Ole Johan Sørensen Schei

Oppgave for graden master i statistikk
Finansteori og forsikringsmatematikk
Universitetet i Bergen, Norge
15. september 2009

Regresjonsmodeller i skipsforsikring

Denne oppgaven er skrevet i LaTeX 2ε med dokumentklassen «uib-mi-master», laget av Karl Ove Hufthammer. Den ble kompilert med pdfTeX-1.40.3 den 15. september 2009. Brødteksten er satt i 11 punkts URW Palladio med kapitéler. Matematikken er satt i URW Palladio og Pazo Math, overskrifter i HV Math og programkode i Bera Mono.

Takk

Først og fremst vil jeg takke min veileder, Jostein Paulsen, for å ha kommet med en interessant oppgave og for å ha kommet med gode råd underveis.

Jeg vil også takke Geir Drage Berentsen for hjelp med teoretiske problemer og Arne Johannes Holmin for å ha hjulpet meg med programmeringen. Eivind Reikerås og Tor Erik Melseth fortjener ros for godt samarbeid og evinnelige politiske diskusjoner gjennom bachelor- og mastergraden.

Det er på sin plass å takke Trygve Nilsen for å ha vært en god ressursperson gjennom masterstudiene og Hans Julius Skaug for hjelp med funksjonen glmm.admb. Jeg vil takke Karl Ove Hufthammer for å ha tatt seg tid til å lage masterklassen i latek. Det har spart meg, og mange andre masterstudenter, for mye ekstraarbeid.

Mine foreldre og mine 2 brødre fortjener ros for å alltid ha vært gode støttespillere. Ellers vil jeg takke Koop, Beinet, Stern, Monica og de andre på Kroepeliens/Jonas Reins gate som har bidratt til en god studietid.

Innhold

Forklaring av notasjon

$\xrightarrow{d}$	konvergens i fordeling
$\xrightarrow{p}$	konvergens i sannsynlighet
iid.	identisk og uavhengig fordelt
trace($\boldsymbol{A}$)	summen av diagonalen til matrisen $\boldsymbol{A}$
max	maksimum
$\mathbb{R}$	den reelle tallinjen
$\mathbb{R}^p$	det p-dimensjonale rommet $\underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{p \text{ ganger}}$
$\forall$	for alle
$\lvert\boldsymbol{A}\rvert$	determinanten til matrisen $\boldsymbol{A}$
$\boldsymbol{A}'$	matrisen $\boldsymbol{A}$ transponert
$1_{(a=x)}$	Indikatorfunksjon: = 1 hvis $a = x$ = 0 hvis $a \neq x$
$\Gamma()$	gamma funksjonen definert ved $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,\mathrm{d}t$
$\mathcal{N}(\mu, \sigma^2)$	normalfordelingen med forventning μ og varians σ^2
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{A})$	multivariat normalfordeling med dimensjon p, forventningsvektor $\boldsymbol{\mu}$ og kovariansmatrise $\boldsymbol{A}$
$F(n_1, n_2)$	F-fordelingen med n_1 og n_2 frihetsgrader
$Y \sim f$	den tilfeldige variabelen Y har fordelingsfunksjon f
SME	sannsynlighetsmaksimeringsestimatoren
x!	fakultet av x

1 Introduksjon

I denne masteroppgaven skal jeg undersøke et datasett innen skipsforsikring vha. ulike regresjonsmodeller. Hovedfokuset vil være på modeller med tilfeldige effekter, såkalte generaliserte lineære miksede modeller. Dette er en type modeller som kan være veldig kompliserte, men kan gi god tilpasning til spesielle typer datasett.

1.1 Datagrunnlag

Datasettet jeg undersøker består av 62 569 rekker og 16 kolonner. Kolonnene inneholder følgende informasjon:

Claim Number Antall forsikringskrav til det aktuelle skipet det året.

Days.covered Fraksjon av året skipet er forsikret.

Age Skipets alder det gjeldende året.

GT Bruttotonn, indeks som gir et mål på skipets innvendige volum.

Sum.Insured Forsikringssum.

HP.prime.mover Antall hestekrefter.

Stroke Indikator for type maskineri. Verdi 0 for en 2-takter og verdi 1 for en 4-takter.

UW.year Tegningsår, dvs. hvilket år skipet ble forsikret.

Ship type Indikator for hva slags type skip det er.

Claim size 1 ((Størrelse på forsikringskrav+egenandel)/forsikringssum) for det første kravet. Egenandelen er en nedre grense for hva forsikringsselskapet dekker. Hvis skadebeløpet er mindre enn egenandelen vil ikke forsikringsselskapet dekke skaden.

Claim size 2 Det samme som Claim size 1, men for det andre kravet.

Claim size 3 Det samme som Claim size 1, men for det tredje kravet.

Claim size 4 Det samme som Claim size 1, men for det fjerde kravet.

Claim size 5 Det samme som Claim size 1, men for det femte kravet.

Basic.Ded Egenandel dividert med forsikringssum.

Datasettet er ikke helt komplett, i den grad av at det er manglende verdier i noen av rekkene. Dette blir fremhevet med verdien -1 i datasettet. Kolonner med manglende verdier er gitt i følgende tabell:

Kolonne	Antall mangelfulle verdier
Age	343
GT	54
HP.prime.mover	1676
Stroke	8906

Tabell 1.1: Manglende verdier.

Tabellen viser at det spesielt for Stroke er mange verdier som mangler. For alle beregningene i masteroppgaven har jeg fjernet alle rader som inneholder mangelfulle verdier. Det gjelder også for tabellen nedenfor.
Antall forsikringskrav, dvs. skadefrekvensen, er den mest interessante kolonnen og den jeg skal undersøke mest. Skadefrekvens er heltallsdata. Jeg vil i denne masteroppgaven undersøke skadefrekvens for tankskip og alle typer

skip(dvs. hele datasettet). For tankskip har jeg 12 318 observasjoner, mens jeg for alle typer skip har 53 588 observasjoner. I dette datasettet er det særdeles mange observasjoner for skadefrekvens med verdi 0, noe den følgende tabellen viser.

Type	Gjennomsnitt	Største verdi	Skadeprosent	Varians	Sum
Tankskip	0.11	5	9.06 %	0.12	1 231
Alle skip	0.09	5	9.92 %	0.11	5 926

Tabell 1.2: Skadefrekvens til datasettet.

I tabellen viser skadeprosenten hvor mange observasjoner som er ulik 0 for skadefrekvens, dvs. hvor mange skip som har rapportert skader til forsikringsselskapet. Kolonnen for gjennomsnitt viser at tankskip har høyere andel rapporterte skader enn alle typer skip.

1.2 Temaer i oppgaven

I kapittel 2 vil jeg undersøke og modellere frekvensen til antall forsikringskrav. En del grunnleggende teori som er nødvendig for analysen vil også bli gjennomgått.
Kapittel 3 fokuserer hovedsakelig på skadebeløp, og jeg vil bl.a. forklare lineære og lineært miksede modeller.
Kapittel 4, 5 og 6 handler om generalisert lineært miksede modeller. Kapittel 4 viser grunnleggende teori for slike modeller og anvender en modell på dataene for skadefrekvens. I kapittel 5 ser jeg på en metode som gjør det mulig å bruke ulike typer modeller på dataene for skadefrekvens. Kapittel 6 tar for seg hypotesetesting og datagenerering av generalisert lineært miksede modeller.

Alle beregninger i denne masteroppgaven er blitt gjort med statistikkprogrammet R.
Jeg har valgt å bruke punktum som desimaltegn i denne masteroppgaven.

2

Ulike regresjonsmodeller for skadefrekvens

I dette kapittelet skal jeg modellere frekvensen til antall forsikringskrav. For å gjøre dette vil jeg bruke vanlige regresjonsmodeller for heltallsdata. Jeg vil nå introdusere noen slike modeller, og så bruke disse på datasettet mitt. VIF, AIC og BIC blir også forklart siden det blir brukt på modellene.

2.1 Generaliserte lineære modeller(GLM)

GLM ble introdusert av Nelder og Wedderburn (1972). Temaet ble så utarbeidet videre i boken til McCullagh og Nelder (1983).
En generalisert lineær modell inneholder følgende:

– Responsvariabler $Y_1, \ldots, Y_n$ som alle deler den samme fordelingen fra den eksponensielle familien. Fordelingen må være på kanonisk form. Fordelingsfunksjonen er da på formen

$$f(y_i; \theta_i, \varphi) = \exp\left(\frac{y_i\theta_i - A(\theta_i)}{\varphi} + c(y_i, \varphi)\right),\ i = 1, \ldots, n.$$

Her er $A(\theta_i)$ og $c(y_i, \varphi)$ kjente funksjoner, θ_i er den naturlige parameteren og er en funksjon av parametrene i modellen, mens φ er en skaleringsparameter.

- En parametervektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ og en matrise med forklaringsvariabler(kovariater) $\boldsymbol{X}$.

- En glatt og invertibel link-funksjon, g, som binder sammen forventningsvektoren $\boldsymbol{\mu}$ og forklaringsvariablene på følgende måte:

$$g(\mu_i) = \xi_i = \boldsymbol{X_i}'\boldsymbol{\beta}\,,$$

hvor ξ_i er definert som den lineære prediktoren.
Forventningen og variansen til Y_i henger sammen på følgende måte:

$$\mu_i = \mathbb{E}[Y_i] = A'(\theta_i),\ \mathrm{Var}(Y_i) = \varphi A''(\theta_i) = \varphi \nu(\mu_i).$$

Her er $\nu(\mu_i)$ definert som $A''(\theta_i)$ og kalles ofte varians-funksjonen, siden den viser hvordan variansen til Y_i avhenger av μ_i.

For GLM og andre regresjonsmodeller er det vanlig å bruke et eller flere offset ledd. Et offset ledd, heretter kalt motvekt, er en komponent av den lineære prediktoren som er kjent på forhånd. Man trenger ikke å beregne noen parametere til motvekten siden den er kjent på forhånd. Det er altså en forklaringsvariabel i den lineære predikatoren som ikke har en tilhørende parameter. Motvekten blir holdt fast mens andre forklaringsvariabler blir evaluert.
I første omgang vil jeg bruke Days.covered som motvekt. Days.covered er en andel av tiden hvert skip er forsikret. Det er naturlig å forvente høyere skadefrekvens til et skip dess lengre skipet har vært forsikret. Derfor vil jeg multiplisere forventningsverdien til hver modell med motvekten Days.covered. I R blir motvekter lagt til i den lineære prediktoren. Derfor må jeg transformere Days.covered med hver modell sin link-funksjon før jeg bruker den som motvekt, slik at Days.covered blir multiplisert med μ. I alle modeller for skadefrekvens vil jeg bruke logaritmen som link-funksjon til μ. Derfor vil jeg i praksis bruke log(Days.covered) som motvekt i disse modellene. For å vise dette lar jeg ϖ betegne Days.covered og η betegne motvekten i en GLM med

logaritmen som link-funksjon. Da får jeg

$$\begin{aligned} \omega_i \mu_i &= \omega_i g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}) = \omega_i e^{\boldsymbol{X}_i'\boldsymbol{\beta}} = e^{\boldsymbol{X}_i'\boldsymbol{\beta}+\log(\omega_i)} \ , \\ \Rightarrow \xi_i &= \boldsymbol{X_i}'\boldsymbol{\beta} + \log(\omega_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + \eta_i \ , \eta_i = \log(\omega_i). \end{aligned}$$

2.1.1 Poisson regresjon

Poisson regresjon er en generalisert lineær modell med logaritmen som link-funksjon. Responsvariablene er Poisson fordelt. Poisson fordelingen er gitt ved

$$\mathbb{P}(Y = y; \mu) = \frac{e^{-\mu}\mu^y}{y!}; \ y = 0,1,\ldots; \ 0 < \mu < \infty \ .$$

Dette er kanskje den mest vanlige fordelingen for å modellere heltallsdata. Poisson fordelingen skiller seg blant annet fra andre fordelinger ved at forventningsverdien er lik variansen, dvs. $\mathbb{E}[Y] = \text{Var}(Y)$. Den generaliserte lineære modellen for Poisson-regresjon er definert ved

$$\mathbb{E}[Y_i] = \mu_i = e^{\boldsymbol{X_i}'\boldsymbol{\beta}+\eta_i}; \ Y_i \sim \text{Poisson}(\mu_i) \ ; \ \ i = 1,\ldots,n \ .$$

Her er η_i motvekten til modellen og link-funksjonen er gitt som

$$\log(\mu_i) = \boldsymbol{X_i}'\boldsymbol{\beta} + \eta_i \ .$$

Et stort problem med Poisson modellen er når det er overdispersjon i datasettet, dvs. når $\text{Var}(Y_i) > \mathbb{E}[Y_i]$. Poisson modellen er uegnet til å modellere datasett hvor variansen er mye større enn forventningen. Tabell 1.2 på side 3 gir at det er litt overdispersjon i datasettet, men ikke nok til at man kan avfeie Poisson modellen.
I R kan man gjøre Poisson regresjon ved å bruke funksjonen glm. Denne funksjonen ble i R opprinnelig introdusert av Simon Davies.

2.1.2 Kvasi-likelihood og Kvasi-Poisson

Kvasi-likelihood ble introdusert av Wedderburn (1974). Min referanse på området vil være McCullagh og Nelder (1989).
Jeg forutsetter at komponentene i responsvektoren $\boldsymbol{Y}$ er uavhengige med forventningsvektor $\boldsymbol{\mu}$ og kovariansmatrise $\varphi \boldsymbol{V}(\boldsymbol{\mu})$. Her kan φ være ukjent

og $\boldsymbol{V}(\boldsymbol{\mu})$ består av kjente funksjoner. Som tidligere er $\boldsymbol{\mu}$ en funksjon av en kovariatmatrise $\boldsymbol{X}$ og en parametervektor $\boldsymbol{\beta}$. Siden komponentene til $\boldsymbol{Y}$ er forutsatt å være uavhengige må matrisen $\boldsymbol{V}$ være diagonal.
Jeg ser først på et enkelt komponent Y i responsvektoren $\boldsymbol{Y}$, der $\nu(\mu)$ er leddet i $\boldsymbol{V}(\mu)$ som samsvarer med Y. Under forutsetningene ovenfor vil funksjonen

$$U = u(\mu; Y) = \frac{Y - \mu}{\varphi \nu(\mu)} ,$$

dele flere av de samme egenskapene som karakteriserer en derivert log likelihood. Disse egenskapene er

$$\mathbb{E}[U] = 0, \;\; \mathrm{Var}(U) = \frac{1}{\varphi \nu(\mu)} , \;\; -\mathbb{E}\left[\frac{\mathrm{d}U}{\mathrm{d}\mu}\right] = \frac{1}{\varphi \nu(\mu)} .$$

Mye første-ordens asymptotisk teori knyttet til likelihood funksjoner er basert på disse egenskapene. Det følgende integralet vil derfor til en viss grad oppføre seg som en log likelihood funksjon for μ.

$$Q(\mu; y) = \int_Y^\mu \frac{Y - t}{\varphi \nu(t)} \, \mathrm{d}t .$$

$Q(\mu; y)$ kalles kvasi-likelihood for μ basert på informasjonen Y. Siden komponentene til $\boldsymbol{Y}$ er forutsatt å være uavhengige blir kvasi- likelihood for hele datasettet summen av de uavhengige bidragene:

$$\boldsymbol{Q}(\boldsymbol{\mu}; \boldsymbol{Y}) = \sum_{i=1}^{n} Q_i(\mu_i; Y_i).$$

Kvasi-likelihood er en måte å beregne hvor godt kovariater passer til responsvariabler når man ikke har en likelihood funksjon å forholde seg til. Man trenger bare å spesifisere forholdet mellom forventningen og variansen til $\boldsymbol{Y}$ opp til en proposjonal konstant. Ved å bare stille disse forutsetningene lar man formen til fordelingsfunksjonen være helt fri. Men på grunn av de få forutsetningene så er det begrenset hvor mye inferens man kan få fra metoden. Man kan få et estimat av $\boldsymbol{\beta}$ og standardfeil til estimatet, men siden man ikke har noen eksakt likelihood funksjon så kan man ikke sammenligne ulike modeller f.eks. med AIC. Modellsammenligning kan bare skje med nøstede modeller.
Det er ingen fordelingsfunksjon for kvasi-likelihood man kan bruke til ulike

formål, men kvasi-likelihood kan gi en pekepinn på hvilken fordeling som passer til dataene.

Eksempel 2.1.1: Kvasi-Poisson

Hvis $\boldsymbol{y}$ er uavhengig Poisson fordelt med parameter $\boldsymbol{\mu}$ vil $\boldsymbol{V(\mu)}$ være en diagonalmatrise der diagonalen er lik $\boldsymbol{\mu}$. Jeg får dermed at

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu \frac{y-t}{\varphi t}\,\mathrm{d}t \\ &= \frac{1}{\varphi}\left[y\log(t) - t\right]\rfloor_y^\mu \\ &= \frac{1}{\varphi}\left(y\log(\mu) - \mu - y\log(y) + y\right) \\ \boldsymbol{Q(\mu; y)} &= \frac{1}{\varphi}\sum_{i=1}^{n}\left(y_i\log(\mu_i) - \mu_i - y_i\log(y_i) + y_i\right) \end{aligned}$$

Estimater til parametervektoren $\boldsymbol{\beta}$ kan beregnes ved å løse $\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{Q(\mu; y)} = 0$.

Zeileis *et al.* (2008) hevder at Poisson modellen ofte er brukbar til å beskrive forventningen $\boldsymbol{\mu}$, men underestimerer variansen i dataene. Det kan føre til problemer siden jeg blant annet er interessert i å modellere de største skadefrekvensene. Det kan også føre til at standardfeilene til den estimerte parametervektoren $\hat{\boldsymbol{\beta}}$ blir unaturlig lave, slik at man får et feil bilde av hvor gode parametrene er. Derfor kan det være nyttig å sammenligne Poisson modellen med en kvasi-Poisson modell. De estimerte verdiene til $\boldsymbol{\beta}$ vil bli de samme, men standardfeilene til kvasi-Poisson er justert for at variansen til $\boldsymbol{y}$ kan være større enn forventningen. Justeringen kommer av at mens man i Poisson modellen forutsetter at $\varphi = 1$ så tillater kvasi-Poisson at φ er forskjellig fra 1. Kvasi-Poisson, og kvasi-likelihood generelt, tillater dermed overdispersjon i datasettet man undersøker.
I R kan man finne kvasi-likelihood for Poisson modellen direkte fra glm-funksjonen nevnt tidligere. Der vil φ bli estimert utifra observasjonene.

2.1.3 Negativ Binomisk regresjon

Negativ Binomisk fordeling er i Venables og Ripley (2002) uttrykt ved

$$\mathbb{P}(Y = y; \varsigma, \mu) = \frac{\Gamma(y+\varsigma)}{\Gamma(\varsigma)\, y!} \frac{\mu^y\, \varsigma^\varsigma}{(\mu+\varsigma)^{y+\varsigma}};\ y = 0,1,2,\ldots;\ \varsigma > 0;\ \mu \geq 0\,. \qquad (2.1)$$

Denne fordelingen kan utledes ved å se på den marginale fordelingen til Y når $Y|\gamma \sim \text{Poisson}(\mu\gamma)$ og $\gamma\varsigma \sim \text{Gamma}(\varsigma,1)$. Her er γ en tilfeldig variabel. Da vil Y få marginalfordelingen ovenfor.
Forventning og varians til denne fordelingen er gitt ved $\mathbb{E}[Y] = \mu$ og $\text{Var}(Y) = \mu + \frac{\mu^2}{\varsigma}$. Den Negativt Binomiske fordelingen tillater større spredning enn det Poisson fordelingen gjør siden $\text{Var}(Y) > \mathbb{E}[Y]$. Negativ Binomisk regresjon er derfor et godt alternativ for datasett der det er overdispersjon, hvor ς i modellen indikerer graden av overdispersjon. Når ς er kjent er NB(μ, ς) en GLM. Men jeg vil for denne modellen behandle ς som en vanlig parameter.
I resten av masteroppgaven vil jeg ofte bruke "NB"som en forkortelse for Negativ Binomisk modell.

Lawless (1987) ser på modellen NB(μ,a) hvor $a = \frac{1}{\varsigma}$. Når $a \to 0$ vil NB(μ,a) i grensen omgjøres til Poisson(μ). Poisson modellen er dermed et spesialtilfelle av den Negativt Binomiske modellen.

