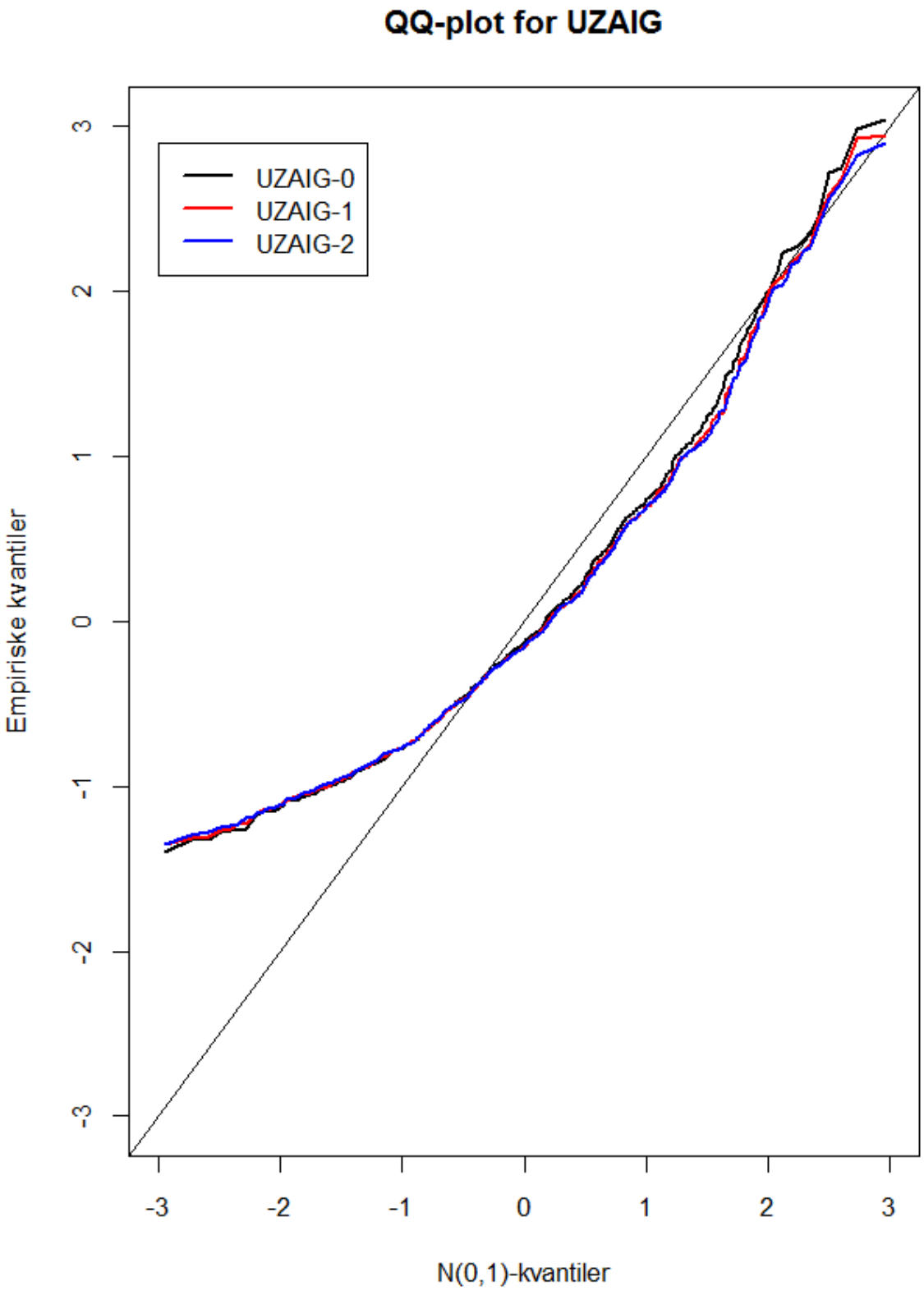
Figur 9.2 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UPOIIG-modellene



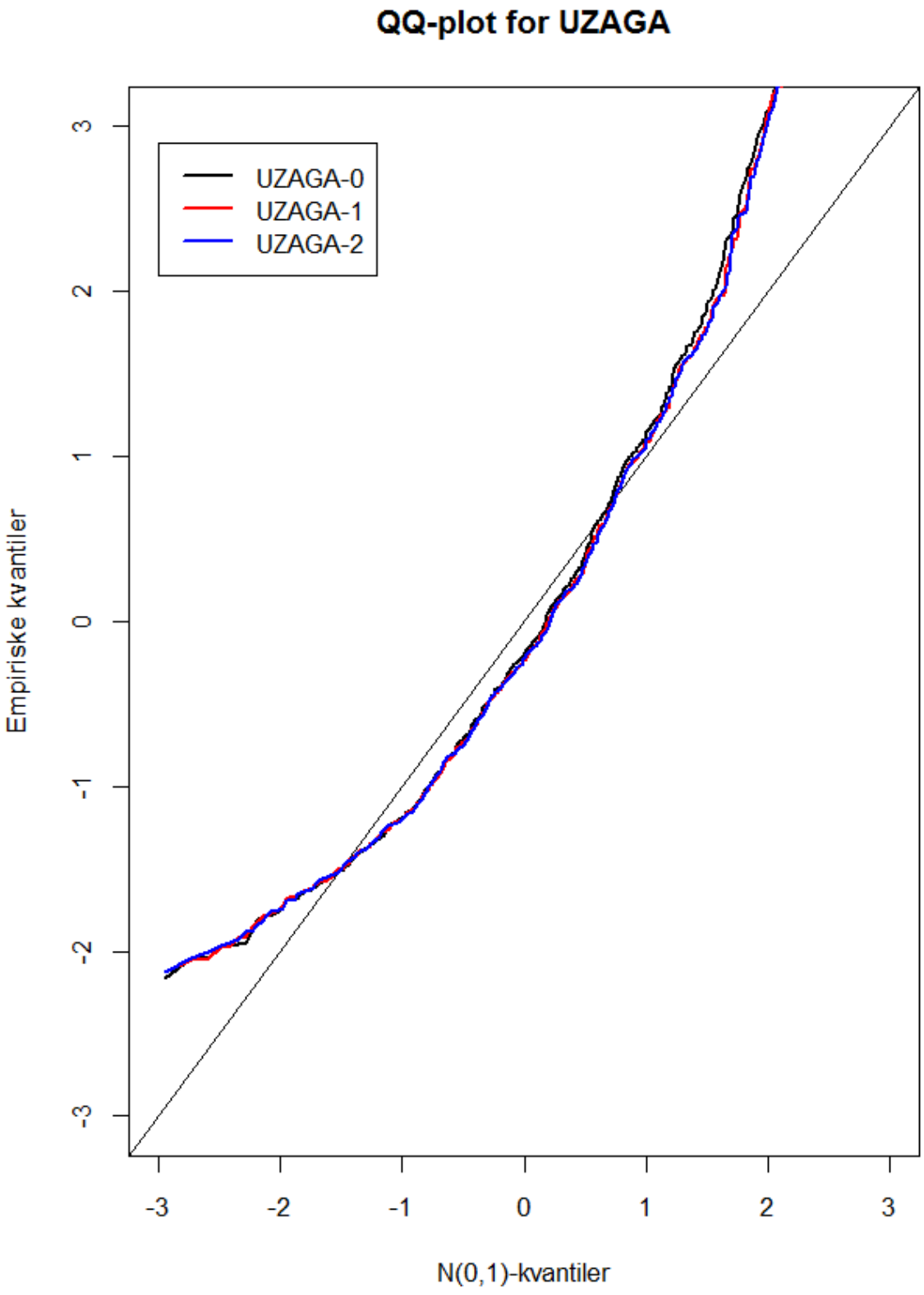
Figur 9.3 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UPOILOGLOG-modellene