I R kan man bruke funksjonen glm.nb til å lage Negativt Binomiske regresjonsmodeller. Funksjonen kommer fra pakken MASS som ble introdusert av Venables og Ripley (2002).
I glm.nb kan man tilegne ς en egen verdi. Hvis man ikke gjør det, vil R-funksjonen bestemme en verdi for ς ved å bruke en moment estimator fra en Poisson GLM. Skaleringsparameteren φ antas å være lik 1.
I glm.nb kan man velge om link-funksjonen skal være identitet, kvadratrot eller logaritmen. Jeg har valgt å bruke logaritmen. Dette fører til at forventningen til regresjonsmodellen blir på samme form som for Poisson modellen. Det er da lettere å sammenligne estimerte parametre og standardfeil for de 2 modellene.

2.2 Nullforhøyde modeller

Nullforhøyde regresjonsmodeller er et godt alternativ når man har datasett med flere null-observasjoner enn de klassiske regresjonsmodellene greier å fange opp. Mine referanser her er hovedsakelig Cameron og Trivedi (1998) og Zeileis *et al.* (2008).
En nullforhøyd regresjonsmodell tar utgangspunkt i at null-observasjonene kommer fra mer enn en plass. Hvis man for eksempel spør mennesker i Bergen om de har sett en fotballkamp den siste uken får man null-observasjoner fra 2 plasser. Null-observasjonene kommer fra dem som aldri ser på fotballkamper, og fra dem som ser på fotballkamper men som ikke gjorde det den uken.

En nullforhøyd diskret sannsynlighetsfordeling er en blanding av en Bernoulli koeffisient, ϑ, på punktet 0 og en diskret sannsynlighetsfordeling f. For responsvariabler $\boldsymbol{Y}$ forutsetter man at

$$\begin{aligned}
&Y_i = 0 \text{ med sannsynlighet } \vartheta_i \text{ ,} \\
&Y_i \sim f_{Y_i}(y) \text{ med sannsynlighet } (1-\vartheta_i) \text{ ,} \\
&\mathbb{P}(Y_i = y) = \vartheta_i \cdot 1_{(y=0)} + (1-\vartheta_i) \cdot f_{Y_i}(y); \;\; i = 1,\ldots,n; \; y = 0,1,\ldots \text{ .}
\end{aligned} \tag{2.2}$$

Man kan velge om ϑ skal være en konstant, dvs. en enslig parameter, eller avhenge av en eller flere kovariater(slik at ϑ får indeks). Formelt vil ϑ avhenge av kovariatmatrisen $\boldsymbol{z}$ og parametervektoren γ gjennom den kanoniske link-funksjonen $H(\vartheta_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}$.
Fordelingsfunksjonen f avhenger av kovariatmatrisen $\boldsymbol{X}$ og parametervektoren $\boldsymbol{\beta}$. Jeg vil bare se på tilfellene hvor f har Poisson eller Negativt Binomisk fordeling. Her er μ_i forventningsparameteren til fordelingen f. Da blir forholdet mellom μ_i, $\boldsymbol{X}_i$ og $\boldsymbol{\beta}$ bestemt gjennom den kanoniske link-funksjonen $g(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta}$. Jeg vil heretter kalle nullforhøyd Poisson for ZIP og nullforhøyd Negativ Binomisk ZINB.

I sammenheng med skipsforsikring kan man se på null-observasjoner som at noen skip er "sikre"og vil aldri komme utfor ulykker, og det vil dermed komme null-observasjoner fra disse skipene. Andre skip er "usikre", og skadefrekvensen til disse skipene vil følge fordelingsfunksjonen f. Man kan dermed se på koeffisienten ϑ som en størrelse som utifra et skips kovariater bestemmer hvor sannynlig det er at det aktuelle skipet er "sikkert"

Generelt har jeg 2 sett med parametre som må beregnes, $\boldsymbol{\beta}$ og $\boldsymbol{\gamma}$. For Negativ Binomisk fordeling vil parameteren ς bli behandlet som en støyparameter. Støyparametre blir nærmere forklart i neste seksjon.

For ZIP og ZINB får jeg følgende forventning for Y_i :

$$\mathbb{E}[Y_i] = \vartheta_i \cdot 0 + [1 - \vartheta_i]\mu_i = [1 - H^{-1}(\boldsymbol{z}_i'\boldsymbol{\gamma})]g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}), \;\; i = 1, \ldots, n \, . \tag{2.3}$$

Variansen til ZIP og ZINB blir

$$\begin{aligned} \text{ZIP: } \operatorname{Var}(Y_i) &= (1 - \vartheta_i)\mu_i(1 + \vartheta_i\mu_i) \geq \mathbb{E}[Y_i], \\ \text{ZINB: } \operatorname{Var}(Y_i) &= (1 - \vartheta_i)\mu_i\left(1 + \mu_i\left(\vartheta_i + \frac{1}{\varsigma}\right)\right) \geq \mathbb{E}[Y_i]. \end{aligned}$$

Fra ligningene for variansen ser man at ZIP og ZINB tillater overdispersjon.

For en ZIP er sannsynlighetstettheten gitt ved

$$\begin{aligned} \mathbb{P}(Y_i = 0) &= \vartheta_i + [1 - \vartheta_i]e^{-\mu_i} \, , \\ \mathbb{P}(Y_i = r) &= [1 - \vartheta_i]\frac{e^{-\mu_i}\mu_i^r}{r!}, \; r = 1,2,\ldots; \;\; i = 1, \ldots, n \, . \end{aligned}$$

Lambert (1992) introduserte ZIP hvor $\mu_i = \mu(\boldsymbol{X}_i, \boldsymbol{\beta})$ og ϑ_i er parametrisert via en logit link-funksjon av kovariatene $\boldsymbol{z}_i$. ϑ_i blir da på formen $\vartheta_i = \frac{\exp(\boldsymbol{z}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{z}_i'\boldsymbol{\gamma})}$ hvor $\boldsymbol{\gamma}$ er parametervektoren tilhørende $\boldsymbol{z}_i$. Fordelen med denne link-funksjonen er at ϑ_i alltid blir positiv. ZIP ble diskutert i artikler før 1992, men Lambert (1992) var den første som lot ϑ avhenge av kovariater.

Det kan virke naturlig at man skal være forsiktig med å ha de samme kovariatene i $\boldsymbol{z}_i$ og $\boldsymbol{X}_i$ på grunn av multikollinearitet(multikollinearitet blir forklart i avsnitt 2.4). Men multikollinearitet i nullforhøyde modeller er ikke noe tema i litteraturen jeg har tatt utgangspunkt i (Hall (2000), Lambert (1992), Cameron og Trivedi (1998) og Zeileis *et al.* (2008)). Lambert (1992) hevder at i utgangspunktet kan $\boldsymbol{z}_i = \boldsymbol{X}_i$ for ZIP hvis man ønsker det. Hall (2000) hevder, også for ZIP, at for en kovariat som forekommer i både $\boldsymbol{z}_i$ og $\boldsymbol{X}_i$ så vil hypotesetester si hvorvidt kovariatens $\boldsymbol{\beta}$ og $\boldsymbol{\gamma}$ verdier er meningsfulle eller ikke.

Det viktigste er at man velger kovariater slik at de estimerte parametrene gir mening i forhold til dataene man har. Så kan man eventuelt undersøke om det

er multikollinearitet mellom kovariatene i z_i og X_i hver for seg.

I R kan man modellere nullforhøyde regresjonsmodeller ved å bruke funksjonen zeroinfl fra pakken pscl. Dette er en forholdsvis ny pakke, så zeroinfl er ikke like anvendelig som glm og glm.nb. Man kan velge om sannsynlighetsfordelingen skal være Poisson, Negativt Binomisk eller Geometrisk.
For zeroinfl vil standard utforming til ϑ være en Bernoulli koeffisient med logit link-funksjon. Jeg har brukt denne utformingen på mine data, men man har flere andre muligheter i zeroinfl. Det er mulig å få litt bedre tilpasning til datasettet for noen av modellene ved å bruke en annen link-funksjon, men den optimale link-funksjonen varierer avhengig av hvilken modell det er snakk om. Derfor bruker jeg logit på alle de nullforhøyde modellene, slik at det blir lettere å sammenligne resultatene. For μ bruker jeg logaritmen som link-funksjon.

2.3 Likelihood teori, AIC og BIC

Her vil jeg først gå gjennom litt flerparameter likelihood-teori, for så å forklare AIC og BIC. Referansen her er hovedsakelig Pawitan (2001).

Definisjon 2.3.1: Likelihood funksjon

For observasjoner $y_1, y_2, \ldots, y_n$ fra en fordeling $f_Y(y;\boldsymbol{\theta})$, hvor $\boldsymbol{\theta} \in \mathbb{R}^p$, er likelihood funksjonen $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{Y_i}(y_i;\boldsymbol{\theta})$. Her antar jeg at observasjonene er uavhengige.
Log likelihood funksjonen er logaritmen til $L(\boldsymbol{\theta})$, dvs. $\log(L(\boldsymbol{\theta}))$.

Teorem 2.3.2: Konsistens

La $\boldsymbol{Y_n} = (Y_1, \ldots, Y_n)$, og la $\boldsymbol{W_n} = W_n(\boldsymbol{Y_n})$ være en estimator for $\boldsymbol{\theta}$. Da er $\boldsymbol{W_n}$ konsistent for $\boldsymbol{\theta}$ hvis

$$\boldsymbol{W_n} \xrightarrow{p} \boldsymbol{\theta} \text{ når } n \to \infty \, .$$

Dette vil si at $\mathbb{P}(|\boldsymbol{W_n} - \boldsymbol{\theta}| > \epsilon) \to 0$ for $\epsilon > 0$ når $n \to \infty$.

Definisjon 2.3.3: Score funksjon

Forutsatt at $\log(L(\boldsymbol{\theta}))$ er differensierbar er score funksjonen definert som

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log(L(\boldsymbol{\theta})) \, .$$

Score funksjonen er en vektor for fordelinger med flere parametre. SME til $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, finner man ved å sette ligningen ovenfor lik 0. Det er ikke alltid at SME kan beregnes analytisk, og man må da prøve å finne en numerisk løsning.

Definisjon 2.3.4: Fisher informasjon

Fisher informasjonen er en matrise definert som

$$I(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log(L(\boldsymbol{\theta}))\ .$$

Fisher informasjonen er den hessiske matrisen til log likelihood funksjonen med negativt fortegn.

Teorem 2.3.5

Anta at $\hat{\boldsymbol{\theta}}$ er en konsistent estimator for $\boldsymbol{\theta}$. Under visse betingelser vil den asymptotiske fordelingen til $\hat{\boldsymbol{\theta}}$ være gitt ved

$$I(\hat{\boldsymbol{\theta}})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_p(0, \boldsymbol{I}_p)\ .$$

Her er $\boldsymbol{I}_p$ identitetsmatrisen med dimensjon $p \times p$.

For beregningene i denne masteroppgaven vil jeg anta at SME, $\hat{\boldsymbol{\theta}}$, er en konsistent estimator av $\boldsymbol{\theta}$. Dette fører til jeg kan estimere standardfeilen til $\hat{\theta}_j$ med $\text{se}(\hat{\theta}_j) = \sqrt{I^{jj}}$ hvor I^{jj} er ledd j i diagonalen til $I(\hat{\boldsymbol{\theta}})^{-1}$.

Definisjon 2.3.6: Profil Likelihood

La $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ hvor $\boldsymbol{\theta}_1 \in \mathbb{R}^q$, $\boldsymbol{\theta}_2 \in \mathbb{R}^r$ og $p = q + r$. Profil likelihood til $\boldsymbol{\theta}_1$ blir da

$$L(\boldsymbol{\theta}_1) = \max_{\boldsymbol{\theta}_2} L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2(\boldsymbol{\theta}_1)),$$

hvor maksimeringen av $\boldsymbol{\theta}_2$ skjer ved at $\boldsymbol{\theta}_1$ holdes fast. Her behandles $\boldsymbol{\theta}_2$ som en vektor med støyparametre, d.v.s. en parametervektor uten umiddelbar interesse, men som må beregnes for å kunne analysere $\boldsymbol{\theta}_1$.

Å sammenligne statistiske modeller er vanligvis lett hvis modellene er nøstede. Nøstede modeller er f.eks.

$$\text{Modell 1: } \mu = \beta_0 + \beta_1 X_1 \,,$$
$$\text{Modell 2: } \mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \,,$$

hvor begge modellene tilhører samme statistiske fordeling. Modellene er nøstet siden Modell 1 er et spesialtilfelle av Modell 2. Å velge hvilke av modellene som er best her er ekvivalent med hypotesetesting av parametrene.
I tilfeller der man vil sammenligne ikke-nøstede modeller er AIC veldig nyttig, f.eks. hvis de 2 modellene ovenfor hadde hatt ulik fordeling.

AIC, eller Akaike Information Criterium, ble introdusert av Akaike i 1971. AIC til en parametervektor $\boldsymbol{\theta}$ er definert som

$$\text{AIC}(\boldsymbol{\theta}) = -2\log(L(\hat{\boldsymbol{\theta}})) + 2p \,.$$

Her er $\hat{\boldsymbol{\theta}}$ SME til $\boldsymbol{\theta}$ estimert fra et datasett med n observasjoner. L er likelihood funksjonen og p er dimensjonen til $\boldsymbol{\theta}$.
AIC er ikke et mål for å teste modeller, men et verktøy for å sammenligne modeller. Fremgangsmåten er enkel; modellen med lavest AIC vinner. $2p$-leddet i AIC straffer log likelihood funksjonen for hvor mange parametere det er i modellen.
Hvis 2 modeller er nøstede så vil modellen med flest parametre alltid gi størst $\log(L(\hat{\boldsymbol{\theta}}))$, selv om de ekstra parametrene kan være meningsløse. Derfor er det bedre å bruke AIC enn $\log(L(\hat{\boldsymbol{\theta}}))$ når man skal sammenligne nøstede modeller.
I R ligger AIC som en egen funksjon. Det er uproblematisk å bruke den på modellene jeg så langt har nevnt. AIC er ikke definert for kvasi-likelihood siden kvasi-likelihood ikke har noen likelihood funksjon.

Nå vil jeg gi en teoretisk begrunnelse for å bruke AIC. La $y_1, \ldots, y_n$ være et iid. utvalg fra en ukjent fordeling g og anta at man vil tilpasse en fordeling $f(y;\boldsymbol{\theta})$ til observasjonene, dvs. estimere $\boldsymbol{\theta}$. Hvis man bruker SME så maksimerer man

$$\frac{1}{n} l(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \log(f(y_i;\boldsymbol{\theta})) \xrightarrow{p} \mathbb{E}_g[\log(f(y;\boldsymbol{\theta}))],\ n \to \infty,$$

på grunn av store talls lov.
La så $\hat{\boldsymbol{\theta}}_k$ være SME til modellen f_k. Man er ute etter en modell som maksimerer $\mathbb{E}[\log(f_k(S, \boldsymbol{\theta}_k))]$ hvor $S \sim g$. Men siden $\boldsymbol{\theta}_k$ er ukjent så velger man istedet modellen som maksimerer $Q_k = \mathbb{E}[\log(f_k(S, \hat{\boldsymbol{\theta}}_k))]$. Her er forventningen tatt mhp. S og $\hat{\boldsymbol{\theta}}_k$, og man forutsetter at disse størrelsene er uavhengige av hverandre.
La $\boldsymbol{\theta}_{k0}$ være løsningen til $\frac{\partial}{\partial \boldsymbol{\theta}_k}\lambda(\boldsymbol{\theta}_k) = 0$ hvor $\lambda(\boldsymbol{\theta}_k) = \mathbb{E}[\log(f_k(S, \boldsymbol{\theta}_k))]$. $\boldsymbol{\theta}_{k0}$ er parametervektoren som blir estimert av $\hat{\boldsymbol{\theta}}_k$. La så

$$\boldsymbol{J}_k = \mathbb{E}\left[\frac{\partial \log(f_k(S, \boldsymbol{\theta}_k))}{\partial \boldsymbol{\theta}_k} \frac{\partial \log(f_k(S, \boldsymbol{\theta}_k))}{\partial \boldsymbol{\theta}_k'}\right]\Bigg|_{\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k0}},$$
$$\boldsymbol{\mathcal{I}}_k = -\mathbb{E}\left[\frac{\partial^2 \log(f_k(S, \boldsymbol{\theta}_k))}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k'}\right]\Bigg|_{\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k0}}.$$

Pawitan (2001) gir at $\boldsymbol{J}_k = \boldsymbol{\mathcal{I}}_k$ hvis f_k er den sanne modellen, dvs. hvis $f_k = g$. Ved å bruke Taylor utvikling av $\log(L(\boldsymbol{\theta}_k))$ og $\lambda(\boldsymbol{\theta}_k)$ kan man vise at

$$nQ_k \approx \mathbb{E}[\log(L(\hat{\boldsymbol{\theta}}_k))] - \text{trace}(\boldsymbol{J}_k \boldsymbol{\mathcal{I}}_k^{-1}).$$

Utifra dette befester Pawitan (2001) at høyre side av ligningen over er asymptotisk forventningsrett for nQ_k. AIC formelen er basert på å forutsette at $\boldsymbol{J}_k \approx \boldsymbol{\mathcal{I}}_k$ slik at $\text{trace}(\boldsymbol{J}_k \boldsymbol{\mathcal{I}}_k^{-1}) \approx p$. AIC er dermed et estimat av $-2nQ_k$. Å minimere AIC kan dermed sammenlignes med å maksimere Q_k.
I praksis kan den virkelige verdien til $\text{trace}(\boldsymbol{J}_k \boldsymbol{\mathcal{I}}_k^{-1})$ være veldig forskjellig fra p slik at AIC ikke er en god estimator av $-2nQ_k$. Men dette gjør ikke AIC ugyldig som vurderingskriterium for modeller. Selv om man i denne begrunnelsen forutsetter at $y_1, \ldots, y_n$ skal være iid. så gjelder AIC formelen også utenfor denne forutsetningen.

Det finnes alternativer til AIC, det mest kjente er Bayes Information Criterium(BIC). BIC er definert ved

$$\text{BIC}(\boldsymbol{\theta}) = -2\log(L(\hat{\boldsymbol{\theta}})) + p\log(n)\,.$$

BIC fungerer på samme måte som AIC, ved at modellen med minst BIC "vinner". Men BIC slår hardere ned på mange parametre enn det AIC gjør. I denne masteroppgaven vil jeg bruke AIC til å velge kovariater for modellene, og AIC og BIC til å sammenligne de ulike modellene.

2.4 VIF

Multikollinearitet defineres som når det er høy korrelasjon mellom to eller flere av forklaringsvariablene i en regresjonsmodell. Den ideelle situasjonen er at man har en samling av uavhengige kovariater som er høyt korrelert med responsvariabelen, men lite korrelert med hverandre. Selv om det ikke er noen høye korrelasjonverdier mellom de enkelte kovariatene så kan det fremdeles være multikollinearitet.
En veldig høy grad av multikollinearitet kan føre til at $\boldsymbol{X}'\boldsymbol{X}$ ikke er invertibel. Da er det ikke mulig å estimere noen av koeffisientene i en regresjonsmodell. I tillegg hevder Demidenko (2004) at AIC fungerer dårlig som vurderingskriterium hvis det er multikollinearitet i en eller flere av modellene man undersøker. Til å undersøke graden av multikollinearitet i en modell kan man bruke Variance inflation factor(VIF). Jeg kommer ofte til å ha så mange som 6 kovariater i modellene mine, så det vil være nødvendig å undersøke om det er multikollinearitet mellom kovariatene jeg bruker.

Definisjon 2.4.1: Fremgangsmåte og definisjon til VIF:

Referansen her er Fox og Monette (1992).
Jeg tar utgangspunkt i en lineær modell på formen

$$\begin{aligned} y_i =& \beta_0 + X_{i,1}\beta_1 + \cdots + X_{i,p-1}\beta_{p-1} + \epsilon_i;\ i = 1, \ldots, n \\ \epsilon_i \sim& \mathcal{N}(0, \sigma^2)\ , \end{aligned}$$

hvor $\boldsymbol{y}$ er en vektor med responsvariabler, $\boldsymbol{X}$ er en matrise med kovariater og $\boldsymbol{\epsilon}$ er en vektor med feilledd.

Jeg antar nå at jeg har estimater for parametervektoren $\boldsymbol{\beta}$. De estimerte verdiene for $\boldsymbol{y}$ blir da

$$\hat{y}_i = \hat{\beta}_0 + X_{i,1}\hat{\beta}_1 + \cdots + X_{i,p-1}\hat{\beta}_{p-1},\ i = 1, \ldots, n\ .$$

Jeg introduserer R^2 som er gitt ved.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\ .$$

VIF er en verdi som kan beregnes for hvert av estimatene $\hat{\beta}_j$, $j = 1, \ldots, p-1$. For å beregne VIF($\hat{\beta}_j$) gjør man følgende:

1 Tilpass en lineær regresjonsmodell til kovariaten $\boldsymbol{X}_j$ basert på de andre kovariatene: $X_{i,j} = \alpha_0 + X^*_{i,1}\alpha_1 + \cdots + X^*_{i,p-2}\alpha_{p-2} + \epsilon_i,\ i = 1, \ldots, n$. $\boldsymbol{X}^*$ er kovariatmatrisen $\boldsymbol{X}$ uten kovariaten $\boldsymbol{X}_j$.

2 Finn estimater til parametervektoren $\boldsymbol{\alpha}$ ved hjelp av minste kvadraters metode. Jeg har da $\hat{X}_{i,j} = \hat{\alpha}_0 + X^*_{i,1}\hat{\alpha}_1 + \cdots + X^*_{i,p-2}\hat{\alpha}_{p-2},\ i = 1, \ldots, n$.

3 VIF til $\hat{\beta}_j$ kan da beregnes ved

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_j^2},\quad R_j^2 = \frac{\sum_{i=1}^n (\hat{X}_{i,j} - \bar{X}_j)^2}{\sum_{i=1}^n (X_{i,j} - \bar{X}_j)^2}\ .$$

Hvis $\hat{\boldsymbol{X}}_j$ er et godt estimat av $\boldsymbol{X}_j$ så blir $\text{VIF}(\hat{\beta}_j)$ stor. Da kan $\boldsymbol{X}_j$ forklares vha. de andre kovariatene og det er grunn til å tro at det er multikollinearietet mellom kovariatene.
Denne fremgangsmåten kan også brukes på GLM siden man ovenfor bare studerer forholdet mellom de ulike kovariatene.

Det er ikke fastsatt noen eksakte grenser for når VIF-verdier er høye, men en tommelfingerregel pleier å være at det er stor grad av multikollinearitet hvis VIF>10(Marquardt (1970)).
VIF-verdiene blir vanligvis litt endret hvis man tar med en eller flere motvekter i modellene man undersøker. For mine modeller ble disse forskjellene så små at jeg ser helt bort fra dem.
I R ligger VIF som en funksjon i pakken "car". Denne funksjonen er kompatibel med bl.a. funksjonene lm, glm og glm.nb, men ikke med zeroinfl.

2.5 Estimerte verdier

Jeg vil nå bruke regresjonsmodellene jeg har definert på datasettet mitt. Hvis man vil ha en nærmere forklaring på hvordan modellene brukes i R vil jeg anbefale artikkelen Regression Models for Count Data in R av Zeileis *et al.* (2008). I alle modellene bruker jeg log(Days.covered) som motvekt. Ved å gjøre dette blir AIC og BIC litt lavere. Motvekten blir lagt til i den lineære prediktoren til hver modell.
I alle tabeller over estimerte parametre i denne masteroppgaven vil standardfeil stå i parentes bak de estimerte parameterverdiene.

2.5.1 Tankskip

Jeg ser nå bare på skip av typen tankskip.

Antall krav, Claim number, er responsvariabelen i en regresjonsmodell. Jeg lar Age, GT, Sum.Insured, HP.prime.mover og Stroke være kovariater, mens Days.covered blir brukt som motvekt. HP.prime.mover vil heretter bli forkortet til "HP". I tillegg bruker jeg Intercept som kovariat. Intercept er en vektor hvor alle elementene i vektoren er like 1.
Det er veldig stor spredning av verdier i kovariatene ovenfor. Jeg vil derfor heller bruke logaritmen til GT, Sum.Insured og HP for å minske spredningen, mens jeg bruker log(Age+2) istedet for Age. Tabellen nedenfor viser at spredningen da blir betydelig redusert. I resten av masteroppgaven vil jeg alltid bruke log-verdiene hvis jeg skal ha GT, Sum.Insured, HP eller Age som kovariater.

Kolonne	Min.	Gj.snitt	Maks
Sum.Insured	1 040 000	26 630 000	218 500 000
Age	0.00	9.59	40.50
HP	600	15 210	60 040
GT	348	48 700	235 000
Stroke	0.00	0.11	1.00
log(Sum.Insured)	13.85	16.81	19.20
log(Age+2)	0.69	2.23	3.75
log(HP)	6.39	9.45	11.00
log(GT)	5.85	10.32	12.37

Tabell 2.1: Oversikt over kovariater.

For tankskip er det høy korrelasjon mellom log(GT) og log(HP), noe bl.a VIF-funksjonen bekrefter. Jeg bruker derfor ikke log(HP) som kovariat i modellene. For nullforhøyd Poisson(ZIP) modellen lar jeg ϑ avhenge av kovariatene log(Sum.Insured), log(GT) og log(Age+2) i tillegg til Intercept. Jeg får følgende verdier for modellene jeg har nevnt sålangt:

	Poisson	Kvasi-Poisson	NB	ZIP
Verdier for β:				
Intercept	-4.73(0.92)	-4.73(1.01)	-4.77(0.97)	-13.10(1.68)
log(Age+2)	0.29(0.05)	0.29(0.06)	0.29(0.06)	0.63(0.14)
log(GT)	-0.15(0.04)	-0.15(0.04)	-0.15(0.04)	-0.30(0.09)
log(Sum.Insured)	0.19(0.06)	0.19(0.06)	0.20(0.06)	0.77(0.12)
Stroke	0.25(0.09)	0.25(0.10)	0.26(0.10)	0.29(0.10)
$\hat{\varsigma}$			1.06(0.20)	
Verdier for γ:				
Intercept				-23.88(4.90)
log(Age+2)				0.86(0.36)
log(GT)				-0.37(0.21)
log(Sum.Insured)				1.50(0.32)
Antall parametre	5	5	6	9
AIC	8 210.4	Eksisterer ikke	8 168.3	8 163.7
BIC	8 247.5	Eksisterer ikke	8 212.8	8 230.5

Tabell 2.2: Estimerte verdier for tankskip.

Dispersjonsparameteren for kvasi-Poisson, φ, ble beregnet til å være 1.199. Standardfeilene til kvasi-Poisson modellen ble såpass lik standardfeilene til Poisson modellen at det ikke førte til endringer i valget av kovariater.

2.5.2 Hele datasettet

Jeg ser nå på alle typer skip. For disse observasjonene er det ikke lenger høy korrelasjon mellom log(GT) og log(HP). Jeg kan nå bruke alle 5 kovariatene i regresjonsmodellene uten at VIF gir uttrykk for multikollinearitet. Det viser seg også at bruk av alle kovariatene gir minst AIC.

	Poisson	Kvasi-Poisson	NB	ZIP
Verdier for β:				
Intercept	-4.32(0.30)	-4.32(0.32)	-4.39(0.32)	-8.89(0.44)
log(Age+2)	0.15(0.02)	0.15(0.02)	0.15(0.02)	0.17(0.02)
log(GT)	-0.11(0.02)	-0.11(0.02)	-0.12(0.02)	
log(Sum.Insured)	0.07(0.02)	0.07(0.02)	0.08(0.02)	0.40(0.02)
log(HP)	0.15(0.02)	0.15(0.03)	0.16(0.02)	
Stroke	0.42(0.03)	0.42(0.03)	0.42(0.03)	0.47(0.03)
$\hat{\varsigma}$			1.08(0.09)	
Verdier for γ:				
Intercept				-14.13(1.19)
log(GT)				0.48(0.06)
log(Sum.Insured)				0.89(0.08)
log(HP)				-0.63(0.09)
Antall parametre	6	6	7	8
AIC	38 225.0	Eksisterer ikke	37 997.0	37 923.8
BIC	38 278.3	Eksisterer ikke	38 059.2	37 994.9

Tabell 2.3: Estimerte verdier for alle typer skip.

For kvasi-Poisson ble dispersjonsparameteren, φ, beregnet til å være 1.138. I denne tabellen og tabellen for tankskip ser man at standardfeilene til kvasi-poisson modellen er like store eller litt større enn standardfeilene til Poisson modellen.
For begge datasettene viser verdien til AIC og BIC at Negativ Binomiske modeller gir bedre tilpasning enn Poisson, og at Nullforhøyd Poisson modell gir best tilpasning av modellene. Parameterestimater til Nullforhøyd Negativ Binomisk(ZINB) modell blir oppgitt i kapittel 3.4, det ble ikke plass til alle modellene i tabellene ovenfor.

3
Modeller for skadebeløp

I dette kapittelet vil jeg fokusere på skadebeløpene og egenandelene til datasettet. Jeg vil også undersøke om det er gunstig å ta hensyn til egenandelene når jeg skal undersøke skadefrekvensen til hvert enkelt skip. Skadebeløp ansees som kontinuerlige data, og det er derfor naturlig å se på lineære modeller.

3.1 Lineær og lineært mikset modell

Referansen her er McCulloch og Searle (2001).
En grunnleggende lineær modell er på formen $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Her er $\boldsymbol{Y}$ en vektor bestående av responsvariabler med kovariansmatrise $\boldsymbol{R}$, $\boldsymbol{X}$ er en kovariatmatrise og $\boldsymbol{\beta}$ er en parametervektor. Det er n variabler i $\boldsymbol{Y}$ og p elementer i $\boldsymbol{\beta}$, mens $\boldsymbol{X}$ har n rader og p kolonner.
Hvis man forutsetter at variansen til hvert element i $\boldsymbol{Y}$ er den samme, og at kovariansen mellom hvert par av elementer er 0, så kan man uttrykke variansen som $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$. $\boldsymbol{I}_n$ er her en identitetsmatrise.
For lineære modeller forutsetter man at $\boldsymbol{Y}$ er multinormalfordelt, dvs. på formen

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim \mathcal{N}_n(0,\, \boldsymbol{R}). \tag{3.1}$$

Parametrene $\boldsymbol{\beta}$ og σ^2 kan beregnes ved SME. $\boldsymbol{\beta}$ kan også beregnes ved minste kvadraters metode, noe som vil gi samme estimat som SME.
For lineære modeller med normalfordelte feilledd trenger man ikke ta med motvekter i modellene. Man kan bare trekke motvektene fra verdiene til responsvariablene og bruke residualene istedet for y-verdiene. Opprinnelig har man forventningsvektoren med elementer $\mu_i = \boldsymbol{X}_i'\boldsymbol{\beta}$, men man ønsker nå å arbeide med $\mu_i^\star = \boldsymbol{X}_i'\boldsymbol{\beta} + \eta_i$ hvor η er motvekten til modellen. Ifølge den lineære modellen vil $Y_i \approx \mu_i^\star$, så man kan dermed flytte om på ligningen og heller arbeide med $\mu_i \approx Y_i - \eta_i = Y_i^\star$.
I R kan man beregne lineære modeller ved å bruke funksjonen lm.

En ren random effect modell, dvs. en modell som bare består av tilfeldige effekter, er på formen $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{Z}\boldsymbol{u}$. Her er $\boldsymbol{Y}$ igjen en vektor av responsvariabler og $\boldsymbol{Z}$ er en kovariatmatrise med n rader og q kolonner. Men nå er $\boldsymbol{u}$ en vektor bestående av tilfeldige variabler $u_1, \ldots, u_q$ hvor $\mathbb{E}(\boldsymbol{u}) = \boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{B}$ d.v.s. $\boldsymbol{B}$ er en kovariansmatrise for elementene i $\boldsymbol{u}$. For en slik modell er jeg bare interessert i hvor mye $\boldsymbol{Y}$ varierer, siden forventningen til $\boldsymbol{Y}$ marginalt er definert til å være lik 0.

En grunnleggende forskjell mellom en random effect modell og en vanlig lineær modell er at observasjonene i en random effect modell ikke er uavhengige. Hvis jeg har en grunn til å tro at observasjonene kommer fra m ulike grupper, hvor observasjoner innen hver gruppe er korrelerte, så kan jeg innføre tilfeldige effekter. Jeg har da grunn til å tro at observasjonene i en gruppe er blitt påvirket forskjellig av ytre faktorer enn observasjonene i en annen gruppe. Man gir da observasjonene to indekser istedet for en: $Y_{1,1}, \ldots, Y_{1,n_1}, \ldots, Y_{m,1}, \ldots, Y_{m,n_m}$. Her brukes den første indeksen til å skille de ulike gruppene fra hverandre og den andre til å skille mellom de ulike observasjonene i hver gruppe.
Forutsetningen om tilfeldige faktorer kan være en praktisk måte å innføre en struktur for varians-kovarians i modellen. Blant annet tillater modeller med tilfeldige effekter overdispersjon blant observasjonene man undersøker.
I den "rene"random effect modellen vil det være q tilfeldige effekter som påvirker kovariatmatrisen. Disse effektene endrer seg ikke for kovariater i samme gruppe, men de er forskjellige fra en gruppe til en annen. Modellen

bør derfor egentlig skrives på formen

$$\mathbb{E}[Y_{ij}|\boldsymbol{u}_i] = \boldsymbol{Z}'_{ij}\boldsymbol{u}_i\ ;\ \ i = 1,\ldots,m;\ \ j = 1,\ldots,n_i\ .$$

Vanligvis blir de ulike vektorene med tilfeldige effekter forutsatt til å være iid. I utregninger ser jeg derfor bort fra indeksen som skiller disse vektorene fra hverandre.

Hvis man har en lineær regresjonsmodell som består av både faste($\boldsymbol{\beta}$) og variable($\boldsymbol{u}$) komponenter så har man en lineært mikset modell. Jeg får her

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim \mathcal{N}_n(0,\boldsymbol{R}),\ \ \boldsymbol{u} \sim \mathcal{N}_q(0,\boldsymbol{B}), \\ &\text{hvor } \boldsymbol{Y} = (Y_{1,1},\ldots,Y_{1,n_1},\ldots,Y_{m,1},\ldots,Y_{m,n_m}).\end{aligned} \tag{3.2}$$

Jeg antar at $\boldsymbol{u}$ og $\boldsymbol{\epsilon}$ er uavhengige. Her er $\boldsymbol{u}, \boldsymbol{X}, \boldsymbol{\beta}$ og $\boldsymbol{Z}$ definert som tidligere. Men kolonnene(kovariatene) i $\boldsymbol{Z}$ er en delmengde av kovariatene i $\boldsymbol{X}$. Dette medfører at kovariatene i $\boldsymbol{X}$ som modellen tillater å variere tilfeldig med $\boldsymbol{u}$ blir bestemt av kovariatene i $\boldsymbol{Z}$. Her er $\boldsymbol{B}$ per definisjon symmetrisk og positiv definit. I tillegg får jeg at $\text{Var}(\boldsymbol{Y}|\boldsymbol{u}) = \boldsymbol{R}$ mens $\boldsymbol{Y}$ marginalt er gitt ved

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}\ ,\ \boldsymbol{R} + \boldsymbol{ZBZ}'). \tag{3.3}$$

Fordelingen gir at $\boldsymbol{u}$ bare påvirker $\boldsymbol{Y}$ marginalt gjennom variansen til $\boldsymbol{Y}$. Parametrene jeg må estimere i denne modellen er $\boldsymbol{\beta}$, elementene i $\boldsymbol{R}$ og elementene i $\boldsymbol{B}$. Under forutsetningen om at betinget kovarians mellom ulike elementer i $\boldsymbol{y}$ er lik 0 trenger jeg bare beregne σ^2 for $\boldsymbol{R}$.
Jeg vil ofte bruke forkortelsen lmm for lineært miksede modeller.

I denne masteroppgaven vil jeg bruke UW.year(kalles heretter tegningsår), dvs. årene skipene ble forsikret, som tilfeldig effekt. Dette vil si at jeg lar tegningsår være $\boldsymbol{u}$. Tegningsårene går fra 1995 til 2004, slik at kovariatene i $\boldsymbol{Z}$ varierer tilfeldig med de 10 årstallene. Grunnen til at jeg bruker tegningsår er at hvert år har sin egen kredibilitetsfaktor. Det virker naturlig at denne kredibilitetsfaktoren fører til at skip med samme tegningsår er korrelerte.
Når man gjør beregninger pleier de statistiske modellene å si fra hvis det er urimelig å bruke en spesifikk variabel som tilfeldig effekt. I R blir gjerne elementene i kovariansmatrisen $\boldsymbol{B}$ beregnet til å være lik 0 hvis man har valgt

en $\boldsymbol{u}$ som ikke passer til dataene.

For å beregne den lineært miksede modellen i R har jeg brukt funksjonen lmer i pakken lme4. Parametrene blir der beregnet ved REML (restricted maximum likelihood). Denne funksjonen kan også brukes for generaliserte linært miksede modeller som jeg skal se på senere.
Det har vist seg at log likelihood verdien som lmer oppgir for lmm ikke er direkte sammenlignbar med vanlige lineære modeller. Når jeg selv prøvde å beregne log likelihood verdien utifra parameterestimatene fra lmer så fikk jeg en høyere verdi enn det lmer oppgir. Derfor bruker jeg lmer til å beregne parameterestimater og standardfeil for lmm, mens jeg beregner AIC for lmm manuelt. For å beregne log likelihood i AIC bruker jeg en endimensjonal Laplace approksimasjon, en metode som blir forklart nærmere i kapittel 4.

3.1.1 Estimerte verdier

Jeg skal nå bruke en lineær modell og en lineær mikset modell på skadebeløpene i datasettet mitt. Den lineært miksede modellen er ikke sentral i masteroppgaven, men fungerer som en introduksjon til modeller med tilfeldige effekter.
For å kunne tilpasse modeller til dataene er skadebeløpene i datasettet dividert med forsikringssummen til hvert enkelt skip, slik at man ender opp med såkalt skadegrad. Dette er hensiktsmessig siden forsikringssummen, og dermed det aktuelle skadebeløpet, til hvert skip er veldig varierende. Skadegradene er forutsatt å være iid.

Skadegradene i datasettet er kolonnene Claim size 1 til Claim size 5. Siden dette er ulike kolonner vil jeg nå slå sammen skadegrader i de ulike kolonnene til en kolonne. Jeg får da en vektor med alle skadegradene uavhengig om det er den første, andre eller femte skaden til skipet. Dette vil føre til at skip med flere enn en skade blir representert mer enn skip med en skade, dvs. at verdiene til noen skips kovariater går igjen flere ganger i datasettet. I en forsikringssammenheng virker det naturlig at "skrøpelige"skip blir overrepresentert.

I det nye datasettet tar jeg bare med rapporterte skadegrader, dvs. at jeg fjerner alle observasjoner der antall skader, Claim number, er lik 0. Jeg tar heller ikke med skadegrader med verdi større enn 0.6. Disse er såkalte totalskader

som pleier å bli behandlet separert fra resten av datasettet med egne modeller. Jeg kommer ikke til å studere totalskader i denne masteroppgaven.

I denne seksjonen, og i den neste, vil jeg forutsette at skadegradene er lognormalt fordelt. Dette vil si at hvis $\boldsymbol{Y}$ er vektoren med skadegradene, så er elementene i $\boldsymbol{T} = \log(\boldsymbol{Y})$ normalfordelte. Forutsetningen er dermed at hvert enkelt element i $\boldsymbol{T}$ er gitt ved $T_i \sim \mathcal{N}(\boldsymbol{X}_i'\boldsymbol{\beta}, \sigma^2)$. Forventningen til Y blir da $\mathbb{E}(Y_i) = \exp(\boldsymbol{X}_i'\boldsymbol{\beta} + \frac{\sigma^2}{2});\ i = 1, \ldots, n$.
Forsøk med å la elementene i $\boldsymbol{Y}$ være normalfordelte ga både dårlige estimerte verdier og ustabile beregninger. Normalfordelingen kan være problematisk å bruke siden den er symmetrisk og korthalet.

Etter å ha slått sammen kolonnene for skadegrad i datasettet sitter jeg igjen med 5 877 observasjoner som kan brukes til å tilpasse modeller. For tankskip alene er det 1 221 observasjoner. Jeg tilpasser først lineære modeller til tankskip, for så å tilpasse modeller til alle typer skip.
For den lineært miksede modellen bruker jeg de samme faste kovariatene som for den lineære modellen. For de tilfeldige effektene i modellen har jeg ikke prøvd kombinasjoner hvor $\boldsymbol{Z}$ består av mer enn en kovariat og intercept. Dette gjør jeg for å holde modellene forholdsvis enkle, siden den lineært miksede modellen i denne oppgaven bare er en oppvarming til GLMM. Litt prøving og feiling viser at bare en kovariat i $\boldsymbol{Z}$ gir best tilpasning både for tankskip og alle typer skip. De lineært miksede modellene blir da på formen

$$\begin{aligned} &T_{ij} = \boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}u_i + \epsilon_{ij},\ \epsilon \sim \mathcal{N}(0, \sigma^2),\ \ u \sim \mathcal{N}(0, \psi^2), \\ &T_{ij} = \log(Y_{ij}),\ i = 1, \ldots, 10;\ \ j = 1, \ldots, n_i\ . \end{aligned}$$

Her er Z enten en konstant(intercept) eller en kovariat. Siden det bare er ett element i kovariansmatrisen $\boldsymbol{B}$ til de tilfeldige effektene kan jeg beskrive variansen for de tilfeldige effektene i modellen med parameteren ψ^2.