Figur 9.4 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UPOILOGGA-modellene



Figur 9.5 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UZAIG-modellene



Figur 9.6 - QQ-plot av Z-verdier mot standardnormalfordelingen for alle varianter av UZAIG-modellene

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2000	UPOIIG-0	70 441 217	5 019	1 277	314
2000	UPOIIG-1	69 780 993	4 741	930	268
2000	UPOIIG-2	70 400 004	5 052	1 334	320
2000	UPOILOG-0	70 895 495	5 184	1 484	402
2000	UPOILOG-1	70 045 528	4 835	1 043	323
2000	UPOILOG-2	69 757 247	4 774	986	327
2000	UPOILOGGA-0	69 126 971	4 003	-30	157
2000	UPOILOGGA-1	69 185 403	3 863	-215	134
2000	UPOILOGGA-2	69 010 553	3 918	-126	134
2000	UPOILOGLOG-0	69 144 941	4 131	138	157
2000	UPOILOGLOG-1	69 160 403	3 898	-168	134
2000	UPOILOGLOG-2	68 969 446	3 909	-135	134
2000	UZAGA-0	69 661 470	4 660	820	136
2000	UZAGA-1	69 308 904	4 435	536	119
2000	UZAGA-2	68 994 046	4 368	466	118
2000	UZAIG-0	69 661 470	4 660	820	212
2000	UZAIG-1	69 312 754	4 438	540	182
2000	UZAIG-2	69 114 843	4 357	444	175

Tabell 9.1 - Årstabell 2000 – Grønn er lavest og rød er høyest (i absoluttverdi). Dette gjelder også for de andre årstabellene.

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2001	UPOIIG-0	162 984 233	5 548	527	292
2001	UPOIIG-1	162 679 767	5 141	10	229
2001	UPOIIG-2	162 877 798	5 561	558	296
2001	UPOILOG-0	163 329 331	5 779	816	383
2001	UPOILOG-1	162 908 226	5 365	295	299
2001	UPOILOG-2	162 810 418	5 335	270	304
2001	UPOILOGGA-0	163 023 938	4 563	-743	140
2001	UPOILOGGA-1	163 424 217	4 335	-1 043	119
2001	UPOILOGGA-2	163 228 707	4 391	-956	120
2001	UPOILOGLOG-0	162 908 175	4 659	-617	140
2001	UPOILOGLOG-1	163 225 852	4 438	-906	119
2001	UPOILOGLOG-2	163 037 001	4 466	-856	120
2001	UZAGA-0	162 670 664	5 199	82	125
2001	UZAGA-1	162 677 774	4 900	-301	105
2001	UZAGA-2	162 577 644	4 872	-325	104
2001	UZAIG-0	162 670 664	5 199	82	193
2001	UZAIG-1	162 680 229	4 897	-306	157
2001	UZAIG-2	162 543 182	4 873	-324	156

Tabell 9.2 - Årstabell 2001

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2002	UPOIIG-0	189 417 371	5 975	399	299
2002	UPOIIG-1	189 261 715	5 814	199	275
2002	UPOIIG-2	189 389 369	6 005	446	306
2002	UPOILOG-0	189 764 885	6 225	713	395
2002	UPOILOG-1	189 663 944	6 060	505	364
2002	UPOILOG-2	189 818 194	6 089	548	378
2002	UPOILOGGA-0	189 790 650	4 852	-1 053	135
2002	UPOILOGGA-1	189 787 063	4 861	-1 039	126
2002	UPOILOGGA-2	189 695 220	4 874	-1 013	128
2002	UPOILOGLOG-0	189 582 151	4 979	-885	135
2002	UPOILOGLOG-1	189 749 601	4 885	-1 009	126
2002	UPOILOGLOG-2	189 621 033	4 906	-970	128
2002	UZAGA-0	189 149 771	5 459	-260	118
2002	UZAGA-1	189 171 620	5 347	-400	111
2002	UZAGA-2	189 115 419	5 344	-397	111
2002	UZAIG-0	189 149 771	5 459	-260	183
2002	UZAIG-1	189 171 399	5 347	-401	170
2002	UZAIG-2	189 085 481	5 352	-388	171