Tankskip

Også i det nye datasettet for tankskip er det høy korrelasjon mellom log(GT) og log(HP). For den lineære modellen bruker jeg log(HP) og log(Sum.Insured) som kovariater i $\boldsymbol{X}$. Det er denne kombinasjonen av kovariatene som gir lavest AIC og liten multikollinearitet. Jeg tar også med intercept i $\boldsymbol{X}$.

I den lineært miksede modellen bruker jeg de samme kovariatene i $\boldsymbol{X}$ som for den lineære modellen. I $\boldsymbol{Z}$ tar jeg bare med kovariaten log(Sum.Insured). Litt prøving og feiling viser at det er denne kovariaten som gir best tilpasning.

Tabellen nedenfor gir de estimerte verdiene for den lineære modellen og den lineært miksede modellen.

	Vanlig lineær modell	Lineært mikset modell
Intercept	6.42(0.53)	6.42(0.53)
log(Sum.Insured)	-0.86(0.04)	-0.86(0.04)
log(HP)	0.43(0.05)	0.43(0.05)
$\hat{\sigma}$	0.83	0.83
$\hat{\psi}$		0.001
Antall parametre	4	5
AIC	3 032.0	3 034.0
BIC	3 052.4	3 059.5

Tabell 3.1: Estimerte verdier for tankskip.

I dette tilfellet gir en vanlig lineær modell en bedre tilnærming enn en lineært mikset modell. Det er bare den ekstra parameteren for varians, ψ, som skiller modellene fra hverandre. Hvis jeg hadde tatt med flere desimaler i tabellen ovenfor ville man sett at estimatene er litt forskjellige, men det kommer nok av forskjellige beregningsmetoder.
Det kan hende at R beregner ψ til å være lik 0. Dette vil i så fall bety at modellen er bedre uten de tilfeldige effektene man har valgt. Man bør da enten endre kovariatene i $\boldsymbol{Z}$ eller bruke en vanlig lineær modell. For tankskip ovenfor er $\hat{\psi}$ veldig lav, samtidig som AIC og BIC er høyere for den lineært miksede modellen enn for den lineære modellen. Derfor virker det unødvendig å bruke en lineært mikset modell i dette tilfellet.

Alle typer skip

Jeg tar nå utgangspunkt i det modifiserte datasettet med 5 877 observasjoner. For den lineære modellen bruker jeg kovariatene Stroke, log(GT), log(Sum.Insured) og log(Age+2) i tillegg til intercept. Denne kombinasjonen gir lavest AIC, og VIF-funksjonen viser at det ikke er noe problem med multikollinearitet.

Etter prøving og feiling kom jeg frem til at log(Age+2) som eneste element i $\boldsymbol{Z}$ gir minst AIC for den lineært miksede modellen. Resultatene er gitt i tabellen under.

	Lineær modell 2	Lineært mikset modell 2
Intercept	6.60(0.28)	6.51(0.28)
Stroke	0.21(0.03)	0.21 (0.03)
log(GT)	0.24(0.01)	0.24(0.01)
log(Sum.Insured)	-0.80(0.01)	-0.80(0.01)
log(Age+2)	0.15(0.02)	0.15(0.02)
$\hat{\sigma}$	0.85	0.85
$\hat{\psi}$		0.03
Antall parametre	6	7
AIC	14 856.5	14 823.0
BIC	14 896.5	14 869.7

Tabell 3.2: Estimerte verdier for alle typer skip.

Her er AIC og BIC en del lavere for den lineært miksede modellen enn for den lineære modellen. Det er naturlig at ψ er større her enn for tankskip siden log(Sum.Insured) er betydelig større enn log(Age+2), dvs. $\boldsymbol{Z}$ er større for modellen til tankskip enn $\boldsymbol{Z}$ er for denne modellen. Tabell 2.1 på side 18 viser forskjellen mellom disse to kovariatene.

3.2 Lineær modell betinget mhp. egenandel

Hvert skip har en egenandel, i datasettet er det kolonnen Basic.Ded. Et skip får ikke dekket skader hos forsikringsselskapet hvis skadebeløpet er mindre enn skipets egenandel. Forsikringsselskapet mangler derfor data om størrelseordenen til de små skadene som aldri blir rapportert. Dataene jeg har for skadebeløp er derfor ufullstendige siden datasettet ikke inneholder skadebeløp mindre enn den respektive egenandelen. Dataene for skadebeløp går derfor under betegnelsen venstre-trunkerte observasjoner.
Jeg skal her tilpasse en modell som tar hensyn til at det ikke er registrert noen skadebeløp som er mindre enn skipenes respektive egenandeler. Egenandelene er også dividert med forsikringssummen til hvert skip slik at man kan sammenligne direkte med skadegrad. For å tilpasse modellen ser jeg på forde-

lingen til en skadegrad, Y_i, betinget på at skadegraden er større enn skipet sin egenandel, d_i. Jeg antar at skadegradene er uavhengig av egenandelene.

Det er lett å vise at den betingede tettheten er gitt ved

$$f_{Y_i|Y_i>d_i}(y_i) = \frac{f_{Y_i}(y_i)}{\bar{F}_{Y_i}(d_i)} = \frac{f_{Y_i}(y_i)}{1 - F_{Y_i}(d_i)} \; ; \; i = 1, \ldots, n. \tag{3.4}$$

Her er f sannsynlighetsfordelingen, F er den kumulative fordelingsfunksjonen og $\bar{F}$ er overlevelsesfunksjonen $\bar{F}_Y(d) = 1 - F_Y(d)$.
Som tidligere forutsetter jeg at Y_i har lognormal fordeling, så jeg lar $T_i = \log(Y_i)$ og $z_i = \log(d_i)$. Da blir den betingede tettheten

$$f_{T_i|T_i>z_i}(t_i) = \frac{f_{T_i}(t_i)}{1 - F_{T_i}(z_i)}, \; T_i \sim \mathcal{N}(\boldsymbol{X}_i'\boldsymbol{\beta}, \sigma^2) \; . \tag{3.5}$$

Jeg vil nå finne estimater for parametrene $\boldsymbol{\kappa} = (\beta_0, \cdots, \beta_{p-1}, \sigma)$ for denne betingede fordelingen. Dette kan gjøres vha. SME.
Likelihood og log likelihood for verdiene $T_1, \ldots, T_n$ blir

$$\begin{aligned} L(\boldsymbol{\kappa}; T_1, \ldots, T_n | T_1 > z_1, \ldots, T_n > z_n) = L(\boldsymbol{\kappa}) &= \prod_{i=1}^{n} \frac{f_{T_i}(t_i)}{1 - F_{T_i}(z_i)} \; , \\ l(\boldsymbol{\kappa}) = \log(L(\boldsymbol{\kappa})) &= \sum_{i=1}^{n} \Big(\log(f_{T_i}(t_i)) - \log(1 - F_{T_i}(z_i)) \Big). \end{aligned} \tag{3.6}$$

Det neste steget ville vært å derivere $l(\boldsymbol{\kappa})$ med hensyn på hver enkelt parameter, sette ligningene lik 0 og løse dem for å finne estimater for parametrene. Men for normalfordelingen kan jeg ikke gjøre dette analytisk. Derfor bruker jeg funksjonen optim i R til å finne estimater for $\boldsymbol{\kappa}$ numerisk.
I optim er det mulig å få beregnet et numerisk estimat av den hessiske matrisen til funksjonen man optimerer. Jeg kan derfor også finne standardfeil til estimatene vha. optim.

3.2.1 Estimerte verdier

Jeg bruker de samme kovariatene for denne modellen som jeg brukte for den vanlige lineære modellen. Når jeg skal bruke optim til å beregne parametre må jeg oppgi noen innledende verdier for parametrene. Siden jeg har de samme kovariatene som for den vanlige lineære modellen, lar jeg de innledende verdiene være de estimerte verdiene fra den lineære modellen.

For å finne standardfeil til de estimerte parametrene bruker jeg den estimerte hessiske matrisen.

For tankskip har jeg som før 2 kovariater i tillegg til intercept: log(Sum.Insured) og log(HP). For alle typer skip har jeg 4 kovariater i tillegg til intercept: Stroke, log(GT), log(Age+2) og log(Sum.Insured).

	Tankskip	Alle typer skip
Intercept	6.23(0.88)	6.56(0.47)
Stroke		0.26(0.04)
log(GT)		0.22(0.02)
log(Age+2)		0.10(0.03)
log(Sum.Insured)	-0.83(0.07)	-0.80(0.03)
log(HP)	0.36(0.08)	
$\hat{\sigma}$	1.04(0.03)	1.05(0.01)
Antall parametre	4	6
AIC	2 661.7	13 129
BIC	2 682.1	13 169

Tabell 3.3: Estimerte verdier for lineært betinget modell.

Fra tabellen ovenfor ser man at alle kovariatene er signifikante.
I tabellen har jeg oppgitt AIC og BIC for den lineært betingede modellen. Men det er ikke helt åpenbart om disse verdiene kan brukes til å sammenligne den lineære modellen og den lineært betingede modellen. Begge modellene sin marginale fordeling er lognormal. Justeringen av lognormal modell mhp. egenandelene(den lineært betingede modellen) kan sammenlignes med de nullforhøyde modellene i avsnitt 2.2, noe som kan brukes som argument for at en slik sammenligning er meningsfull. Men siden jeg i denne masteroppgaven fokuserer på å modellere skadefrekvens så velger jeg å ikke legge noen vekt på AIC og BIC verdiene i tabell 3.3.

For å beregne en estimert verdi for den lineært betingede modellen kan man gjøre følgende for $i = 1, \ldots, n$:

$$\begin{aligned}
\mathbb{E}[Y_i | Y_i > d_i] &= \int_{d_i}^{\infty} y \, f_{Y_i|Y_i>d_i}(y) \, \mathrm{d}y \\
&= \int_{d_i}^{\infty} y \frac{f_{Y_i}(y)}{1 - F_{Y_i}(d_i)} \, \mathrm{d}y \\
&= \frac{1}{1 - F_{Y_i}(d_i)} \int_{d_i}^{\infty} y \, f_{Y_i}(y) \, \mathrm{d}y \\
&= \frac{1}{1 - F_{Y_i}(d_i)} \int_{d_i}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(\log(y) - \boldsymbol{X}_i'\boldsymbol{\beta})^2}{2\sigma^2} \right) \mathrm{d}y
\end{aligned}$$

Integralet ovenfor kan ikke løses analytisk, så jeg har brukt funksjonen integrate i R til å finne en numerisk løsning.

Det kan nå virke fristende å plotte observasjoner og estimerte verdier opp mot hverandre. Men selv for tankskip er det så mange observasjoner at det blir vanskelig å se forskjeller utifra plott. Derfor har jeg heller samlet informasjon om observerte og estimerte verdier for skadegrad i to tabeller nedenfor.

Modell	Min.	1 kvantil	Gj.snitt	3 kvantil	Maks	Sum
Observert	0.0026	0.0090	0.0330	0.0334	0.5979	40.40
Lineær	0.0047	0.0183	0.0305	0.0363	0.2086	37.24
Betinget lineær	0.0068	0.0189	0.0331	0.0405	0.3205	40.48

Tabell 3.4: Observerte og estimerte verdier for skadegrad til tankskip.

Modell	Min.	1 kvantil	Gj.snitt	3 kvantil	Maks	Sum
Observert	0.0020	0.0096	0.0409	0.0415	0.6	240.39
Lineær	0.0032	0.0194	0.0377	0.0473	0.2335	221.85
Betinget lineær	0.0044	0.0204	0.0413	0.0512	0.3139	243.18

Tabell 3.5: Observerte og estimerte verdier for skadegrad til alle typer skip.

Tabellene viser at den linært betingede modellen gir høyere estimerte verdier enn det den lineære modellen gir. Blant annet ligger gjennomsnitt og sum for den lineært betingede modellen betydelig nærmere de observerte verdiene enn det den linære modellen gjør. For et forsikringsselskap er det viktig at disse to størrelsene ligger nærme de virkelige verdiene.

3.3 Beregning av motvekt

Jeg vil nå bruke parametrene fra den lineært betingede modellen til å lage en motvekt (offset). Denne motvekten vil senere bli brukt i regresjonsmodeller for skadefrekvensen. Referansen her er Paulsen *et al.* (2008) og forelesningsnotater i kurset StatLIKELIHOOD.

La $\tilde{N}_1, \ldots, \tilde{N}_n$ være virkelig skadefrekvens for poliser $i = 1, \ldots, n$. Jeg antar at hver skadefrekvens er Poisson fordelt på formen $\tilde{N}_i \sim \text{Poisson}(\mu_i)$. La så $\tilde{Y}_{i,1}, \ldots, \tilde{Y}_{i,\tilde{N}_i}$ være skadegradene som tilhører $\tilde{N}_i$ for $i = 1, \ldots, n$. $\tilde{N}_i = 0$ betyr at det ikke har skjedd noen skade i polise i.
Forsikringsselskapet vil kun få registrert skader hvor skadegraden, $\tilde{Y}_i$, er større enn skipets egenandel, d_i. Selskapet vil dermed ha informasjonen

$$\begin{gathered} N_1, \ldots, N_n; \;\; Y_{1,1}, \ldots, Y_{1,N_1}, \ldots, Y_{n,1}, \ldots, Y_{n,N_n} \,, \\ \text{hvor } N_i \leq \tilde{N}_i, \; Y_{i,j} > d_i \; \forall \; j = 1, \ldots, N_i; \; i = 1, \ldots, n \,. \end{gathered}$$

Her er Y et registrert krav hos selskapet, N er observert skadefrekvens, mens $\tilde{Y}$ vil være den virkelige skadegraden uavhengig om den er registrert eller ikke.
La $f_{\tilde{Y}}(y; \boldsymbol{\theta})$ være fordelingen til en skadegrad. Da vil fordelingen til et observert krav være gitt ved

$$f_{Y_i|Y_i>d_i}(y; \boldsymbol{\theta}) = \frac{f_{\tilde{Y}_i}(y; \boldsymbol{\theta})}{1 - F_{\tilde{Y}_i}(d_i; \boldsymbol{\theta})} = \frac{f_{\tilde{Y}_i}(y; \boldsymbol{\theta})}{\mathbb{P}(\tilde{Y}_i > d_i; \boldsymbol{\theta})} \,. \tag{3.7}$$

Denne fordelingen kan sees direkte i sammenheng med formel 3.4 på side 28.

Jeg lar ρ være sannsynligheten for at en skade blir registrert hos forsikringsselskapet. Da er $\rho = \mathbb{P}(\tilde{Y} > d; \boldsymbol{\theta}) = 1 - F_{\tilde{Y}}(d; \boldsymbol{\theta})$ hvor $\boldsymbol{\theta}$ er parametervektoren tilhørende $\tilde{Y}$. Hvis jeg antar at $N|\tilde{N} = h \sim \text{Binomisk}(\rho, h)$ så går det an å vise at $N \sim \text{Poisson}(\rho\mu) = \text{Poisson}([1 - F_{\tilde{Y}}(d; \boldsymbol{\theta})]\mu)$. Dette er nærmere forklart i avsnitt A.1.
Motvekten jeg vil beregne er ρ. Utifra et skips kovariater forklarer ρ hvor sannsynlig det er at det registreres en skade for dette skipet når en skade har skjedd.

Tidligere har jeg beregnet parametervektoren $\hat{\boldsymbol{\kappa}}$ for formel 3.5 på side 28. Dermed kan jeg la $\boldsymbol{\theta} = \boldsymbol{\kappa}$ og ved hjelp av sammenhengen nedenfor kan jeg

beregne verdier for ρ:

$$\begin{aligned}\rho &= 1 - F_{\tilde{Y}}(d;\boldsymbol{\kappa}) = 1 - \mathbb{P}(\tilde{Y} \leq d) = 1 - \mathbb{P}(\log(\tilde{Y}) \leq \log(d)) \\ &= 1 - \mathbb{P}(T \leq z) = 1 - F_T(z;\boldsymbol{\kappa}). \\ \rho_i &\approx 1 - \Phi\Big(\frac{\log(d_i) - \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\Big); \;\; i = 1, \ldots, n\,. \end{aligned} \tag{3.8}$$

Her uttrykker Φ den kumulative fordelingsfunksjonen til standardnormalfordelingen.
For å beregne ρ trenger jeg dermed bare å bruke parameterverdier for den lineært betingede modellen og verdier fra kovariatene. I R er det en egen funksjon for den kumulative fordelingsfunksjonen til normalfordelingen, så det er uproblematisk å beregne verdier for ρ.

I hele oppgaven min har jeg brukt logaritmen som den kanoniske linkfunksjonen for regresjon til skadefrekvensen. Sammenhengen mellom forventningsvektoren, kovariatmatrisen, parametervektoren og vektoren med motvekter($\boldsymbol{\eta}$) er på formen $\boldsymbol{\mu} = \exp(\boldsymbol{X\beta} + \boldsymbol{\eta})$. Siden jeg er interessert i å multiplisere ρ med μ så vil jeg i praksis bruke $\log(\rho)$ som motvekt i modellene. Sammenhengen blir da på formen $\boldsymbol{\rho\mu} = \boldsymbol{\rho}\exp(\boldsymbol{X\beta}) = \exp(\boldsymbol{X\beta} + \log(\boldsymbol{\rho}))$.

I kapittel 2 brukte jeg log(Days.covered) som motvekt i regresjonsmodellene. Denne motvekten vil jeg fortsatt bruke siden det forbedrer modellene. Når jeg sammenligner modeller med og uten $\log(\rho)$ som motvekt så mener jeg η =log(Days.covered) mot η^* =log(Days.covered)+ $\log(\rho)$. I resten av masteroppgaven vil modeller med η^* som motvekt ha $*$ på slutten av navnene sine, mens modeller med η som motvekt ikke har det.

3.4 Bruk av motvekt på Negativ Binomisk modell

Her vil jeg bruke motvekten $\log(\rho)$ på den Negativt Binomiske modellen fra kapittel 2. I tillegg vil jeg bruke Nullforhøyde Negativt Binomiske modeller(ZINB) på datasettet mitt, både med og uten $\log(\rho)$. Resultatene for Poisson modeller med $\log(\rho)$ som motvekt vil bli oppgitt i det neste kapittelet.
For ZINB modeller oppgir ikke funksjonen zeroinfl i R standardfeil til $\hat{\varsigma}$. Den oppgir istedet standardfeil til $\log(\hat{\varsigma})$. Jeg bruker derfor Delta-metoden til å beregne standardfeil til $\hat{\varsigma}$ vha. estimatet til $\hat{\varsigma}$ og standardfeilen til $\log(\hat{\varsigma})$.

Tankskip	NB*	ZINB	ZINB*
Verdier for $\boldsymbol{\beta}$:			
Intercept	-4.73(0.98)	-1.55(0.43)	-1.50(0.43)
log(Age+2)	0.35(0.06)	0.24(0.05)	0.28(0.05)
log(GT)	-0.15(0.04)	-0.12(0.03)	-0.11(0.03)
log(Sum.Insured)	0.20(0.06)		
Stroke	0.27(0.10)	0.34(0.10)	0.35(0.10)
$\hat{\varsigma}$	1.23(0.26)	1.27(0.29)	1.44(0.35)
Verdier for γ:			
Intercept		31.56(8.59)	33.65(9.94)
log(Sum.Insured)		-2.13(0.58)	-2.28(0.67)
Antall parametre	6	7	7
AIC	8 108.0	8 153.5	8 095.8
BIC	8 152.5	8 205.4	8 147.7

Tabell 3.6: Estimerte verdier for tankskip.

	NB*	ZINB	ZINB*
Verdier for $\boldsymbol{\beta}$:			
Intercept	-5.06(0.33)	-7.65(0.47)	-7.29(0.50)
log(Age+2)	0.23(0.02)	0.17(0.02)	0.23(0.02)
log(GT)	-0.12(0.02)		
log(Sum.Insured)	0.13(0.02)	0.31(0.02)	0.30(0.03)
log(HP)	0.16(0.02)		
Stroke	0.37(0.03)	0.50(0.03)	0.39(0.03)
$\hat{\varsigma}$	1.26(0.11)	2.00(0.32)	2.43(0.48)
Verdier for γ:			
Intercept		-19.33(1.78)	-14.50(1.76)
log(GT)		0.75(0.11)	0.86(0.13)
log(Sum.Insured)		1.19(0.12)	0.84(0.12)
log(HP)		-1.00(0.17)	-1.02(0.19)
Antall parametre	7	9	9
AIC	37 573.0	37 879.4	37 504.7
BIC	37 635.2	37 959.4	37 584.7

Tabell 3.7: Estimerte verdier for alle typer skip.

Nå kan man sammenligne verdiene for ZINB med verdiene fra tabell 2.2 på side 19 og tabell 2.3 på side 20. Tabellene viser at ZINB har lavere AIC- og BIC verdi enn NB for både tankskip og alle typer skip. Dette viser at Nullforhøyd Negativ Binomisk modell gir best tilpasning til datasettet av de modellene jeg har undersøkt så langt.