Tabell 9.3 - Årstabell 2002

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2003	UPOIIG-0	142 033 082	5 283	547	274
2003	UPOIIG-1	142 186 653	5 317	588	279
2003	UPOIIG-2	141 962 140	5 292	565	278
2003	UPOILOG-0	143 799 477	6 077	1 508	500
2003	UPOILOG-1	144 088 164	6 151	1 595	522
2003	UPOILOG-2	144 242 106	6 182	1 644	543
2003	UPOILOGGA-0	142 246 121	5 434	733	124
2003	UPOILOGGA-1	141 876 589	4 274	-716	125
2003	UPOILOGGA-2	141 713 813	4 308	-662	127
2003	UPOILOGLOG-0	141 815 273	4 310	-671	124
2003	UPOILOGLOG-1	141 812 017	4 340	-632	125
2003	UPOILOGLOG-2	141 689 142	4 327	-639	127
2003	UZAGA-0	141 815 729	5 137	367	119
2003	UZAGA-1	142 000 227	5 210	456	123
2003	UZAGA-2	141 828 661	5 213	471	124
2003	UZAIG-0	141 815 729	5 137	367	185
2003	UZAIG-1	141 998 989	5 212	458	193
2003	UZAIG-2	141 848 298	5 229	488	196

Tabell 9.4 - Årstabell 2003

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2004	UPOIIG-0	233 081 335	6 003	-331	244
2004	UPOIIG-1	233 117 149	6 186	-100	272
2004	UPOIIG-2	233 054 848	5 991	-339	247
2004	UPOILOG-0	233 079 648	6 100	-208	297
2004	UPOILOG-1	233 280 438	6 262	-7	338
2004	UPOILOG-2	233 613 525	6 325	77	357
2004	UPOILOGGA-0	233 084 448	5 913	-445	128
2004	UPOILOGGA-1	233 698 942	5 376	-1 130	140
2004	UPOILOGGA-2	233 805 468	5 320	-1 198	141
2004	UPOILOGLOG-0	233 907 205	5 251	-1 295	128
2004	UPOILOGLOG-1	233 655 716	5 407	-1 091	140
2004	UPOILOGLOG-2	233 782 938	5 347	-1 164	141
2004	UZAGA-0	233 141 600	5 768	-629	114
2004	UZAGA-1	233 055 664	5 960	-383	127
2004	UZAGA-2	233 084 991	5 967	-367	127
2004	UZAIG-0	233 141 600	5 768	-629	174
2004	UZAIG-1	233 057 312	5 960	-384	196
2004	UZAIG-2	233 024 086	5 986	-346	201

Tabell 9.5 - Årstabell 2004

År	Modell	Kvadrert avvik	Absolutt avvik	Aggregert avvik	Rot av forventet varians
2005	UPOIIG-0	215 261 753	5 888	135	278
2005	UPOIIG-1	216 056 707	6 471	851	373
2005	UPOIIG-2	214 999 473	5 865	115	280
2005	UPOILOG-0	215 488 195	6 117	418	362
2005	UPOILOG-1	217 563 959	6 954	1 422	562
2005	UPOILOG-2	218 614 081	7 110	1 611	593
2005	UPOILOGGA-0	215 801 651	4 870	-1 160	135
2005	UPOILOGGA-1	215 294 724	5 410	-466	170
2005	UPOILOGGA-2	215 349 590	5 356	-525	174
2005	UPOILOGLOG-0	215 558 731	5 022	-963	135
2005	UPOILOGLOG-1	215 292 387	5 446	-421	170
2005	UPOILOGLOG-2	215 326 675	5 435	-431	174
2005	UZAGA-0	215 110 804	5 550	-289	121
2005	UZAGA-1	215 389 346	6 106	408	158
2005	UZAGA-2	215 293 841	6 174	497	160
2005	UZAIG-0	215 110 804	5 550	-289	187
2005	UZAIG-1	215 395 464	6 107	410	257
2005	UZAIG-2	215 254 282	6 178	501	265

Tabell 9.6 - Årstabell 2005

9.3 Kommentarer til resultatene

Når man ser QQ-plottene og årstabellene i sammenheng, peker det seg ikke ut en klar vinner av tittelen "beste modell". Hvilke modeller som treffer best med estimatene, varierer fra år til år. Jeg er imidlertid interessert i å prøve å avdekke, og eventuelt forklare, noen av mønstrene i disse resultatene. Jeg ønsker også å sette modellene opp mot hverandre og forsøke å avgjøre hvilken modell som er den beste. Det er verdt å merke seg at modellene med postfiks 0, som er estimert uten forklaringsvariabler, ser ut til å ha svært like QQ-plot som de mer fleksible modellene med postfiks 1 og 2. Dette kan være et tegn på at valg av sannsynlighetsfordeling har større betydning for kvaliteten på prediksjonene enn funksjonell form av forklaringsvariabler.

9.3.1 UPOILOG- og UPOIIG-modellene

QQ-plottene i figur 9.1-9.2 er de mest sentrale å se på for å avgjøre hvilken modell som best predikerer fordelingen til *U*. QQ-plottene viser at UPOILOG-modellene treffer dårlig. UPOIIG-modellene treffer noe bedre, men fortsatt relativt svakt, etter QQ-plottene å dømme. Fra årstabellene fremgår det også at UPOIIG-modellene treffer noe bedre enn UPOILOG-modellene. Alle disse modellene forutsetter uavhengighet mellom skadefrekvens og skadepris, og modellerer skadepris unimodalt. Et generelt problem ved disse modellene er at de predikerer for høy varians. Dette fremgår av årstabellene, og er årsaken til at *Z*-verdiene for disse modeller ligger så langt fra standardnormalfordeling. Nevneren i uttrykket for *Z*-ene summerer opp for høy varians, hvilket fører til at *Z*-ene i for stor grad blir konsentrert rundt 0. Når jeg skal velge ut den beste modellen i denne kategorien etter QQ-plottene og årstabellene, faller valget på UPOIIG-2. Den scorer relativt bra på årstabellene for flere årganger. Når det gjelder QQ-plot, har de 3 UPOIIG-modellene såpass like QQ-plot at forskjellen er neglisjerbar.

9.3.2 UPOILOGLOG- og UPOILOGGA-modellene

Disse modellene, som er basert på uavhengighet mellom skadefrekvens og gjennomsnittlig skadepris, samt bimodal modellering av skadepris, har *Z*-verdier med systematisk høyere kvantiler enn standardnormalfordelingen. Dette fremgår av QQ-plottene der alle QQ-punktene ligger over linjen $y = x$ for alle de 6 modellene i denne kategorien. Årsaken til dette er at disse modellene har en tendens til å predikere for lave verdier av $\mathrm{E}(U)$. Det gir videre for

høye verdier av *Z*-ene, og dermed for høye kvantiler. Man kan også se fenomenet i årstabellene, der UPOILOGLOG- og UPOILOGGA-modellene systematisk har negative verdier for aggregert avvik. Ofte har også disse negative verdiene høy absoluttverdi. Til tross for dette er det verdt å merke seg at laveste verdi av absolutt avvik, oppnås av en modell i denne kategorien i alle årgangene. Variansene for disse modellene er av moderat lav størrelse. Jeg legger også merke til at formen på QQ-plottene ikke er veldig langt unna rette linjer med stigningstall 1 (som $y = x$). Dette kan tyde på at variansene er relativt bra estimert ved disse modellene, mens forventningsverdiene systematisk er for lave. En totalvurdering tilsier at den beste modellen i denne kategorien er UPOILOGLOG-2. Dette er den modellen som totalt sett scorer best på årstabellene blant disse 6. QQ-plottene er svært like for alle UPOILOGLOG-modellene. UPOILOGGA-0 skiller seg ut med et relativt bra QQ-plot, men scorer relativt svakt på årstabellene.

9.3.3 UZAIG- og UZAGA-modellene

Disse modellene har *Z*-verdier som ser forholdsvis normalfordelte ut, etter QQ-plottene å dømme. Disse modellene scorer også jevnt over bra på årstabellene, dersom man gjør en totalvurdering av alle årganger. Modellene er basert på direkte estimering av fordelingen til *U,* hvilket ser ut til å være den avgjørende fordelen her. Hvilke av QQ-plottene som er nærmest linjen $y = x$ totalt sett, er noe diffust. UZAGA-modellene treffer bedre for lave kvantiler, mens UZAIG-modellene treffer bedre for høye kvantiler. UZAGA-modellene ser også ut til å predikere systematisk lavere varians enn UZAIG-modellene, uten at predikert forventning er mer treffsikker. Dette, sammen med en totalvurdering av QQ-plot og årstabeller, gjør at jeg regner UZAIG-2 som den beste modellen i denne kategorien. Denne modellen gjør det rimelig bra etter de fleste kriteriene, og har laveste kvadrerte avvik i 3 av årgangene.

---

10 Modellene brukt til prissetting

Hva ville blitt resultatet dersom man, i år 2000, hadde vedtatt en premieformel basert på en av prediksjonsmodellene fra denne oppgaven, for så å benytte denne til premiefastsettelse i et forsikringsselskap? Det skal jeg forsøke å teste i dette kapittelet.

10.1 Om "simulert tidsløp"

Etter drøftingen i delkapittel 9.3 sitter jeg igjen med 3 modeller som kandidat til "beste modell". Disse er UPOIIG-2, UPOILOGLOG-2 og UZAIG-2. Jeg "simulerer" tidsløpet for disse modeller ved å estimere parameterne basert på data fra år 2000. Dette danner grunnlaget for prediksjoner for år 2001. Parameterne estimeres så på nytt basert på data fra både år 2000 og år 2001, for så å brukes til å predikere utbetalinger i år 2002. Slik fortsetter jeg til det er prediksjoner for 2001-2005, der samtlige prediksjoner i en årgang er basert på historikken til tidligere årganger fra og med år 2000. Målet her er ikke, som i kapittel 9, å teste treffsikkerheten, men å teste brukbarheten, i et rent pragmatisk forsikringsperspektiv. Parallelt med prediksjonene regner jeg ut en retrospektiv premie. Som nevnt i delkapittel 3.9, bruker jeg standardavviksprinsippet, $R = b\sqrt{\mathrm{Var}(U)}$, som regel for å bestemme risikotillegget. For å beregne administrasjonskostnadene, *A*, markedstilpasningen, *M*, og ønsket fortjeneste, *F*, ser jeg på forsikringsselskapenes overordnede forretningsmodell. Et viktig element i så måte er å bruke lønnsomhetsmålet *Combined Ratio* - CR, forenklet definert som

$$\mathrm{CR} = \frac{\text{Utbetalinger} + \text{Kostnader}}{\text{Premieinntekter}}.$$ [27]

CR er "combined" i den forstand at den kan videre deles inn i *Kostnadsprosent* og *Skadeprosent*, der begge disse størrelser er prosentandeler av total premieinntekt. Det gir regelen $\mathrm{CR} = \textit{Kostnadsprosent} + \textit{Skadeprosent}$. Jeg antar videre at forsikringsselskapets styre opererer med en fastlagt kostnadsprosent. For å gjøre regnestykket mer realistisk bruker jeg gjennomsnittet av faktisk kostnadsprosent fra de 4 største norske forsikringsselskapene (Gjensidige, If, Tryg og Sparebank1) for 2011.[28] Kostnadsprosenten blir da 18,25 %. Jeg antar

[27] For mer om forretningsmodellen i et forsikringsselskap, og grundigere forklaring av *Combined Ratio* og andre forsikringstermer, se http://en.wikipedia.org/wiki/Insurance.