Det er verdt å merke seg at estimatet til parameteren ς øker markert når man innfører $\log(\rho)$ som motvekt, både for Negativ Binomisk modell og ZINB. Som tidligere nevnt indikerer ς graden av overdispersjon i datasettet. En økning i ς når $\mu($ og ϑ for ZINB) holdes fast fører til at overdispersjonen minker i modellene. Økningen i $\hat{\varsigma}$ kan komme av at $\log(\rho)$ som motvekt forklarer en del av null-observasjonene, slik at det blir mindre overdispersjon i datasettet. Men siden parametrene i μ og ϑ også endrer seg som følge av innføringen av $\log(\rho)$ så kan man ikke garantere at $\log(\rho)$ har ført til mindre overdispersjon.

For å undersøke om $\log(\rho)$ fører til mindre overdispersjon har jeg beregnet forventning og varians for den vanlige NB modellen. Jeg har så tatt gjennomsnitt av beregnet forventning og varians for alle observasjonene i datasettet.

Modell	Forventning	Varians	Varians/forventning
NB Tankskip	0.1002	0.1110	1.1079
NB* Tankskip	0.1000	0.1097	1.0967
NB alle skip	0.1108	0.1237	1.1169
NB* alle skip	0.1106	0.1222	1.1049

Tabell 3.8: Tabell som viser overdispersjon for Negativ Binomisk modell.

Tabellen viser at overdispersjonen minker litt for NB modellen etter å ha innført $\log(\rho)$ som motvekt. For ZINB minker overdispersjonen litt for alle typer skip når man har innført $\log(\rho)$ som motvekt, mens overdispersjonen øker litt for tankskip. Dermed er det vanskelig å bestemme hvordan $\log(\rho)$ påvirker overdispersjonen for disse modellene.

4

GLMM: generalisert lineært miksede modeller

Jeg utvider nå GLM til å inneholde tilfeldige effekter. Referansen her er McCulloch og Searle (2001).
Som vanlig har jeg en vektor $\boldsymbol{y}$ med n observasjoner. Observasjonene kan deles inn i m grupper og jeg har grunn til å tro at observasjonene innen hver gruppe er korrelerte. Jeg har som tidligere kovariatmatriser $\boldsymbol{X}$ og $\boldsymbol{Z}$ som henholdsvis har dimensjon $n \times p$ og $n \times q$, en parametervektor $\boldsymbol{\beta}$ med p elementer og en vektor $\boldsymbol{u}$ som inneholder q tilfeldige variabler. Kovariatmatrisen $\boldsymbol{Z}$ bør være en delmengde av $\boldsymbol{X}$, slik at $q \leq p$.
Mange av egenskapene til GLM vil også gjelde for GLMM.

Definisjon 4.0.1: GLMM

Jeg forutsetter at $\boldsymbol{Y}$ gitt $\boldsymbol{u}$ består av uavhengige elementer. Fordelingen til elementene kommer fra den eksponensielle familien:

$$\begin{aligned} Y_i|\boldsymbol{u} &\sim \text{ uavh.} f_{Y_i|\boldsymbol{u}}(y_i)\ , \\ f_{Y_i|\boldsymbol{u}}(y_i) &= \exp\left(\frac{y_i\theta_i - A(\theta_i)}{\varphi} + c(y_i,\varphi)\right)\ , \\ \mathbb{E}[Y_i|\boldsymbol{u}] &= \mu_i\ , \\ g(\mu_i) &= \xi_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u},\ \ i = 1,2,...,n\ . \end{aligned}$$

$g(.)$ er den kjente link-funksjonen som er glatt og invertibel, og ξ_i er den lineære prediktoren.
Jeg tilegner også en egen fordeling til $\boldsymbol{u}$: $\boldsymbol{u} \sim f_{\boldsymbol{U}}(\boldsymbol{u})$.

For GLMM får jeg, som for GLM, følgende forhold mellom den betingede forventningen og variansen til $\boldsymbol{Y}$:

$$\mu_i = \frac{\partial A(\theta_i)}{\partial \theta_i},\quad \text{Var}(Y_i|\boldsymbol{u}) = \varphi\frac{\partial^2 A(\theta_i)}{\partial \theta_i^2} = \varphi\nu(\mu_i)\ .$$

Jeg får følgende ligninger for marginal forventning og varians:

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|\boldsymbol{u}]] = \mathbb{E}[\mu_i] = \mathbb{E}[g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u})]\ .$$

$$\begin{aligned} \text{Var}(Y_i) &= \text{Var}(\mathbb{E}[Y_i|\boldsymbol{u}]) + \mathbb{E}[\text{Var}(Y_i|\boldsymbol{u})] \\ &= \text{Var}(\mu_i) + \mathbb{E}[\varphi\nu(\mu_i)] \\ &= \text{Var}(g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u})) + \varphi\,\mathbb{E}[\nu(g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}))]\ . \end{aligned}$$

For $i \neq j$ får jeg følgende ligning for marginal kovarians:

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \text{Cov}(\mathbb{E}[Y_i|\boldsymbol{u}], \mathbb{E}[Y_j|\boldsymbol{u}]) + \mathbb{E}[\text{Cov}(Y_i, Y_j|\boldsymbol{u})] \\ &= \text{Cov}(\mu_i, \mu_j) + 0 \\ &= \text{Cov}(g^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}), g^{-1}(\boldsymbol{X}_j'\boldsymbol{\beta} + \boldsymbol{Z}_j'\boldsymbol{u}))\ . \end{aligned}$$

For å forenkle disse uttrykkene må jeg definere formen til g og/eller den betingede fordelingen til $\boldsymbol{Y}$. Det er spesielt interessant å legge merke til $\text{Var}(Y_i)$ og $\text{Cov}(Y_i, Y_j)$. I de siste leddene der er det kun $\boldsymbol{u}$ som er stokastisk.

Ved hjelp av fordelingene jeg definerte ovenfor kan likelihood funksjonen til observasjonene $\boldsymbol{y}$ nå uttrykkes:

$$\begin{aligned} L =& f_Y(\boldsymbol{y}) = \int_{\boldsymbol{u}} f_{Y,U}(\boldsymbol{y},\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} = \int_{\boldsymbol{u}} f_{Y|U}(\boldsymbol{y}|\boldsymbol{u}) f_U(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \\ =& \int_{\boldsymbol{u}} \left(\prod_{i=1}^{n} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) \right) f_U(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \end{aligned} \tag{4.1}$$

Fordelingene til både $\boldsymbol{y}$ og $\boldsymbol{u}$ inneholder parametre som jeg må maksimere vha. likelihood funksjonen. I likelihood funksjonen ovenfor integrerer jeg over den q-dimensjonale fordelingen til $\boldsymbol{u}$. Det er derfor ingen triviell måte å finne estimatorer for parametrene i modellen. I de enkleste tilfellene kan numerisk integrasjon brukes direkte til å finne estimatorer.

4.1 Estimering av parametre

For å beregne GLMM i R kan man bl.a. bruke funksjonen lmer i pakken "lme4"eller glmm.admb i pakken "glmmADMB".

PQL

Jeg vil nå vise hvordan man kan komme frem til en forholdsvis enkel algoritme for å beregne parameterestimater for GLMM vha. Laplace approksimasjon. Dette gjør jeg for å vise hvor problematisk det kan være å beregne parametre til en slik modell. Denne utledningen står forklart i McCulloch og Searle (2001).

Det jeg er ute etter er ligninger som kan brukes til å estimere parametervektoren $\boldsymbol{\beta}$ og eventuelle parametre fra fordelingen til $\boldsymbol{u}$. Jeg tar utgangspunkt i den generelle modellen ovenfor og viser noen egenskaper til $f_{Y_i|\boldsymbol{u}}(y_i)$:

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \left(\frac{\partial^2 A(\theta_i)}{\partial \theta_i^2} \right)^{-1} = \frac{1}{\nu(\mu)}\,.$$

$$\frac{\partial \mu_i}{\partial \boldsymbol{u}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \boldsymbol{u}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \left(\frac{\partial}{\partial \boldsymbol{u}} (\boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{u}) \right) = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \boldsymbol{Z}_i'\,.$$

Jeg tar så utgangspunkt i log likelihood for $f_{Y_i|\boldsymbol{u}}(y_i)$ og deriverer denne mhp. $\boldsymbol{u}$:

$$l(\boldsymbol{\beta}; \boldsymbol{y}|\boldsymbol{u}) = l_1 = \sum_{i=1}^{n} \left(\frac{[y_i\theta_i - A(\theta_i)]}{\varphi} - c(y_i, \varphi) \right).$$

$$\begin{aligned}
\frac{\partial l_1}{\partial \boldsymbol{u}} &= \frac{1}{\varphi} \sum_{i=1}^{n} \left[y_i \frac{\partial \theta_i}{\partial \boldsymbol{u}} - \frac{\partial A(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{u}} \right] \\
&= \frac{1}{\varphi} \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \boldsymbol{u}} \\
&= \frac{1}{\varphi} \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{u}} \\
&= \frac{1}{\varphi} \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{\nu(\mu_i)} \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \boldsymbol{Z}_i' \\
&= \frac{1}{\varphi} \sum_{i=1}^{n} (y_i - \mu_i) w_i \frac{\partial g(\mu_i)}{\partial \mu_i} \boldsymbol{Z}_i' \,, \\
&\text{hvor } w_i = \left[\nu(\mu_i) \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^2 \right]^{-1}.
\end{aligned}$$

På matriseform får jeg

$$\frac{\partial l_1}{\partial \boldsymbol{u}} = \frac{\partial}{\partial \boldsymbol{u}} \log f_{Y|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) = \frac{1}{\varphi} \boldsymbol{Z}' \boldsymbol{W} \triangle (\boldsymbol{y} - \boldsymbol{\mu}).$$

Her er $\boldsymbol{W}$ en $n \times n$ matrise hvor diagonalen består av elementene $w_i, : \ i = 1, \ldots, n$, alle andre elementer er lik 0. $\triangle$ er også en $n \times n$ matrise men her er diagonalen $\frac{\partial g(\mu_i)}{\partial \mu_i}$; $i = 1, \ldots n$, alle andre elementer er lik 0.
Ved å beregne den deriverte av log likelihood funksjonen mhp. $\boldsymbol{\beta}$ kan man ved samme fremgangsmåte som ovenfor vise at

$$\frac{\partial l_1}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y|\boldsymbol{U}}(\boldsymbol{y}|\boldsymbol{u}) = \frac{1}{\varphi} \boldsymbol{X}' \boldsymbol{W} \triangle (\boldsymbol{y} - \boldsymbol{\mu}).$$

Jeg tar nå utgangspunkt i en Laplace approksimasjon av den marginale log likelihood til $\boldsymbol{y}$. En første ordens Laplace approksimasjon er gitt ved

$$\int_u e^{h(u)}\,\mathrm{d}u \approx e^{h(u_0)}(2\pi)^{q/2}\left| -\frac{\partial^2 h(u)}{\partial u \partial u'}\Big|_{u=u_0}\right|^{-\frac{1}{2}},$$

hvor u_0 er løsningen på ligningen

$$\frac{\partial h(u)}{\partial u}\Big|_{u=u_0} = 0\,. \tag{4.2}$$

Ved å ta logaritmen til Laplace approksimasjonen ovenfor får jeg følgende uttrykk

$$\log\int_u e^{h(u)}\,\mathrm{d}u \approx h(u_0) + \frac{q}{2}\log(2\pi) - \frac{1}{2}\log\left| -\frac{\partial^2 h(u)}{\partial u \partial u'}\Big|_{u=u_0}\right|. \tag{4.3}$$

Jeg skriver så om den marginale log likelihood funksjonen til y slik den passer til Laplace approksimasjonen:

$$\begin{aligned} l &= \log\int_u f_{Y|U}(y) f_U(u)\,\mathrm{d}u \\ &= \log\int_u e^{(\log f_{Y|U}(y) + \log f_U(u))}\,\mathrm{d}u \\ &= \log\int_u e^{h(u)}\,\mathrm{d}u\,, \\ h(u) &= \log f_{Y|U}(y) + \log f_U(u). \end{aligned}$$

For å finne estimater for parametrene i log likelihood funksjonen må jeg løse formel 4.2 og finne et uttrykk for $|-\frac{\partial^2 h(u)}{\partial u \partial u'}|$.

For å komme videre med beregningene antar jeg at vektoren $u = [u_1, u_2, \ldots, u_q]'$ er multivariat normalfordelt på formen

$$\begin{gathered} u \sim \mathcal{N}_q(0, B). \text{ Tettheten er da} \\ f_U(u) = \frac{1}{(2\pi)^{q/2}|B|^{1/2}}\exp\{-u'B^{-1}u/2\}, \\ -\infty < u_j < \infty,\ \ j = 1, \ldots, q\,. \end{gathered}$$

Det er dermed $\boldsymbol{B}$ og $\boldsymbol{\beta}$ jeg må estimere for å optimere likelihood funksjonen. Nå er det mulig å derivere $h(\boldsymbol{u})$ med hensyn på $\boldsymbol{u}$:

$$h(\boldsymbol{u}) = \log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}) - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{B}|) - \frac{1}{2}\boldsymbol{u}'\boldsymbol{B}^{-1}\boldsymbol{u},$$
$$\frac{\partial h(\boldsymbol{u})}{\partial \boldsymbol{u}} = \frac{\partial}{\partial \boldsymbol{u}}\log f_{\boldsymbol{Y}|\boldsymbol{U}}(\boldsymbol{y}) - \boldsymbol{B}^{-1}\boldsymbol{u} = \frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\triangle(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{B}^{-1}\boldsymbol{u}.$$

For å finne $\boldsymbol{u}_0$ må jeg sette ligningen ovenfor lik 0:

$$\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\triangle(\boldsymbol{y}-\boldsymbol{\mu}) - \boldsymbol{B}^{-1}\boldsymbol{u} = 0.$$

I ligningen ovenfor er $\boldsymbol{W}$, $\triangle$ og $\boldsymbol{\mu}$ funksjoner av $\boldsymbol{u}$, mens $\boldsymbol{B}$ er kovariansmatrisen til $\boldsymbol{u}$. Denne ligningen er nødvendig for å beregne $\boldsymbol{\beta}$ og $\boldsymbol{B}$, noe jeg vil komme tilbake til senere.

Nå må jeg finne et uttrykk for den andrederiverte til $h(\boldsymbol{u})$:

$$\frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}'} = -\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\triangle\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{u}} - \boldsymbol{B}^{-1} + \underbrace{\frac{1}{\varphi}\boldsymbol{Z}'\Big(\frac{\partial}{\partial \boldsymbol{u}}\boldsymbol{W}\triangle\Big)(\boldsymbol{y}-\boldsymbol{\mu})}_{(\star)}.$$

For noen fordelinger til $Y_i|\boldsymbol{u}$, blant annet Poisson og Binomisk, vil $\boldsymbol{W}\triangle = \boldsymbol{I}_N$ slik at $(\star) = 0$. Generelt vil forventningen til $(\boldsymbol{y}-\boldsymbol{\mu})$ betinget på $\boldsymbol{u}$ være lik 0. Derfor velger jeg å se bort fra $(\star)$, dvs. la $(\star) = 0$, slik at ligningene ikke blir for kompliserte.

Siden $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{u}} = \triangle^{-1}\boldsymbol{Z}$ får jeg følgende ligning:

$$\begin{aligned}
-\frac{\partial^2 h(\boldsymbol{u})}{\partial \boldsymbol{u}\partial \boldsymbol{u}'} &= \frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\triangle\triangle^{-1}\boldsymbol{Z} + \boldsymbol{B}^{-1} \\
&= \frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{B}^{-1} \\
&= (\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{I}_N)\boldsymbol{B}^{-1}.
\end{aligned}$$

Nå har jeg de 2 uttrykkene jeg trenger, så jeg setter de inn i Laplace approksimasjonen og får

$$\begin{aligned} l &\approx \log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}_0) - \frac{1}{2}\boldsymbol{u}_0'\boldsymbol{B}^{-1}\boldsymbol{u}_0 - \frac{q}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{B}|) \\ &\quad - \frac{1}{2}\log\left|(\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{I}_N)\boldsymbol{B}^{-1}\right| + \frac{q}{2}\log(2\pi) \\ &= \log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}_0) - \frac{1}{2}\boldsymbol{u}_0'\boldsymbol{B}^{-1}\boldsymbol{u}_0 - \frac{1}{2}\log\left|\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{I}_N\right|. \end{aligned}$$

Jeg deriverer nå log likelihood funksjonen ovenfor med hensyn på $\boldsymbol{\beta}$ for å få en ny ligning:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}_0) \underbrace{-\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\beta}}\log\left|\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}\boldsymbol{B} + \boldsymbol{I}_N\right|}_{(\star\star)}. \tag{4.4}$$

I $(\star\star)$ er det $\boldsymbol{W}$ som avhenger av $\boldsymbol{\beta}$. Jeg forutsetter at $\boldsymbol{W}$ endrer seg lite som funksjon av $\boldsymbol{\beta}$. Dette virker litt suspekt, men blir gjort når man har en slik Laplace transformasjon i Breslow og Clayton (1993) og McCulloch og Searle (2001). Under denne forutsetningen lar jeg $(\star\star) = 0$ slik at ligningen får følgende form:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} \approx \frac{\partial}{\partial \boldsymbol{\beta}}\log f_{\boldsymbol{Y}|\boldsymbol{u}}(\boldsymbol{y}|\boldsymbol{u}_0) = \frac{1}{\varphi}\boldsymbol{X}'\boldsymbol{W} \triangle (\boldsymbol{y} - \boldsymbol{\mu}). \tag{4.5}$$

Ved å sette ligningen ovenfor lik 0 har jeg enda en ligning jeg kan bruke til å beregne ukjente parametre. Etter alle beregningene har jeg endt opp med følgende ligninger:

$$\frac{1}{\varphi}\boldsymbol{X}'\boldsymbol{W} \triangle (\boldsymbol{y} - \boldsymbol{\mu}) = 0 \tag{4.6}$$

$$\frac{1}{\varphi}\boldsymbol{Z}'\boldsymbol{W} \triangle (\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{B}^{-1}\boldsymbol{u} = 0 \tag{4.7}$$

Disse ligningene må løses samtidig for $\boldsymbol{\beta}$ og $\boldsymbol{u}$ for å estimere $\boldsymbol{\beta}$. Man trenger i tillegg en metode for å estimere $\boldsymbol{B}$. En metode er å bruke estimatet av $\boldsymbol{u}$ til å estimere $\boldsymbol{B}$ vha. vanlig SME.

Under hele denne utledningen har jeg ikke definert hva fordelingen til $Y_i|\boldsymbol{u}$ er. Det eneste som mangler i ligningene ovenfor er å definere link-funksjonen $g()$ og variansfunksjonen $\nu()$. Metoder som løser ligningene ovenfor blir ofte kalt penalized quasi likelihood (PQL) metoder. PQL kan sees i direkte sammenheng med kvasi-likelihood tidligere i masteroppgaven.
I McCulloch (2003) foreslår man en algoritme på følgende form for å finne estimater for $\boldsymbol{\beta}$ og $\boldsymbol{B}$ i en Poisson-Normal modell:

1 Finn startverdier for $\boldsymbol{\beta}$, $\boldsymbol{B}$, φ og $\boldsymbol{u}$.

2 Bruk startverdiene til å løse formel 4.6 på forrige side mhp. $\boldsymbol{u}$ og få et estimat for $\boldsymbol{u}$.

3 Bruk det nye estimatet av $\boldsymbol{u}$ til å estimere $\boldsymbol{B}$.

4 Løs formel 4.7 på forrige side mhp. φ og $\boldsymbol{\beta}$ ved hjelp av de nye estimatene for $\boldsymbol{u}$ og $\boldsymbol{B}$.

5 Begynn igjen fra punkt 2 med de nye estimatene som startverdier. Fortsett til ønsket konvergens er oppnådd.

Når $Y_i|\boldsymbol{u},\ i = 1, \cdots, n$ er Poisson eller Binomisk fordelt så er $\boldsymbol{W}\triangle = \boldsymbol{I}_n$ og ligningene blir lettere å løse. Dette vil også gjelde for Negativt Binomisk fordeling hvis fordelingen er på formen til en GLM.

I praksis fungerer ikke PQL metoder så bra. McCulloch (2003) påpeker at algoritmen ovenfor fungerer godt hvis $Y_i|\boldsymbol{u}$ er tilnærmet normalfordelt. For fordelinger som er langt fra normalfordelt kan PQL metoder fungere dårlig. McCulloch og Searle (2001) hevder at PQL fungerer spesielt dårlig for binære data.
I utledningen til PQL var spesielt overgangen fra formel 4.4 på forrige side til formel 4.5 på forrige side vanskelig å godta uten videre. Denne overgangen kan være en av flere årsaker til at PQL ikke fungerer så bra.