[28] Følgende tall er hentet ut: 16,4% for Gjensidige, 17,3% for If, 16,8% for Tryg og 22,5% for Sparebank 1. Alle disse tallene er offentlig tilgjengelig informasjon, og er hentet fra følgende nettsider: http://gjensidige.com/web/Forsiden/Investorinformasjon/Rapportering,

videre at forsikringsselskapet ønsker 5 øre i fortjeneste per premiekrone. Ved å legge til 5 % på kostnadsprosenten får man da tallet 23,25 %. Dette tallet representerer andelen av premieinntektene som er satt av til andre ting enn utbetaling til skader. Dermed settes den resterende andelen, 76,75 %, av til skadeutbetalinger.

For hver enkelt polise, *i*, i år *j*, er det en "simulert-tidsløp"-prediksjon, $\mathrm{E}(U_{i,j})$, for hver modell. Jeg definerer grunnpremien for polise (i, j) som $\Pi_{\mathrm{G}} = \frac{\mathrm{E}(U_{i,j})}{0,7675} \approx \mathrm{E}(U_{i,j}) \cdot (1 + 0,3)$, slik at administrasjonstillegget, markedstilpasningen og fortjenesten kan skrives som $A_{i,j} + M_{i,j} + F_{i,j} \approx 0,3\mathrm{E}(U_{i,j})$. Dersom alle prediksjonene stemmer, vil denne grunnpremien være nok til å dekke kostnader på 18,25 % og fortjeneste på 5 % av premieinntektene. Imidlertid må forsikringsselskapet også ta et risikotillegg, ettersom ikke alle prediksjonene treffer i den virkelige verden. Jeg bruker standardavviksprinsippet og setter $b = 2$ %. Det gir risikotillegg $R_{i,j} = 0,02 \cdot \sqrt{\mathrm{Var}(U_{i,j})}$. Den endelige premien, prisen kunden faktisk betaler, blir da

$$\Pi_{i,j} = \frac{\mathrm{E}(U_{i,j})}{0,7675} + 0,02 \cdot \sqrt{\mathrm{Var}(U_{i,j})}.$$

Jeg regner ut denne premien for alle modeller i årene 2001-2005, basert på simulert tidsløp. Resultatet vises i tabell 10.1.

---

http://www.tryg.com/en/governance/general-meeting/next-agm/index.html,

http://www.if.no/web/no/om/Pages/default.aspx

http://investor.sparebank1.no/

10.2 Resultater fra "simulert tidsløp"

År	Modell	Premie	Utbetalinger	Kostnader	Fortjeneste	Skade-prosent	CR
2001	UPOIIG-2	27,3	22,0	5,0	0,2	81 %	99 %
2002	UPOIIG-2	34,1	26,5	6,2	1,4	78 %	96 %
2003	UPOIIG-2	35,0	23,3	6,4	5,3	66 %	85 %
2004	UPOIIG-2	37,2	29,2	6,8	1,2	79 %	97 %
2005	UPOIIG-2	37,8	24,9	6,9	6,0	66 %	84 %
2001	UPOILOGLOG-2	17,6	22,0	3,2	-7,6	125 %	143 %
2002	UPOILOGLOG-2	24,3	26,5	4,4	-6,6	109 %	127 %
2003	UPOILOGLOG-2	26,7	23,3	4,9	-1,4	87 %	105 %
2004	UPOILOGLOG-2	29,2	29,2	5,3	-5,4	100 %	119 %
2005	UPOILOGLOG-2	30,8	24,9	5,6	0,3	81 %	99 %
2001	UZAIG-2	24,9	22,0	4,5	-1,7	89 %	107 %
2002	UZAIG-2	42,8	26,5	7,8	8,6	62 %	80 %
2003	UZAIG-2	46,3	23,3	8,5	14,6	50 %	68 %
2004	UZAIG-2	39,3	29,2	7,2	2,9	74 %	93 %
2005	UZAIG-2	41,4	24,9	7,5	8,9	60 %	79 %

Tabell 10.1 - Resultater av simulert tidsløp for UPOIIG-2, UPOILOGLOG-2 og UZAIG-2. Tall som ikke er prosenttall er i millioner NOK. Kostnader er satt til 18,25 % av premien.

Tabell 10.1 viser at regnskapsmessig er det UPOIIG-2, som over tid ligger nærmest opptil forsikringsselskapets målsetning (skadeprosent 76,75 %). Forretningsmessig er dette en klar fordel, ettersom det gir, på et aggregert nivå, forutsigbarhet og stabilitet over tid. UPOILOGLOG-2 gir for lav premie i hver eneste årgang. Som nevnt i delkapittel 9.3.2 skyldes dette at predikert forventning er systematisk for lav. Dersom denne modellen skal brukes til prising, må koeffisienten *b* i risikotillegget økes. Tabell 10.1 viser også at skadeprosenten for UPOILOGLOG-2 er svært ustabil. Dette er negativt for selskapet ettersom det gir en lite forutsigbar økonomisk situasjon. UZAIG-2 gir over tid høye premier og dermed lave skadeprosenter. Et problem her er at skadeprosentene er generelt for lave, og for ustabile. For høy skadeprosent betyr at forsikringsselskapet taper penger. Det er selvsagt ikke bærekraftig over tid. For lav skadeprosent kan imidlertid bety at forsikringene er overpriset. Da risikerer selskapet at kundene forsvinner til andre selskaper som priser lavere.

10.3 Marked og konkurranse

Som tabell 10.1 viser, er skadeprosenten for UPOIIG-2 relativt stabil rundt målsetningen (76,75 %). Dette gir et fortrinn i markedet ettersom det betyr at forsikringene er riktig priset på aggregert nivå. Dersom en kunde er riktig priset, vil andre forsikringsselskaper ikke kunne lokke til seg kunden med lavere premie, uten selv å tape penger. Imidlertid viser resultatene i kapittel 9 at prediksjonene til UZAIG-2 er mer troverdige på individnivå. Hver enkelt kunde er mest opptatt av prisen på sin egen forsikring. Følgelig vil riktig prising på individnivå være mer verdifullt enn riktig prising på aggregert nivå. Dersom prisingen er riktig på aggregert nivå, men feil på individnivå, betyr det at lavrisikokunder subsidierer høyrisikokunder i samme portefølje. Det betyr videre at konkurrerende forsikringsselskaper vil kunne underby prisen på lavrisikokundene, og dermed overta dem som kunder. Dette kan gi fortjeneste både for kunden og det konkurrerende forsikringsselskapet. Konkurranseaspektet tilsier dermed at riktig prising på individnivå er viktigere enn riktig prising på aggregert nivå. Bruk av UZAIG-2, som etter resultatene i kapittel 9 anses som riktigst priset på individnivå, gir ustabil skadeprosent på aggregert nivå. Dersom man imidlertid opererer med en lengre tidshorisont enn 1 år av gangen, og ser skadeprosenten i for eksempel en 5 års periode, vil skadeprosenten være mer stabil også på aggregert nivå.

10.4 Feilkilder og kommentarer

Prising ved "simulert tidsløp" er en interessant retrospektiv teknikk for å teste ut prisingsmodellene i møte med virkeligheten. Det er imidlertid viktig å være klar over feilkildene og innvendingene mot denne metoden. Her er en noen mulige feilkilder og innvendinger:

- Alle utregninger baserer seg på polisedata og skadedata for virkelige kunder. Premien disse kundene betalte i virkeligheten er ikke tilgjengelig i datasettet. Kundeadferden[29] behandles implisitt som identisk med den historiske kundeadferden, uansett premieutregninger. I virkeligheten vil kundeadferden endre seg når premien endres, og dette er ikke tatt høyde for i delkapitler 10.1-10.3. Dersom jeg for eksempel hadde satt

[29] *Kundeadferd* refererer her til "hvilke kunder som er forsikret i hvilket selskap i hvilket tidsrom". Her ses hele forsikringsmarkedet under ett, og *kundeadferd* refererer også til potensielle kunder, som ikke har observasjoner i datasettet i det hele tatt, fordi de var forsikret i andre selskaper i hele perioden.

$b = 100\ \%$ i formelen for risikotillegget, ville premiene blitt ekstremt høye, og de fleste kundene ville med stor sannsynlighet skiftet forsikringsselskap.

- Kostnader i tabell 10.1, er hele veien satt som 18,25 % av premien, slik at kostnadene over tid behandles som proporsjonale med premieinntektene. Det er tvilsomt om denne proporsjonaliteten er konstant over tid. Ser man for eksempel på skadebehandling, som er en stor del av kostnadene, bør denne delen heller ses som proporsjonal med skadeutbetalingene. Dersom man priser riktigere på individnivå enn konkurrerende selskaper, vil høyrisikokundene bli tilbudt lavere pris hos andre forsikringsselskaper, mens andre forsikringsselskapers lavrisikokunder vil bli tilbudt lavere pris. Det sannsynlige resultatet av dette, dersom man priser riktigere enn konkurrentene, er en vridning i porteføljen i retning av flere lavrisikokunder og færre høyrisikokunder. Dette gir videre færre og mindre kompliserte skadesaker, og dermed lavere kostnadsprosent.
- Målsetningen er satt til 76,75 % hvert av årene, ettersom det vil gi 5 % fortjeneste. Konkurransesituasjonen, finanssituasjonen og andre forhold gjør at disse tallene gjerne vil forandre seg over tid. Samtidig er det naturlig, selv om man i utgangspunktet vedtar en standard prisingsformel, å justere prisene noe etter markedssituasjonen. Skadeprosenten som oppnås i ett år, kan også legge føringer for ønsket skadeprosent året etter.
- Beregning av sannsynlighetsfordelingen til utbetalinger på polise i, U_i, er en oppgave som hører til faggrenene matematikk og statistikk. Når man så har funnet en modell som vedtas, er det å bestemme premien Π_i en oppgave som i større grad er åpen for pragmatiske vurderinger. Det kan gjøres på mange ulike måter, hvorav mange ikke kan rettferdiggjøres fra et vitenskapelig/matematisk fundament. Metodikken jeg har brukt i delkapitler 10.1 og 10.2 er ment som et forslag til prising, og skal ikke forstås som en fasit, eller som den eneste måten man kan gjøre dette på.

11. Avslutning

11.1 Konklusjon

Når jeg ser resultatene fra kapittel 9 under ett, mener jeg at UZAIG-2 er den modellen som gjør det best av de 18 modellene jeg har testet i denne oppgaven. Mer generelt vil jeg hevde at ZAIG-modellene gjør det best. Bortoluzzo et al. (2011) og Heller et al. (2006) har også fått gode resultater ved å anvende denne sannsynlighetsfordelingen for modellering av *U*. Når man sammenlikner de estimerte koeffisienter i delkapitlene 6.3, og 7.8 med de tilsvarende estimerte koeffisienter i delkapittel 8.6, er likheten meget stor. Det betyr at ZAIG-modellene med god presisjon greier å fange opp, og skille ut, effektene på skadefrekvens og skadepris. Følgelig kan det anbefales å bruke ZAIG-modellering for total skadeutbetaling, også dersom man primært ønsker å studere effekter, og ikke nødvendigvis predikere utbetalinger.

Et sentralt resultat fra delkapittel 9 er at modellene med postfiks 2 kommer totalt sett best ut. Disse har funksjonell form på forklaringsvariablene som er bestemt ved AIC-minimerings-algoritmen. Jeg har brukt all tilgjengelig data fra 2000-2005 og kjørt disse gjennom AIC-minimerings-algoritmen for å komme frem til modellene med postfiks 2 (se delkapittel 5.3). Det vil si at de er AIC-minimert *in sample*, på samme datasett som blir brukt til å estimere parameterne. Kryssvalideringen i kapittel 9 er testing *out of sample*, i betydningen at alle parametere estimeres på et annet datasett enn det de brukes til å predikere på. Når da modellene med postfiks 2 gjør det jevnt over bedre i denne *out of sample*-settingen, tyder det på at de funksjonelle formene AIC-minimerings-algoritmen anbefaler, faktisk gjenspeiler virkelige effekter i populasjonen. Dersom dette er representativt, leder det meg til å anbefale AIC-minimering av forklaringsvariablenes funksjonelle form, for å lage prediksjonsmodeller.