Beregningsmetodene i funksjonen lmer i R har blitt endret mye de siste årene. Tidligere kunne man velge å beregne parametrene til GLMM med PQL, Laplace approksimasjon eller en såkalt Lindstrom-Bates algoritme. Men med tiden har man gått over til å bare bruke PQL til å beregne startverdier, for så å fjerne PQL helt fra lmer.

Per dags dato kan man velge mellom 2 beregningsmetoder i lmer for GLMM: Laplace approksimasjon og Adaptive Gauss-Hermite Quadrature(AGQ). Laplace approksimasjonen i lmer er mer komplisert enn Laplace approksimasjonen som blir brukt for PQL.
Som tidligere vist er det nødvendig med flere approksimasjoner for å finne parametre til GLMM vha. PQL. Likelihood verdien som blir produsert er også en approksimasjon. Og selv om man har verdier for parametrene så må man fremdeles løse formel 4.1 på side 37, noe som ikke nødvendigvis er trivielt. Siden Likelihood-verdien ikke er eksakt kan man ikke bruke AIC ukritisk på GLMM.

4.2 Bruk av GLMM på datasettet

Nå vil jeg definere Poisson GLMM og bruke modellen på dataene for skadefrekvens. For å se hvor godt denne modellen fungerer vil jeg sammenligne med resultatene fra Poisson, NB, ZIP og ZINB modellene i kapittel 2 og 3. Jeg skal også undersøke om motvekten definert i formel 3.8 på side 32 fører til at Poisson og ZIP modellene blir bedre.

Poisson GLMM tar utgangspunkt i den generelle modellen definert tidligere i kapittelet. I tillegg antas det følgende:

$$\begin{aligned} Y_k|\boldsymbol{u} &\sim \text{Poisson}(\mu_k); \quad k = 1, \ldots, n\,, \\ \boldsymbol{u} &\sim \mathcal{N}_q(0, \boldsymbol{B}). \end{aligned}$$

Som følge av egenskapene til Poisson fordelingen vil $\mathbb{E}[Y|\boldsymbol{u}] = \text{Var}(Y|\boldsymbol{u})$. Som tidligere lar jeg logaritmen være link-funksjonen for μ, dvs. $g(\mu) = \log(\mu)$. Motvekten(η) har en sentral plass i denne seksjonen, derfor lar jeg den være med i definisjonen til den lineære prediktoren.

Som nevnt tidligere kan man beregne GLMM i R med funksjonene glmm.admb og lmer. Funksjonen glmm.admb konvergerte ikke for de 2 datasettene jeg har tatt utgangspunkt i. Funksjonen lmer konvergerte for begge, men oppga en log likelihood verdi som var veldig tvilsom. lmer oppgir heller ikke standardfeil til varianskomponentene i GLMM.
For å sammenligne glmm.admb og lmer undersøkte jeg et mindre datasett bestående av 1000 observasjoner fra tankskip. I dette datasettet er det 100

observasjoner fra hver gruppe, altså 100 observasjoner for hvert tegningsår. For dette datasettet konvergerte både glmm.admb og lmer for en enkel GLMM. Funksjonene ga omtrent samme parameterestimater, men lmer hadde en mye høyere log likelihood verdi enn det glmm.admb hadde.
Jeg prøvde å beregne log likelihood verdien til formel 4.1 på side 37 med utgangspunkt i parameterestimatene til lmer, og da fikk jeg samme log likelihood verdi som glmm.admb. For å beregne log likelihood verdien brukte jeg funksjonen optim. Dette gikk bra siden jeg allerede hadde parameterestimater fra lmer som optimerte formel 4.1 på side 37.
Douglas Bates, som har hovedansvaret for lmer, skrev i November 2008 at han har tenkt til å endre AIC-verdien som blir gitt for GLMM i lmer. Derfor bruker jeg ikke lmer til å finne AIC. For datasettene mine bruker jeg parameterestimatene fra lmer, men beregner log likelihood verdi og standardfeil(dvs. den hessiske matrisen til log likelihood) selv vha. parameterestimatene fra lmer. Jeg bruker AIC-verdiene i lmer til å finne ut hvilke kovariater det er hensiktsmessig å bruke i modellene.

Jeg har valgt å bare bruke modeller der de tilfeldige effektene påvirker en kovariat, dvs. $q = 1$. Dette er på grunn av 2 ting: Den ene er at modellene i lmer i beste fall bare blir litt bedre hvis $q > 1$. Den andre er at beregningene blir omfattende nok for $q = 1$. Datasettene mine er såpass store at beregningene går sakte nok når man optimerer over et endimensjonalt integral. Når f.eks $q = 2$ får man 2 ekstra parametre og man må optimere over et todimensjonalt integral.

For Poisson GLMM med log-link og bare en kovariat i $\boldsymbol{Z}$ får jeg modellen nedenfor. Kovariaten i $\boldsymbol{Z}$ kan enten være en konstant eller en vanlig kovariat. Fremgangsmåten her er veldig lik et eksempel i McCulloch og Searle (2001) s. 225-226.

$$\begin{aligned} Y_{ij}|u_i &\sim \text{Poisson}(\mu_{ij});\ i = 1,\ldots,10;\ j = 1,\ldots,n_i\ , \\ u_i &\sim \text{iid. } \mathcal{N}(0,\psi^2), \\ \log(\mu_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + Z_{ij}u_i + \eta_{ij}\ , \\ \boldsymbol{U} &= (U_1, U_2, \ldots, U_{10})'\ . \end{aligned} \tag{4.8}$$

Indeks i står her for gruppe nr. i, dvs. i forteller hvilket tegningsår det er snakk om.
Jeg kan utifra dette finne den aktuelle log likelihood:

$$\begin{aligned}
L &= \int_{\boldsymbol{u}} \left(\prod_{i=1}^{10} \prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_i) \right) \prod_{i=1}^{10} f_{U_i}(u_i)\, \mathrm{d}\boldsymbol{u} \\
&= \int_{\boldsymbol{u}} \prod_{i=1}^{10} \left(\left(\prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_i) \right) f_{U_i}(u_i) \right) \mathrm{d}\boldsymbol{u} \\
&= \int_{U_1} \int_{U_2} \cdots \int_{U_{10}} \prod_{i=1}^{10} \left(\left(\prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_i) \right) f_{U_i}(u_i) \right) \mathrm{d}u_1\, \mathrm{d}u_2 \cdots \mathrm{d}u_{10} \\
&= \prod_{i=1}^{10} \int_{U_i} \left(\prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_j) \right) f_{U_i}(u_i)\, \mathrm{d}u_i \,.
\end{aligned}$$

$$\begin{aligned}
l &= \log(L) = \sum_{i=1}^{10} \log \left(\int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} f_{Y_{ij}|U_i}(y_{ij}|u_i) \right) f_{U_i}(u_i)\, \mathrm{d}u_i \right) \\
&= \sum_{i=1}^{10} \log \left(\int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \exp(-\mu_{ij}) \right) \frac{1}{\sqrt{2\pi}\psi} \exp\Big(-\frac{u_i^2}{2\psi^2} \Big)\, \mathrm{d}u_i \right) \\
&= \sum_{i=1}^{10} \log \left(\int_{-\infty}^{\infty} \exp\Big(\sum_{j=1}^{n_i} (Z_{ij}u_i y_{ij} - \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}u_i + \eta_{ij})) \Big) \frac{1}{\sqrt{2\pi}\psi} \exp\Big(-\frac{u_i^2}{2\psi^2} \Big)\, \mathrm{d}u_i \right) \\
&\quad + \sum_{i=1}^{10} \sum_{j=1}^{n_i} ((\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \eta_{ij}) y_{ij} - \log(y_{ij}!)) \\
l &= \sum_{i=1}^{10} \log \left(\int_{-\infty}^{\infty} \exp\Big(\sum_{j=1}^{n_i} (Z_{ij}u_i y_{ij} - \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}u_i + \eta_{ij})) \Big) \frac{1}{\sqrt{2\pi}\psi} \exp\Big(-\frac{u_i^2}{2\psi^2} \Big)\, \mathrm{d}u_i \right) \\
&\quad + \boldsymbol{y}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}) - \sum_{i=1}^{10} \sum_{j=1}^{n_i} \log(y_{ij}!) \,. \qquad (4.9)
\end{aligned}$$

Det problematiske med l her er at man må integrere over u_i. Siden lmer gir paramerestimater for $\boldsymbol{\beta}$ og ψ så kan jeg optimere l med optim mhp. $\boldsymbol{\beta}$ og ψ og få oppgitt log likelihood og standardfeil til estimatene. Jeg bruker en Laplace approksimasjon i optim for å kunne beregne hvert integral.

4.2.1 Estimerte verdier for Tankskip

Man kan her sammenligne verdiene til Poisson og ZIP modellene med tabell 2.2 på side 19. For Poisson GLMM lar jeg Intercept være kovariaten som blir påvirket av de tilfeldige effektene.
ZIP modellen fikk færre signifikante kovariater etter at $\log(\rho)$ ble innført.

	Poisson*	ZIP*	Po GLMM	Po GLMM*
Verdier for β:				
Intercept	-4.71(0.94)	-1.06(0.45)	-4.74(0.96)	-4.74(0.72)
log(Age+2)	0.35(0.05)	0.33(0.05)	0.29(0.06)	0.35(0.05)
log(GT)	-0.15(0.04)	-0.15(0.03)	-0.14(0.04)	-0.14(0.04)
log(Sum.Insured)	0.20(0.06)		0.19(0.06)	0.20(0.05)
Stroke	0.26(0.09)	0.75(0.16)	0.27(0.09)	0.27(0.09)
$\hat{\psi}$			0.12(0.04)	0.12(0.04)
Verdier for γ:				
Intercept		-11.56(3.76)		
Stroke		1.33(0.54)		
log(Sum.Insured)		-0.79(0.25)		
Antall parametre	5	7	6	6
AIC	8 142.3	8 104.4	8 203.7	8 136.5
BIC	8 180.4	8 156.3	8 248.2	8 181.0

Tabell 4.1: Estimerte verdier for tankskip.

Her er Po GLMM en forkortelse for Poisson GLMM.

4.2.2 Estimerte verdier for alle typer skip

Man kan her sammenligne Poisson og ZIP modellen med tabell 2.3 på side 20. For Poisson GLMM er det kovariaten log(Age+2) som blir påvirket av de tilfeldige effektene.

	Poisson*	ZIP*	Po GLMM	Po GLMM*
Verdier for β:				
Intercept	-5.02(0.31)	-8.08(0.52)	-4.31(0.20)	-5.04(0.31)
log(Age+2)	0.23(0.02)	0.23(0.02)	0.15(0.02)	0.22(0.02)
log(GT)	-0.12(0.02)	0.05(0.02)	-0.11(0.02)	-0.12(0.02)
log(Sum.Insured)	0.12(0.02)	0.33(0.03)	0.07(0.01)	0.12(0.02)
log(HP)	0.16(0.02)		0.16(0.02)	0.17(0.02)
Stroke	0.37(0.03)	0.39(0.03)	0.43(0.03)	0.37(0.03)
$\hat{\psi}$			0.02(0.007)	0.02(0.008)
Verdier for γ:				
Intercept		-10.56(1.33)		
log(GT)		0.65(0.08)		
log(Sum.Insured)		0.55(0.10)		
log(HP)		-0.60(0.09)		
Antall parametre	6	8	7	7
AIC	37 757.0	37 531.7	38 218.6	37 742.5
BIC	37 810.3	37 602.8	38 280.8	37 804.7

Tabell 4.2: Estimerte verdier for alle typer skip.

Tabell 4.2 og tabell 4.1 på forrige side, og tabellene i kapittel 2 og 3, viser at $\log(\rho)$ gjør modellene markert bedre. Dette vises ved betydelig lavere AIC og BIC for begge datasettene, og færre signifikante kovariater i ZIP for tankskip. Så det var hensiktsmessig å innføre $\log(\rho)$ som motvekt.
Tabellene viser også at modellene basert på den Negativt Binomiske fordelingen får mindre AIC og BIC enn modellene basert på Poisson fordelingen. Nullforhøyde modeller gir lavere AIC og BIC enn en enkel Poisson GLMM.
Utifra tabellene for Poisson, NB, ZIP, ZINB og Poisson GLMM virker det unødvendig å bruke en Poisson modell med tilfeldige effekter for dette datasettet. Poisson GLMM er komplisert i forhold til den lille forbedringen man får fra å bruke modellen, spesielt når man ser at en Negativt Binomisk modell, som er mindre komplisert, gir en mye bedre modelltilpasning.

Et problem med forståelsen av de nullforhøyde modellene er at koeffisienten ϑ, som avhenger av kovariater og parametervektoren γ, kan sammenlignes med motvekten $\log(\rho)$. Det kan virke som om disse 2 størrelsene kan tolkes på samme måte, og at det dermed er problematisk å bruke begge i samme modell siden det kan være et overlapp i definisjonene. Som tidligere nevnt kan ϑ tolkes som sannsynligheten for at et skip er "sikkert", mens ρ tolkes som hvor sannsynlig det er at en skade blir meldt til forsikringsselskapet når en skade har skjedd.

Både ρ og ϑ bruker kovariatene til skipene når de beregnes. Men ρ blir beregnet med utgangspunkt i skadegradene og egenandelene til skipene, mens ϑ blir beregnet med utgangspunkt i skadefrekvensen til skipene. Parametrene til ZIP og ZINB endres også betydelig når $\log(\rho)$ innføres som motvekt i modellene, og AIC og BIC verdier minker markert. Derfor tror jeg ikke at et eventuelt overlapp er problematisk for modellene.

5
H-likelihood

I dette kapittelet vil jeg se på H-likelihood og undersøke om H-likelihood kan brukes til å undersøke generaliserte lineære miksede modeller(GLMM) for skadefrekvens i datasettet mitt. Referansen her er Lee *et al.* (2006). Forfatterne skiller mellom GLMM, hvor de tilfeldige effektene er normalfordelte, og HGLM hvor de tilfeldige effektene ikke nødvendigvis må være normalfordelte. Jeg vil i dette kapittelet, og i kapittel 6, ofte bruke "miksede modeller"som en fellesbetegnelse på GLMM og HGLM.

H-likelihood er en forholdsvis ny metode, fordelen med metoden er at man slipper å integrere ut de tilfeldige effektene for å beregne parameterestimater til de miksede modellene. Derfor gir H-likelihood inntrykk av å være en mindre komplisert metode sammenlignet med metodene som benytter numerisk integrasjon.
H-likelihood preges av at det ikke er noen god lærebok om emnet. Det er et komplisert emne og Lee *et al.* (2006) har til tider vært vanskelig å forstå.

5.1 Utvidet likelihood

Som for GLMM i forrige kapittel ønsker jeg her at likelihood funksjonen skal kunne håndtere 3 typer objekter:

- Ukjente parametre $\boldsymbol{\theta}$.
- Observerte data $\boldsymbol{y}$.
- Uobserverte tilfeldige størrelser $\boldsymbol{v}$.

Her er $\boldsymbol{v}$ tilfeldige effekter på samme måte som $\boldsymbol{u}$ i kapittel 4, men jeg bruker likevel et annet symbol siden jeg snart vil definere $\boldsymbol{v}$ som en transformasjon av $\boldsymbol{u}$.

Definisjon 5.1.1: Utvidet likelihood funksjon

Den utvidete likelihood funksjonen er definert ved

$$L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{y}, \boldsymbol{v}) = L(\boldsymbol{\theta}; \boldsymbol{y}) L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}|\boldsymbol{y}). \tag{5.1}$$

I definisjonen ovenfor blir ; brukt til å skille mellom parametre og observasjoner. Definisjonen er litt spesiell siden $\boldsymbol{v}$ er på begge sider av ; og blir behandlet både som en vektor med parametre og som en vektor med observasjoner. Men siden $\boldsymbol{v}$ er definert som en vektor med tilfeldige variabler samtidig som den er ukjent så er definisjonen nødvendig.
Det klassiske likelihood prinsippet sier at $L(\boldsymbol{\theta}; \boldsymbol{y})$ inneholder all relevant informasjon fra dataene om de faste parametrene $\boldsymbol{\theta}$.
Det utvidete likelihood prinsippet sier at $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{y}, \boldsymbol{v})$ inneholder all informasjon fra dataene om de uobserverte størrelsene $\boldsymbol{\theta}$ og $\boldsymbol{v}$.
$L(\boldsymbol{\theta}; \boldsymbol{y})$ kan brukes til å estimere $\boldsymbol{\theta}$ når $\boldsymbol{y}$ er kjent, mens $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}|\boldsymbol{y})$ kan brukes til å estimere $\boldsymbol{v}$ når $\boldsymbol{\theta}$ er kjent. $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}|\boldsymbol{y})$ inneholder ikke noe informasjon om $\boldsymbol{\theta}$, og jeg kan derfor ikke estimere $\boldsymbol{\theta}$ ved felles maksimering av $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{y}, \boldsymbol{v})$ mhp. $(\boldsymbol{\theta}, \boldsymbol{v})$. Å estimere $\boldsymbol{\theta}$ på denne måten vil bryte med det klassiske likelihood prinsippet og kan føre til selvmotsigelser. Blant annet har Lee *et al.* (2006) et eksempel som viser hvordan en felles estimering av $(\boldsymbol{\theta}, \boldsymbol{v})$ kan føre til meningsløse estimater.

For senere bruk definerer jeg nå noen størrelser til utvidet likelihood:

$$\text{Utvidet log likelihood: } l_e(\boldsymbol{\theta}, \boldsymbol{v}) = \log(L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{y}, \boldsymbol{v})), \tag{5.2}$$

$$l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}; \boldsymbol{y})),$$

$$l_e(\boldsymbol{\theta}, \boldsymbol{v}) = l(\boldsymbol{\theta}) + \log(f_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{y})) = l(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{v}) + \log(f_{\boldsymbol{\theta}}(\boldsymbol{v})). \tag{5.3}$$

Her er $f_{\boldsymbol{\theta}}(\boldsymbol{v}|\boldsymbol{y})$ fordelingsfunksjonen til $\boldsymbol{v}$ gitt observasjonene $\boldsymbol{y}$.

Siden $l(\boldsymbol{\theta})$ kan være vanskelig å beregne analytisk, dvs. integrere ut $\boldsymbol{v}$, så bruker Lee *et al.* (2006) ofte Laplace approksimasjonen som tilnærmet verdi. En første ordens Laplace approksimasjon til $l(\boldsymbol{\theta})$ mhp. v uttrykkes i dette kapittelet notasjonsmessig med

$$l(\boldsymbol{\theta}) \simeq p_v(l_e) = \left[l_e - \frac{1}{2} \log \left| \frac{D(l_e, v)}{2\pi} \right| \right] \Big|_{v=\hat{v}_\theta}, \quad D(l_e, v) = -\frac{\partial^2 l_e}{\partial v^2} .$$

5.2 Kanonisk vekt og H-likelihood

For noen generelle klasser av modeller kan man bruke utvidet likelihood til å gi en felles inferens for tilfeldige og faste parametre.

Definisjon 5.2.1

Den tilfeldige variabelen v i $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}, \boldsymbol{y})$ er en kanonisk vekt hvis profil likelihood til $\boldsymbol{\theta}$ fra den utvidete likelihood funksjonen er proposjonal med den marginale likelihood funksjonen $L(\boldsymbol{\theta}; \boldsymbol{y})$. Hvis v i $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}, \boldsymbol{y})$ er en kanonisk vekt så kaller vi $L(\boldsymbol{\theta}, \boldsymbol{v}; \boldsymbol{v}, \boldsymbol{y})$ en H-likelihood. Dermed er H-likelihood et spesialtilfelle av utvidet likelihood.
Hvis en kanonisk vekt eksisterer så er definisjonen på H-likelihood umiddelbar. Den kanoniske vekten er informasjonsnøytral om $\boldsymbol{\theta}$.
$H(\boldsymbol{\theta}, \boldsymbol{v})$: H-likelihood.
$h(\boldsymbol{\theta}, \boldsymbol{v})$: H log likelihood.

"H"i H-likelihood står for "hierarkisk".
Hvis den kanoniske vekten eksisterer så har den en del nyttige egenskaper.