11.2 Forslag til anvendelse

Et poeng som resultatene i denne oppgaven illustrerer, er betydningen av valg av statistisk modell. Når et forsikringsselskap jobber for mer risikoriktig prising, er det derfor vel så viktig å fokusere på modellvalg og prisingsprinsipper, som på innhenting av informasjon om kundene. Jeg har ikke tilgang til alle forklaringsvariabler man vanligvis vil ha mulighet til å bruke i en slik modellering i et forsikringsselskap. Det må derfor nevnes at når et stort antall forklaringsvariabler er i bruk på såpass fleksible statistiske modeller, vil det kreve meget stor datakraft for å maksimere likelihooden. Samtidig øker faren for konvergensproblemer når

antall forklaringsvariabler blir høyt. Dette er særlig et problem dersom man opererer med findelte, kategoriske forklaringsvariabler, som for eksempel kommune.[30] En elementær strategi i forhold til riktig prisdifferensiering er å tolke det som ensbetydende med mer findeling av forklaringsvariablene. Jeg vil argumentere for at det er vel så viktig å se på de matematiske modellene, og jobbe med forbedring av disse.

11.3 Forslag til videre studier

- De bimodale FM-modellene gjorde en svært god jobb i modellering av skadepris (se delkapitler 7.5-7.7). Imidlertid gav koblingen mellom disse modellene og Poissonmodellene for skadefrekvens relativt svake resultater. Det kunne vært interessant å teste en utvidelse av de bimodale FM-modellene for skadepris, der nullsannsynlighet legges til. En slik modell kan også sammenlignes med ZAIG- eller ZAGA-modellene, bortsett fra at den positive delen av utfallsrommet får 2 sannsynlighetsfordelinger i stedet for 1.
- I delkapittel 5.5.2 foreslår jeg en mulig sammenheng mellom eksponeringstid, r_i, forventet antall skader per år, λ_i, og forventet antall skader per polise μ_i. Den antatte sammenhengen $\mu_i = r_i \cdot \lambda_i$ finner ikke støtte i datasettet. Jeg foreslår derfor den alternative sammenhengen $\mu_i = \sqrt{r_i} \cdot \lambda_i$, som passer langt bedre med observasjonene i datasettet. Dette impliserer at kunder som har vært forsikret i selskapet i kort tid har en uforholdsmessig høy skadefrekvens. Det kunne vært svært interessant å undersøke hvorvidt min foreslåtte relasjon også støttes av andre data, og eventuelt finne ut noe om årsakene til denne sammenhengen.
- GAMLSS-rammeverket, slik det er implementert i R, tillater å korrigere for korrelerte observasjoner ved bruk av "random effects"-teknikken. Det vil for eksempel si at man kan korrigere for kundeeffekten (ulike observasjoner fra samme kunde kan antas å være korrelerte) ved å innføre en kundeparameter. Denne kundeparameteren kan modelleres som en "random effect", slik at det estimeres en egen sannsynlighetsfordeling for denne. Når man så skal bruke modellen til å predikere

[30] I Norge er det 429 kommuner, og dersom forsikringsselskapet har kunder i alle disse, vil det føre til estimering av 429 parametere for denne ene forklaringsvariabelen.

verdier utenfor kundebasen, gjøres tilfeldige trekk fra den estimerte sannsynlighetsfordelingen for å bestemme kundeparameterens verdi. Dette er et eksempel på hvordan man kan bruke random effects til å analysere datasettet jeg har brukt, eller liknende datasett.

11.4 Forbehold, feilkilder og begrensninger

I en empirisk studie som denne oppgaven, vil det alltid være mulig å finne innvendinger. Det er i denne sammenheng viktig å peke på hvor mulige feilkilder kan ligge, og hvilke begrensninger resultatene og konklusjonene har. Her følger en punktvis oversikt:

- **Feil i datasettet** kan forekomme, ettersom data ofte er registrert manuelt, hvilket gjør at tastefeil og registreringsfeil ikke kan utelukkes.
- **Programmeringsfeil** i R kan ikke utelukkes, ettersom mengden kildekode er meget stor. Slike feil er vanskelig å oppdage, men jeg har forsøkt så godt jeg kan å teste resultatene (også mellomregninger), og påse at de gir mening og virker rimelige.
- **Foreldede data** er et mulig problem i oppgaven. Jeg har skrevet oppgaven i 2012, og datasettet er fra 2000-2005. Fra 2000 til 2012 har det skjedd samfunnsmessige forandringer og bilteknologiske forandringer som kan være relevante her. Imidlertid ser jeg det som rimelig at for eksempel effekten av høy bilalder på U er sammenliknbar i 2000 og 2012. Problemet med foreldede data kan være neglisjerbart, men det er verdt å nevne.
- **Ikke-representative data** er et mulig problem i oppgaven. I årene 2000-2005 var det i gjennomsnitt 1 829 342 biler med kaskoforsikring i Norge (se http://fno.no/no/Hoved/Statistikk/skadeforsikring/). Gjennomsnittlig antall poliser i datasettet for en årgang er 9 088. Det representerer 0,5 % av markedet. Observasjonene i datasettet er tilfeldig valgt innad i forsikringsselskapet jeg har fått datasettet fra. Hvorvidt det dermed er et representativt utvalg fra det norske forsikringsmarkedet, er et åpent spørsmål. Det er imidlertid grunn til å tro at effektene av forklaringsvariablene på responsvariablene ikke er spesifikke for hvert forsikringsselskap, men universelle populasjonseffekter. Ser man på problemet med ikke-representative data i en internasjonal kontekst, er denne feilkilden potensielt større.
- **Få forklaringsvariabler** er tilgjengelige her. Det betyr at det alltid er rom for å stille spørsmålstegn ved de målte effektene, og hvorvidt det er avdekket genuine

årsakssammenhenger, eller bare symptomer på andre underliggende årsakssammenhenger. Andre forklaringsvariabler det kunne vært interessant å ha med, er for eksempel fylke, bilmerke, årlig kjørelengde etc.

- **Feil forklaringsvariabler** er også en mulig feilkilde. Bilalder har klar betydning for skadefrekvensen (se tabell 6.4). Det kan imidlertid godt tenkes at andre, nært beslektede, forklaringsvariabler kan ha større betydning. Eksempler her er antall km bilen har kjørt og hvor mange år nåværende bileier har eid bilen. Personalder, slik det inngår i datasettet for denne oppgaven, er alderen på den som tegner forsikringen. Det kan imidlertid forekomme at bilen kjøres av sønnen eller datteren til forsikringstaker. Dersom jeg hadde hatt tilgang til informasjon om alder på yngste bruker av bilen i datasettet, kan det tenkes at det ville gitt mer forklaringskraft enn personalder slik den fremstår i oppgaven.

---

12. Litteratur

i. **Akaike, H.** (1974). "A new look at the statistical model identification" i *IEEE Transactions on Automatic Control* **19** (6): 716–723.

ii. **Bortoluzzo, A.B., Claro, D.P., Caetano, M.A.L., Artes, R.** (2011) "Estimating Total Claim Size in the Auto Insurance Industry: a Comparison between Tweedie and Zero-Adjusted Inverse Gaussian Distribution." i *BAR, Curtiba,* v.8, n.1 art. 3, pp. 37-47, Jan./Mar. 2011

iii. **Breiman, L., Friedman, J. H.** (1985). "Estimating optimal transformations for multiple regression and correlations (with discussion)". *Journal of the American Statistical Association* **80** (391): 580–619

iv. **Burnham, K. P., Anderson, D.R.** (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag

v. **Casella, G., Berger, L.** (2002) *Statistical Inferrence*, 2. utgave, Duxbury, Pacific Grove

vi. **Cole, T. J. and Green, P. J.** (1992) *Smoothing reference centile curves: the LMS method and penalized likelihood*. Statist. Med., 11, 1305–1319.

vii. **de Jong, P., Heller, G.Z.** (2008) *Generalized Linear Models for Insurance Data*. Camebridge University Press, Camebridge

viii. **Dobson, A.J., Barnett, A.G.** (2008) *An introduction to Generalized Linear Model, third edition*. Boca Raton: Chapman and Hall/CRC.

ix. **Gupta, M. R., Chen, Y.** (2010) "Theory and Use of the EM Algorithm" i *Foundations and Trends in Signal Processing*, Vol.4, No. 3 (2010) sider 223-296.

x. **Haberman, S., Renshaw, A.E.** (1996) "Generalized Linear Models and Actuarial Science." I *The Statistician*, 45 (4), 407-436.

xi. **Hastie, T.J, Tibshirani, R.J.** (1990) *Generalized Additive Models*, Chapman and Hall, New York.

xii. **Heller, G., Stasinopoulos, M., Rigby, B.** (2006) "The zero-adjusted Inverse Gaussian distribution as a model for insurance claims" i *Proceedings of the International Workshop on Statistical Modelling*, Galway, Ireland, 21.

xiii. **Hogg, R.V., Klugman, S.A.** (1984) *Loss Distributions*. Wiley, New York.

xiv. **Hogg, R.V., Tanis, E.A.** (2010) *Probability and Statistical Inference*, 8. Utgave, Pearson Education, Upper Saddle River, New Jersey.

xv. **Johnson, N.L., Kotz, S og Balakrishnan, N.** (1994). *Continous Univariate Distributions*, Volume I, 2. Utgave, Wiley, New York

xvi. **Jørgensen, B., de Souza, M.C.P.** (1994) "Fitting Tweedie's compound Poisson model to insurance claims data." i *Scandinavian Actuarial Journal*, 69-93.

xvii. **Lambert, D.** (1992) "Zero-inflated Poisson Regression with an application to defects in Manufacturing." *Technometrics*, **34**: 1-14.

xviii. **Lange, K.** (1999) *Numerical Analysis for Statisticians.* New York: Springer.

xix. **Le Cam, L.** (1986) "The central limit theorem around 1935", *Statistical Science* 1:1, 78-91

xx. **McCullagh, P, Nelder, J. A.** (1989). *Generalized Linear Models, second edition*. Boca Raton: Chapman and Hall/CRC.

xxi. **Nelder, J.A., Wedderbum R.W.M.** (1972) "Generalized linear models". Side 370-384 I *Journal of the Royal Statistical Society, series A 135*

xxii. **Rigby, R.A., Stasinopoulos, D.M.** (2001) "The GAMLSS project: a flexible approach to statistical modeling". Side 337-345 I Klein, B og Korsholm, L (red.), *New Trends in Statistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, Odense, Danmark.

xxiii. **Rigby, R.A., Stasinopoulos, D.M.** (2005) Generalized Additive Models for Location, Scale and Shape, (with discussion). Appl. Statist., 54, (sider 507-554).

xxiv. **Sundt, B.** (1999) *An Introduction to Non-Life Insurance Mathematics*, 4. Utgave VVW, Karlsruhe.

xxv. **Weisberg, H.I., Tomberlin, T.J.** (1982) "A Statistical Perspective on Actuarial Methods for Estimating Pure Premiums from Cross-Classified Data" i *The Journal of Risk and Insurance*, Vol 49 (4), 539-563

xxvi. **Young, V.R.** (2004) "Premium Principles" i *Encyclopedia of Actuarial Science*, John Wiley & Sons.Ltd.