Definisjon 5.2.2: Egenskaper til den kanoniske vekten

La $I(\hat{\boldsymbol{\theta}})$ være den observerte Fisher informasjonsmatrisen til $\hat{\boldsymbol{\theta}}$, hvor $\hat{\boldsymbol{\theta}}$ er SME fra $L(\boldsymbol{\theta}; \boldsymbol{y})$.
La så $I_h^{-1}(\hat{\boldsymbol{\theta}}, \hat{v}) = \begin{pmatrix} I_h^{11} & I_h^{12} \\ I_h^{21} & I_h^{22} \end{pmatrix}$ være invers av den observerte Fisher informasjonsmatrisen til $(\hat{\boldsymbol{\theta}}, \hat{v})$ fra $H(\boldsymbol{\theta}, v; \boldsymbol{y}, v)$ hvor v er en kanonisk vekt. Her korresponderer I_h^{11} med $\boldsymbol{\theta}$-delen av matrisen. Da får man at

- $\hat{\boldsymbol{\theta}}$, som er SME fra $L(\boldsymbol{\theta}; \boldsymbol{y})$, stemmer overens med $\hat{\boldsymbol{\theta}}$ fra en felles maksimering av $L(\boldsymbol{\theta}, v; \boldsymbol{y}, v)$ mhp. $(\boldsymbol{\theta}, v)$.
- Informasjonsmatrisene for $\hat{\boldsymbol{\theta}}$ fra de to likelihood funksjonene stemmer også overens. Dette betyr at $I^{-1} = I_h^{11}$.

Fra definisjonen ovenfor ser man at inferens fra H-likelihood kan behandles på samme måte som inferens fra vanlig likelihood. Lee og Nelder (2009) hevder at man vha. H-likelihood ikke bare kan få inferens for $\boldsymbol{\theta}$, men også for v i tillegg til felles inferens for $(\boldsymbol{\theta}, v)$.

Den kanoniske vekten fører til at hverken $L(\boldsymbol{\theta}, \hat{v}_{\boldsymbol{\theta}}; v|\boldsymbol{y})$ eller $\hat{v}_\theta$ inneholder noe informasjon om $\boldsymbol{\theta}$. Dette er nødvendig for at det klassiske likelihood prinsippet skal gjelde. Det er da mulig å finne felles inferens for $(\boldsymbol{\theta}, v)$ vha. formel 5.1 på side 50.

Som nevnt ovenfor er den marginale log likelihood $l(\boldsymbol{\theta}; \boldsymbol{y})$ tilnærmet med Laplace approksimasjonen $p_v(l_e)$. Ofte kan man finne at v er kanonisk hvis $I(\hat{v}_\theta)$ i $p_v(l_e)$ ikke inneholder $\boldsymbol{\theta}$. Hvis de faste parametrene består av 2 underrom, $(\boldsymbol{\theta}, \boldsymbol{\phi})$, så er v kanonisk for $\boldsymbol{\theta}$ hvis $I(\hat{v}_{\theta,\phi})$ ikke inneholder $\boldsymbol{\theta}$.
Hvis v er kanonisk for $\boldsymbol{\beta}$ så fører dette blant annet til at Laplace approksimasjonen på h mhp. v ikke fjerner noe informasjon om $\boldsymbol{\beta}$ fra h.

5.3 Hierarkisk GLM

Definisjon 5.3.1: HGLM

En hierarkisk GLM er definert på følgende måte:
La $\boldsymbol{Y}$ være responsvariabler som gitt den tilfeldige variabelen u har en fordeling fra GLM familien. Denne fordelingen oppfyller at

- $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{u}] = \boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{Y}|\boldsymbol{u}) = \varphi\nu(\boldsymbol{\mu})$.
- Kjernen til log likelihood funksjonen er gitt ved $\sum[y_i\theta_i - A(\theta_i)]/\varphi$ hvor $\theta_i = \theta(\mu_i)$ er den kanoniske parameteren.
- Den lineære prediktoren tar formen $g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Zv}$. Her er $v = v(u)$ en monoton funksjon av u. Som tidligere er $\boldsymbol{\beta}$ de faste parametrene, mens $\boldsymbol{X}$ og $\boldsymbol{Z}$ er kovariatmatriser.

Den tilfeldige variabelen u har en fordeling som er konjugert med en GLM familie av fordelinger med parameter λ. Dette betyr at hvis u er en apriori fordeling så vil den aposteriori fordelingen tilhøre samme familie av fordelinger som fordelingen til u.

For å opprettholde invarians av inferens mhp. ekvivalente modeller så må man ovenfor definere H-likelihood på den spesifikke vekten $v(u)$. På denne vekten er de tilfeldige effektene kombinert med de faste parametrene $\boldsymbol{\beta}$ i den lineære prediktoren. Denne vekten kalles svak kanonisk vekt. Den svakt kanoniske vekten kan alltid defineres hvis man kan definere den lineære prediktoren. Grunnen til at den svake kanoniske vekten introduseres er at definisjonen til den kanoniske vekten kan være for restriktiv.
SME til $\boldsymbol{\beta}$ er invariant mhp. ekvivalente modeller. Man kan finne SME til $\boldsymbol{\beta}$ ved felles maksimering hvis v er kanonisk til $\boldsymbol{\beta}$. Så hvis det eksisterer en kanonisk vekt for $\boldsymbol{\beta}$ så tilfredsstiller denne vekten også den svake kanoniske egenskapen ved at estimatoren til $\boldsymbol{\beta}$ fremdeles er invariant mhp. ekvivalente modeller.
For inferens fra HGLM bør H log likelihood defineres som

$$h = \log(f(\boldsymbol{y}|\boldsymbol{v}, \boldsymbol{\beta}, \phi)) + \log(f(\boldsymbol{v}|\lambda)), \tag{5.4}$$

hvor (ϕ, λ) er dispersjonsparametre.

Et viktig aspekt med HGLM er den fleksible spesifikasjonen av fordelingen til de tilfeldige effektene u. Den svake kanoniske vekten gir en representasjon av log likelihood for u som kan skrives som

$$\sum_j [\psi_M \theta_M(u_j) - A_M(\theta_M(u_j))]/\lambda. \tag{5.5}$$

Her er θ_M og A_M kjente funksjoner, mens ψ_M blir tilegnet en verdi slik at λ er den eneste parameteren i formel 5.5. I konjugerte fordelinger vil $\mathbb{E}(u) = \psi_M$ og $\text{Var}(u) = \phi v_M(\psi_M)$. Her er ϕ en funksjon av parameteren λ.

Et spesialtilfelle av HGLM er konjugert HGLM. Ved å la funksjonene $A_M(u) = A(u)$ og $A_M(\theta_M) = A(\theta)$ får man en konjugert log likelihood for de tilfeldige effektene:

$$\sum_j [\psi_M \theta(u_j) - A(\theta(u_j))]/\lambda. \tag{5.6}$$

En HGLM hvor u har en konjugert log likelihood kalles konjugert HGLM. Hvis fordelingen til u er den konjugerte fordelingen til $y|u$ så er modellen en konjugert HLGM.

Når man skal undersøke inferens for de ulike størrelsene i HGLM så oppgir Lee *et al.* (2006) noen grunnprinsipper. De er at man bør bruke H log likelihood funksjonen h for å få inferens om v, den marginale log likelihood funksjonen $l = l(\phi, \lambda, \boldsymbol{\beta}; \boldsymbol{y})$ for å få inferens om $\boldsymbol{\beta}$ og $l(\phi, \lambda, \hat{\boldsymbol{\beta}}; \boldsymbol{y})$ for å få inferens om dispersjonsparametrene. Her er $\hat{\boldsymbol{\beta}}$ SME til $\boldsymbol{\beta}$ og $l(\phi, \lambda, \hat{\boldsymbol{\beta}}; \boldsymbol{y})$ er l innsatt $\hat{\boldsymbol{\beta}}$.
For miksede modeller kan man ikke bruke l til å finne et analytisk uttrykk for $\hat{\boldsymbol{\beta}}$ med mindre l kommer fra en lmm. Derfor foreslår Lee *et al.* (2006) å bruke $p_{\boldsymbol{\beta}}(l)$ som en tilnærming til $l(\phi, \lambda, \hat{\boldsymbol{\beta}}; \boldsymbol{y})$.
Hvis l er vanskelig å beregne analytisk så kan $p_v(h)$ brukes som en approksimasjon. Da kan også $p_{(\boldsymbol{\beta},v)}(h)$ brukes som en approksimasjon på $p_{\boldsymbol{\beta}}(l)$ og dermed på $l(\phi, \lambda, \hat{\boldsymbol{\beta}}; \boldsymbol{y})$. Her er $p_{(\boldsymbol{\beta},v)}(h)$ en Laplace approksimasjon til h mhp. vektoren $(\boldsymbol{\beta}, \boldsymbol{v})$.
For lmm gir Lee *et al.* (2006) at $p_{(\boldsymbol{\beta},v)}(h) = p_{\boldsymbol{\beta}}(l)$. For miksede modeller hvor $y|u$ ikke er normalfordelt er $p_{(\boldsymbol{\beta},v)}(h)$ approksimalt lik $p_{\boldsymbol{\beta}}(l)$.

Utifra teorien jeg har gitt så langt så ligner fremgangsmåten til H likeli-

hood på fremgangsmåten til PQL tidligere i masteroppgaven. Lee *et al.* (2006) hevder at gitt dispersjonsparametrene (ϕ, λ) er PQL estimatorene for $(\boldsymbol{v}, \boldsymbol{\beta})$ de samme som H-likelihood estimatorene som i fellesskap maksimerer h. I Poisson og binomisk GLMM gitt $(\boldsymbol{v}, \boldsymbol{\beta})$ er PQL estimatorene for dispersjonsparametrene forskjellige fra de første ordens H-likelihood estimatorene som maksimerer $p_{(v,\boldsymbol{\beta})}(h)$. Dette kommer av at PQL estimatorene ignorerer $\frac{\partial \hat{v}}{\partial \phi}$ og $\frac{\partial \hat{v}}{\partial \lambda}$, dvs. at PQL ikke tar hensyn til hvordan dispersjonsparametrene (ϕ, λ) påvirker estimatene av de tilfeldige effektene.

5.4 Modeller beregnet ved H-likelihood

Den eneste funksjonen jeg har funnet i R som beregner modeller med H-likelihood er hnlmix fra pakken repeated. Men jeg har ikke studert denne funksjonen siden jeg er interessert i å se hvordan beregningsmetodene for H-likelihood fungerer.

5.4.1 Poisson-log-Gamma

Dette er en modell som blir mye brukt i Lee *et al.* (2006) som eksempel på HGLM. Denne modellen har flere egenskaper som forenkler parameterestimeringen for H-likelihood. Derfor har jeg undersøkt denne modellen som et alternativ til Poisson-Normal modellen i kapittel 4.2.
Modellen jeg undersøker er gitt på følgende form:

$$\begin{aligned} Y_{ij}|u_i &\sim \text{Poisson}(\mu_{ij});\ i = 1,\ldots,10;\ j = 1,\ldots,n_i\ , \\ u_i &\sim \text{iid. Gamma}\Big(\frac{1}{\lambda}, \lambda\Big)\ , \\ v_i &= \log(u_i)\ , \\ \mu_{ij} &= \mathbb{E}[Y_{ij}|u_i] = u_i e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\eta_{ij}} = e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\eta_{ij}+v_i}\ . \end{aligned}$$

Den marginale variansen for Y er $\text{Var}(Y_{ij}) = e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\eta_{ij}}(1 + \lambda e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\eta_{ij}})$.
For denne modellen kan man faktisk integrere ut de tilfeldige effektene, dvs. $\boldsymbol{u} = (u_1, \ldots, u_{10})'$, hvis man setter opp likelihood funksjonen på formen til formel 4.1 på side 37. Dette kan sees i sammenheng med den Negativt Binomiske modellen fra avsnitt 2.1.3. Man får da følgende marginale log

likelihood for $\boldsymbol{Y}$:

$$l(\boldsymbol{\beta}, \lambda; \boldsymbol{y}) = \sum_{i=1}^{10} \Big[-\frac{1}{\lambda}\log(\lambda) - \log\Big(\Gamma\Big(\frac{1}{\lambda}\Big)\Big) + \sum_{j=1}^{n_i}(y_{ij}(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \eta_{ij}) - \log(y_{ij}!)) \\ - \Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}\Big)\log\Big(\frac{1}{\lambda} + \sum_{j=1}^{n_i} e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta}+\eta_{ij}}\Big) + \log\Big(\Gamma\Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}\Big)\Big)\Big] \tag{5.7}$$

Jeg kan dermed bruke denne funksjonen til å estimere $\boldsymbol{\beta}$. Men jeg må også estimere λ og $\boldsymbol{v}$. Det er nyttig å prøve å finne et analytisk estimat av $\boldsymbol{v}$ slik at man kan undersøke om $\boldsymbol{v}$ er kanonisk for $\boldsymbol{\beta}$ og λ.
Jeg ønsker å bruke fremgangsmåten som er beskrevet i kapittel 5.3. Først ser jeg på log likelihood til $\boldsymbol{u}$:

$$f(u_i; \lambda) = \Big(\frac{1}{\lambda}\Big)^{\frac{1}{\lambda}} \frac{1}{\Gamma(\frac{1}{\lambda})} u_i^{\frac{1}{\lambda}-1} e^{-\frac{u_i}{\lambda}},$$

$$l(\lambda; \boldsymbol{u}) = \sum_{i=1}^{10} \Big[\Big(\frac{1}{\lambda} - 1\Big)\log(u_i) - \frac{u_i}{\lambda} - \log(\Gamma\Big(\frac{1}{\lambda}\Big)) - \Big(\frac{1}{\lambda}\Big)\log(\lambda)\Big].$$

Per definisjon er $\mathbb{E}[u_i] = \lambda\frac{1}{\lambda} = 1$ og $\mathrm{Var}(u_i) = \lambda$. Utifra formel 5.5 på side 54 kan man se at $\psi_M = 1$, $\theta_M(u_i) = \log(u_i)$ og $A_M() = \exp()$. Ved å sammenligne med Poisson fordelingen ser man at $A_M() = A()$ og $\theta_M() = \theta() = \log()$. Dette fører til at Poisson-log-Gamma modellen går under betegnelsen konjugert HGLM.
Etter å ha transformert gjennom vekten $u_i = e^{v_i}$ får jeg følgende log likelihood for de tilfeldige effektene:

$$l(\lambda; \boldsymbol{v}) = \sum_{i=1}^{10} \Big[(v_i - e^{v_i})/\lambda - \log\Big(\Gamma\Big(\frac{1}{\lambda}\Big)\Big) - \frac{1}{\lambda}\log(\lambda)\Big].$$

Ved å sette opp H log likelihood som formel 5.4 på side 53 får jeg følgende funksjon:

$$\begin{aligned} h =& l(\lambda; \boldsymbol{v}) + l(\boldsymbol{\beta}; \boldsymbol{y}|\boldsymbol{v}) \\ =& c + \sum_{i=1}^{10}\Big(\sum_{j=1}^{n_i}[y_{ij}v_i - \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \eta_{ij} + v_i)] + (v_i - e^{v_i})/\lambda\Big), \end{aligned} \tag{5.8}$$

hvor c er en konstant. Det er for denne modellen mulig å løse $\frac{\partial h}{\partial v_i}|_{v_i=\hat{v}_i} = 0$ analytisk. Jeg får da følgende estimat for de tilfeldige effektene:

$$\hat{v}_i = \log\Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}\Big) - \log\Big(\frac{1}{\lambda} + \sum_{j=1}^{n_i} \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \eta_{ij})\Big),\ i = 1,\ldots,10\,. \quad (5.9)$$

Det er dermed mulig å undersøke om v er kanonisk for parametrene $(\lambda, \boldsymbol{\beta})$.

$$-\frac{\partial^2 h}{\partial v_i^2} = \sum_{j=1}^{n_i} \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \eta_{ij} + v_i) + \frac{e^{v_i}}{\lambda},$$
$$-\frac{\partial^2 h}{\partial v_i^2}\Big|_{v_i=\hat{v}_i} = \sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}.$$

Utifra dette ser man at v er kanonisk for $\boldsymbol{\beta}$, men ikke for λ.
Hvis man deriverer l mhp. $\boldsymbol{\beta}$ får man følgende ligning:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{10}\Big[\sum_{j=1}^{n_i} y_{ij}\boldsymbol{X}_{ij} - \frac{\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}}{\frac{1}{\lambda} + \sum_{j=1}^{n_i}\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \eta_{ij})}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \eta_{ij})\Big].$$

For denne funksjonen er det ikke mulig å løse $\frac{\partial l}{\partial \boldsymbol{\beta}}\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$ analytisk. Derfor beregner jeg $\boldsymbol{\beta}$ numerisk og bruker $p_{\boldsymbol{\beta}}(l)$ til å finne estimat for λ. Som startverdier for den numeriske beregningen av $\boldsymbol{\beta}$ bruker jeg estimerte verdier fra en Poisson GLM.

$$p_{\boldsymbol{\beta}}(l) = \Big[l - \frac{1}{2}\log(|D(l,\boldsymbol{\beta})/(2\pi)|)\Big]\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},\ D(l,\boldsymbol{\beta}) = -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\,.$$
$$D(l,\boldsymbol{\beta}) = \sum_{i=1}^{10}\frac{\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}}{\frac{1}{\lambda} + \sum_{j=1}^{n_i} e^{(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\eta_{ij})}}\Big[\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}' e^{(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\eta_{ij})}$$
$$-\frac{1}{\frac{1}{\lambda} + \sum_{j=1}^{n_i} e^{(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\eta_{ij})}}\Big(\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}e^{(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\eta_{ij})}\Big)\Big(\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}'e^{(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\eta_{ij})}\Big)\Big],$$

hvor $D(l,\boldsymbol{\beta})$ er en $p \times p$ matrise. For å kunne utføre beregningene i R for dette datasettet er det nødvendig å bruke Stirlings formel på $\log(\Gamma\Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda}\Big))$ i formel 5.7 på forrige side, og dermed også i $p_{\hat{\boldsymbol{\beta}}}(l)$. Stirlings formel er gitt ved $\log(\Gamma(x)) \approx (x - \frac{1}{2})\log(x) + \frac{1}{2}\log(2\pi) - x$. Ved å gjøre dette får man følgende

funksjon for estimering av λ:

$$p_{\hat{\boldsymbol{\beta}}}(l) = \sum_{i=1}^{10} \Big[\sum_{j=1}^{n_i} (y_{ij}(\boldsymbol{X}'_{ij}\hat{\boldsymbol{\beta}} + \eta_{ij}) - \log(y_{ij}!) - y_{ij}) \\ - \Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda} \Big) \log \Big(\frac{1}{\lambda} + \sum_{j=1}^{n_i} e^{(\boldsymbol{X}'_{ij}\hat{\boldsymbol{\beta}} + \eta_{ij})} \Big) + \Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda} - 0.5 \Big) \log \Big(\sum_{j=1}^{n_i} y_{ij} + \frac{1}{\lambda} \Big) \\ + \frac{1}{2}\log(2\pi) - \frac{1}{\lambda} - \frac{1}{\lambda}\log(\lambda) - \log\Big(\Gamma\Big(\frac{1}{\lambda}\Big)\Big) \Big] - \frac{1}{2}\log(|D(l, \hat{\boldsymbol{\beta}})|). \tag{5.10}$$

Estimerte verdier

For tankskip beregner jeg estimater og standardfeil til $\boldsymbol{\beta}$ fra optimering av l. For λ får jeg estimatet og standardfeilen fra optimering av $p_{\hat{\boldsymbol{\beta}}}(l))$.

	Poisson-log-Gamma	Poisson-log-Gamma*
Verdier for $\boldsymbol{\beta}$:		
Intercept	-4.73(0.93)	-4.73(0.95)
log(Age+2)	0.29(0.06)	0.35(0.06)
log(GT)	-0.14(0.04)	-0.14(0.04)
log(Sum.Insured)	0.19(0.06)	0.20(0.06)
Stroke	0.27(0.09)	0.28(0.09)
$\hat{\lambda}$	0.019(0.01)	0.017(0.01)
Antall parametre	6	6
AIC	8205.8	8138.5
BIC	8250.3	8183.1

Tabell 5.1: Estimerte verdier for tankskip.

For alle typer skip konvergerte ikke funksjonen optim i R når jeg skulle estimere λ. Jeg måtte derfor bruke en annen funksjon, noe som førte til at jeg ikke fikk oppgitt standardfeil til $\hat{\lambda}$.
For både tankskip og alle typer skip har jeg brukt formel 5.7 på side 56 når jeg skal beregne AIC og BIC.

	Poisson-log-Gamma	Poisson-log-Gamma*
Verdier for $\boldsymbol{\beta}$:		
Intercept	-4.32(0.30)	-5.04(0.31)
log(Age+2)	0.15(0.02)	0.23(0.02)
log(GT)	-0.11(0.02)	-0.12(0.02)
log(Sum.Insured)	0.07(0.02)	0.12(0.02)
log(HP)	0.16(0.02)	0.17(0.02)
Stroke	0.43(0.03)	0.37(0.03)
$\hat{\lambda}$	0.005	0.005
Antall parametre	7	7
AIC	38 221.6	37 747.0
BIC	38 283.8	37 809.3

Tabell 5.2: Estimerte verdier for alle typer skip.

Siden formel 5.9 på side 57 gir et estimat av de tilfeldige effektene oppgir jeg også $\hat{u} = e^{\hat{v}}$:

	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{u}_4$	$\hat{u}_5$	$\hat{u}_6$	$\hat{u}_7$	$\hat{u}_8$	$\hat{u}_9$	$\hat{u}_{10}$
Tankskip	1.13	0.97	0.91	1.07	1.17	0.89	1.03	0.82	1.06	0.89
Tankskip*	1.12	0.98	0.92	1.08	1.15	0.89	1.00	0.83	1.07	0.90
Alle skip	1.05	1.00	0.97	1.03	1.07	0.92	1.00	0.89	1.02	1.02
Alle skip*	1.06	1.00	0.96	1.02	1.05	0.90	0.97	0.89	1.04	1.06

Tabell 5.3: Estimerte verdier av de tilfeldige effektene $\boldsymbol{u}$.

Siden formel 5.7 på side 56 bare avhenger av parametrene $\boldsymbol{\beta}$ og λ har jeg også sammenlignet H-likelihood metoden med en felles maksimering av formel 5.7 på side 56 mhp. $(\boldsymbol{\beta}, \lambda)$. For de 4 modellene ovenfor ble estimatene og tilhørende standardfeil ganske like. For tankskip ble λ estimert litt lavere i den felles maksimeringen enn i modellene ovenfor(0.016 og 0.015 mot 0.019 og 0.017). Parameteren til Intercept ble estimert litt høyere for 3 av 4 modeller i den felles maksimeringen (forskjellen varierte mellom 0.02 til 0.1), unntaket var tankskip uten $\log(\rho)$ som motvekt.

5.4.2 Poisson-Normal

Jeg vil nå bruke H-likelihood på modellen fra formel 4.8 på side 44. Denne modellen vil heretter bli kalt Poisson-Normal. For denne modellen kan jeg ikke integrere ut de tilfeldige effektene analytisk for å få en marginal log likelihood funksjon for $\boldsymbol{y}$. Derfor er det interessant å se om bruk av H-likelihood forenkler beregningene for denne modellen. I formel 4.8 på side 44 lot jeg kovariatmatrisen for de tilfeldige effektene, $\boldsymbol{Z}$, bestå av kovariaten log(Age+2) for alle typer skip. For tanskip lot jeg $\boldsymbol{Z}$ bare bestå av Intercept. Jeg bruker de samme kovariatene her, slik at jeg kan sammenligne resultatene her med resultatene fra kapittel 4.2.
For denne modellen oppgir Lee *et al.* (2006, side 180) at $u = v$ er den kanoniske vekten. Ved å sette opp modellen får jeg følgende log likelihood for $\boldsymbol{y}|\boldsymbol{v}$ og $\boldsymbol{v}$:

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{v}; \boldsymbol{y}|\boldsymbol{v}) &= \sum_{i=1}^{10}\sum_{j=1}^{n_i}(y_{ij}\log(\mu_{ij}) - \mu_{ij} - \log(y_{ij}!)) \\ &= \sum_{i=1}^{10}\sum_{j=1}^{n_i}(y_{ij}(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}) - \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}) - \log(y_{ij}!)), \\ l(\psi; \boldsymbol{v}) &= \sum_{i=1}^{10}\left(-\frac{v_i^2}{2\psi^2} - \frac{1}{2}\log(2\pi) - \log(\psi)\right). \end{aligned}$$

Ligningen for H log likelihood er som vanlig gitt ved $h = l(\boldsymbol{\beta}, \boldsymbol{v}; \boldsymbol{y}|\boldsymbol{v}) + l(\psi; \boldsymbol{v})$. Første og andrederiverte for h mhp. v_i blir

$$\begin{aligned} \frac{\partial h}{\partial v_i} &= \sum_{j=1}^{n_i}(y_{ij}Z_{ij} - Z_{ij}\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij})) - \frac{v_i}{\psi^2}\,, \\ \frac{\partial^2 h}{\partial v_i^2} &= -\sum_{j=1}^{n_i} Z_{ij}^2\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}) - \frac{1}{\psi^2}\,. \end{aligned}$$

Som tidligere vil $\frac{\partial h}{\partial v_i}|_{v_i=\hat{v}_i} = 0,\ i = 1,\ldots,10$ gi SME til v gitt $\boldsymbol{\beta}$ og ψ. Men denne ligningen kan ikke løses analytisk her, og man må derfor finne et numerisk estimat. Siden $\frac{\partial^2 h}{\partial v_i^2} < 0\ \forall\ v_i$ vil det numeriske estimatet av $\hat{\boldsymbol{v}} = (\hat{v}_1, \ldots, \hat{v}_{10})'$ maksimere h.

For å beregne SME til $\boldsymbol{\beta}$ må jeg bruke $p_v(h)$ siden log likelihood funksjonen l til $\boldsymbol{y}$ ikke kan beregnes analytisk. Jeg får da

$$p_v(h) = c + \sum_{i=1}^{10} \left(\sum_{j=1}^{n_i} (y_{ij}(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}\hat{v}_i) - \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}\hat{v}_i + \eta_{ij})) - \frac{\hat{v}_i^2}{2\psi^2} - \log(\psi) \right.$$
$$\left. + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sum_{j=1}^{n_i} Z_{ij}^2 \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}\hat{v}_i + \eta_{ij}) + \frac{1}{\psi^2} \right) \right).$$

For å beregne SME til ψ kan jeg bruke $p_{(\boldsymbol{\beta},v)}(h)$. Her må jeg beregne $D(h,(\boldsymbol{\beta},\boldsymbol{v}))$ som er en $(p+10)\times(p+10)$ matrise og er satt sammen av følgende funksjoner for $i = 1,\ldots,10$:

$$\begin{aligned}
-\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} &= \sum_{i=1}^{10}\sum_{j=1}^{n_i} \boldsymbol{X}_{ij}\boldsymbol{X}'_{ij} \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}) \,, \\
-\frac{\partial^2 h}{\partial v_i^2} &= \sum_{j=1}^{n_i} Z_{ij}^2 \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}) + \frac{1}{\psi^2} \,, \\
-\frac{\partial^2 h}{\partial v_i \partial\boldsymbol{\beta}} &= -\frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial v_i} = \sum_{j=1}^{n_i} \boldsymbol{X}_{ij} Z_{ij} \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + Z_{ij}v_i + \eta_{ij}).
\end{aligned} \tag{5.11}$$

$D(h,(\boldsymbol{v},\boldsymbol{\beta}))$ er sammensatt av 4 matriser basert på funksjonene ovenfor. Formen til $D(h,(\boldsymbol{v},\boldsymbol{\beta}))$ blir da

$$\begin{bmatrix} -\dfrac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & -\dfrac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{v}'} \\ -\dfrac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{\beta}'} & -\dfrac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}'} \end{bmatrix}. \tag{5.12}$$

Her er $-\frac{\partial^2 h}{\partial\boldsymbol{v}\partial\boldsymbol{v}'}$ en matrise bestående av 0 med unntak av diagonalen som består av formel 5.11 for $i = 1,\ldots,10$. Innsatt estimater får jeg følgende funksjon som kan estimere ψ:

$$p_{(\boldsymbol{v},\boldsymbol{\beta})}(h) = c - \sum_{i=1}^{10}\left(\frac{\hat{v}_i^2}{2\psi^2} + \log(\psi)\right) - \frac{1}{2}\log\left(\left|\frac{D(h,(\boldsymbol{v},\boldsymbol{\beta}))}{(2\pi)}\right|\right)\Bigg|_{(\boldsymbol{v},\boldsymbol{\beta})=(\hat{\boldsymbol{v}},\hat{\boldsymbol{\beta}})}$$

For å utføre beregningene har jeg brukt følgende algoritme. Noh *et al.* (2005) brukte en lignende algoritme for en mer komplisert modell:

1 Velg initialverdier for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{v}^{(0)}$ og $\psi^{(0)}$. La $m = 0$.

2 Beregn $\boldsymbol{v}^{(m+1)}$ ved å maksimere h gitt $\boldsymbol{\beta}^{(m)}$ og $\psi^{(m)}$.

3 Beregn $\boldsymbol{\beta}^{(m+1)}$ ved å maksimere $p_v(h)$ gitt $\boldsymbol{v}^{(m+1)}$ og $\psi^{(m)}$.

4 Beregn $\psi^{(m+1)}$ ved å maksimere $p_{(v,\boldsymbol{\beta})}(h)$ gitt $\boldsymbol{v}^{(m+1)}$ og $\boldsymbol{\beta}^{(m+1)}$.

5 Hvis ønsket konvergens er oppnådd så avslutt. Hvis ikke ønsket konvergens er oppnådd, la $m = m + 1$ og begynn på nytt fra punkt 2.

Ønsket konvergens lar jeg være at den maksimale endringen i estimatene mellom iterasjon m og $m - 1$ skal være 10^{-6}. Noh *et al.* (2005) brukte også dette kriteriet. Som initialverdier har jeg brukt estimerte verdier fra en Poisson GLM for $\boldsymbol{\beta}^{(0)}$, nullvektoren $\mathbf{0}$ for $\boldsymbol{v}^{(0)}$ og verdien 0.1 for $\psi^{(0)}$.
Noh *et al.* (2005) bruker en form for IWLS-algoritme til å utføre hver maksimering. Jeg har valgt å ikke gjøre dette, siden det er et godt utvalg av maksimeringsmetoder i R. Hver maksimering skjer med funksjonen optim i R, med unntak for ψ hvor jeg må bruke optimize for alle typer skip. Standardfeil er beregnet vha. optim, gjennom funksjonen $p_v(h)$ for $\boldsymbol{\beta}$ og gjennom funksjonen $p_{(v,\boldsymbol{\beta})}(h)$ for ψ.

Estimerte verdier

Tabell 5.4 på neste side oppgir estimater og standardfeil for Poisson-Normal modellen beregnet ved H-likelihood. Jeg er interessert i å sammenligne estimatene her med Poisson GLMM estimatene fra tabell 4.1 på side 46 og tabell 4.2 på side 47. Derfor har jeg ikke beregnet log likelihood verdi, og dermed ikke AIC og BIC, for disse modellene. Siden jeg bare vil sammenligne estimater har jeg bare sett på modellen hvor Days.covered er eneste motvekt.
For tankskip holdt det med 20 iterasjoner for å oppnå ønsket konvergens. For alle typer skip greide jeg pga. manglende datakraft bare å gjøre 4 iterasjoner og oppnådd konvergens varierte mellom 10^{-3} og 10^{-5} for disse estimerte verdiene.

Når man sammenligner tabell 5.4 på neste side med verdiene til Poisson GLMM i kapittel 4 så er estimatene og standardfeilene til $\boldsymbol{\beta}$ ganske like. Den eneste store forskjellen er standardfeilen til parameteren Intercept for alle

typer skip(0.1 i forskjell). Variansparametrene er også forholdsvis like både for tankskip og alle typer skip.

	Tankskip	Alle typer skip
Verdier for $\boldsymbol{\beta}$:		
Intercept	-4.74(0.92)	-4.31(0.30)
log(Age+2)	0.29(0.05)	0.14(0.02)
log(GT)	-0.14(0.04)	-0.11(0.02)
log(Sum.Insured)	0.19(0.06)	0.07(0.02)
log(HP)		0.16(0.02)
Stroke	0.27(0.09)	0.43(0.03)
$\hat{\psi}$	0.13(0.03)	0.02
Antall parametre	6	7

Tabell 5.4: Estimerte verdier for Poisson-normal modellen.

Nedenfor er en tabell med de estimerte verdiene av de tilfeldige effektene for tankskip. For å kunne sammenligne med Poisson-log-Gamma modellen har jeg oppgitt $\exp(\hat{\boldsymbol{u}})$ siden $u = v$ for Poisson-normal. Estimatene for de to modellene er jevnt over litt forskjellige, men likevel ganske like.

	$e^{\hat{u}_1}$	$e^{\hat{u}_2}$	$e^{\hat{u}_3}$	$e^{\hat{u}_4}$	$e^{\hat{u}_5}$	$e^{\hat{u}_6}$	$e^{\hat{u}_7}$	$e^{\hat{u}_8}$	$e^{\hat{u}_9}$	$e^{\hat{u}_{10}}$
Tankskip	1.15	0.98	0.92	1.08	1.18	0.90	1.04	0.83	1.07	0.91

Tabell 5.5: Estimerte verdier av de tilfeldige effektene for tankskip.

5.4.3 Negativ Binomisk-Normal

I tidligere kapitler har jeg sett at Negativt Binomiske modeller gir en bedre tilpasning for dette datasettet enn det Poisson modeller gjør. Derfor vil jeg her undersøke hvor godt en Negativt Binomisk modell med tilfeldige effekter passer til datasettet. Teorien jeg bruker her vil være den samme som for de to

modellene tidligere i kapittelet. Modellen jeg vil undersøke her er på formen

$$\begin{aligned} Y_{ij}|u_i &\sim \text{Neg.bin}(\mu_{ij}, \varsigma);\ i = 1,\ldots,10;\ j = 1,\ldots,n_i\ , \\ u_i &\sim \text{iid. } \mathcal{N}(0,\sigma^2)\ , \\ v_i &= v(u_i) = u_i\ , \\ \mu_{ij} &= \mathbb{E}[Y_{ij}|u_i] = \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + \eta_{ij} + v_i)\ . \end{aligned}$$

For denne modellen har jeg ikke funnet noen definisjon på den kanoniske vekten i litteraturen til H-likelihood. I tabell 6.2 fra Lee *et al.* (2006, side 180) blir $u = v$ brukt i de tre modellene der u er normalfordelt. Siden u her også er normalfordelt er dette et argument for å la $u = v$. I tillegg fører $u = v$ til at den marginale forventningen til Y blir den samme som for Poisson-Normal siden $\mathbb{E}[Y|u] = \mu$ for både Poisson-Normal og NB-Normal. Derfor antar jeg at $v = v(u) = u$ er den kanoniske vekten for NB-Normal.
Fra tidligere i masterogaven vet jeg at NB-modellen er en GLM hvis ς er en konstant. Det virker derfor naturlig å anta at ς er konstant for NB-Normal modellen siden denne modellen da passer inn under definisjonen på HGLM. Men det er også rimelig å anta at estimatet av ς vil bli påvirket av de tilfeldige effektene. Derfor vil jeg undersøke NB-Normal modellen både når ς behandles som en konstant og når ς behandles som en parameter på lik linje med σ.

For å beregne et estimat av $\boldsymbol{v}$ finner jeg først H log likelihood, som får følgende form for denne modellen:

$$\begin{aligned} h =& l(\boldsymbol{\mu}, \varsigma; \boldsymbol{y}|\boldsymbol{v}) + l(\sigma; \boldsymbol{v}) \\ =& \sum_{i=1}^{10} \Bigg[\sum_{j=1}^{n_i} (\log(\Gamma(y_{ij} + \varsigma)) - \log(\Gamma(\varsigma)) - \log(y_{ij}!) + y_{ij}(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij}) \\ & + \varsigma \log(\varsigma) - (y_{ij} + \varsigma)\log(e^{\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij}} + \varsigma)) - \frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{v_i^2}{2\sigma^2} \Bigg] \end{aligned}$$

Ved å derivere denne funksjonen mhp. v_i får man følgende ligninger:

$$\begin{aligned} \frac{\partial h}{\partial v_i} &= \sum_{j=1}^{n_i} \left[y_{ij} - \frac{(y_{ij} + \varsigma)\exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij})}{\exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij}) + \varsigma} \right] - \frac{v_i}{\sigma^2}\ , \\ -\frac{\partial^2 h}{\partial v_i^2} &= \frac{1}{\sigma^2} + \sum_{j=1}^{n_i} \frac{(y_{ij} + \varsigma)\varsigma \exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij})}{(\exp(\boldsymbol{X}'_{ij}\boldsymbol{\beta} + v_i + \eta_{ij}) + \varsigma)^2}\ . \end{aligned} \tag{5.13}$$

Her er $\frac{\partial^2 h}{\partial v_i^2} < 0 \; \forall \; v_i$ siden $\varsigma > 0$. Dermed vil $\hat{v}_i$ fra $\frac{\partial h}{\partial v_i}|_{v_i=\hat{v}_i} = 0$ gi maksimum for h.
For denne modellen kan ikke den marginale log likelihood funksjonen l til $\boldsymbol{y}$ beregnes analytisk. Jeg bruker derfor $p_{\boldsymbol{v}}(h)$ til å finne et estimat av $\boldsymbol{\beta}$.

$$
\begin{aligned}
p_{\boldsymbol{v}}(h) = \sum_{i=1}^{10} \Bigg[\sum_{j=1}^{n_i} & (\log(\Gamma(y_{ij}+\varsigma)) - \log(\Gamma(\varsigma)) - \log(y_{ij}!) + y_{ij}(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + \hat{v}_i + \eta_{ij}) \\
& + \varsigma \log(\varsigma) - (y_{ij}+\varsigma)\log\left(e^{\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\hat{v}_i+\eta_{ij}} + \varsigma\right)) - \log(\sigma) - \frac{\hat{v}_i^2}{2\sigma^2} \\
& - \frac{1}{2}\log\Big(- \frac{\partial^2 h}{\partial v_i^2}\Big|_{v_i=\hat{v}_i}\Big)\Bigg].
\end{aligned}
$$

Jeg bruker også denne funksjonen til å beregne AIC og BIC verdier for ulike varianter av NB-Normal modellen.
For å finne et estimat av σ, og ς når det er aktuelt, bruker jeg $p_{(\boldsymbol{v},\boldsymbol{\beta})}(h)$. Jeg trenger da følgende ligninger i tillegg til formel 5.13 på forrige side

$$
\begin{aligned}
-\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} &= \sum_{i=1}^{10}\sum_{j=1}^{n_i} \boldsymbol{X}_{ij}\boldsymbol{X}_{ij}' \frac{(y_{ij}+\varsigma)\varsigma \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + v_i + \eta_{ij})}{(\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + v_i + \eta_{ij}) + \varsigma)^2}, \\
-\frac{\partial^2 h}{\partial v_i \partial \boldsymbol{\beta}} &= -\frac{\partial^2 h}{\partial \boldsymbol{\beta}\partial v_i} = \sum_{j=1}^{n_i} \boldsymbol{X}_{ij} \frac{(y_{ij}+\varsigma)\varsigma \exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + v_i + \eta_{ij})}{(\exp(\boldsymbol{X}_{ij}'\boldsymbol{\beta} + v_i + \eta_{ij}) + \varsigma)^2}.
\end{aligned}
$$

Ved hjelp av disse ligningene innsatt i formel 5.12 på side 61 kan jeg nå uttrykke $D(h,(\boldsymbol{v},\boldsymbol{\beta}))$. Innsatt estimater får jeg da følgende funksjon som kan estimere σ

$$
\begin{aligned}
p_{(\boldsymbol{v},\boldsymbol{\beta})}(h) = c + \sum_{i=1}^{10} \Bigg[\sum_{j=1}^{n_i} & (\log(\Gamma(y_{ij}+\varsigma)) - \log(\Gamma(\varsigma)) + y_{ij}(\boldsymbol{X}_{ij}'\hat{\boldsymbol{\beta}} + \hat{v}_i) + \varsigma\log(\varsigma) \\
& - (y_{ij}+\varsigma)\log\left(e^{\boldsymbol{X}_{ij}'\hat{\boldsymbol{\beta}}+\hat{v}_i+\eta_{ij}} + \varsigma\right)) - \log(\sigma) - \frac{\hat{v}_i^2}{2\sigma^2}\Bigg] \\
& - \frac{1}{2}\log\left(\left|\frac{D(h,(\boldsymbol{v},\boldsymbol{\beta}))}{(2\pi)}\right|\right)\Bigg|_{(\boldsymbol{v},\boldsymbol{\beta})=(\hat{\boldsymbol{v}},\hat{\boldsymbol{\beta}})}.
\end{aligned}
$$

For å utføre beregningene bruker jeg den samme algoritmen som for Poisson-Normal modellen i forrige delkapittel. Den eneste forskjellen er at jeg estimerer σ og ς samtidig vha. $p_{(\boldsymbol{v},\boldsymbol{\beta})}(h)$ i punkt 4 når ς skal behandles som en parameter.

Som startverdier har jeg brukt estimerte verdier fra en vanlig NB modell for $\boldsymbol{\beta}$ og ς, verdien 0.1 for σ og nullvektoren $\mathbf{0}$ for $\boldsymbol{v}$.

Estimerte verdier

For tankskip holdt det med 20 iterasjoner for å oppnå ønsket konvergens, dvs. at største endring mellom estimatene fra 2 iterasjoner skal være 10^{-6}. Her NB-N forkortelse for Negativ Binomisk-Normal modell, og NB-N1 er forkortelse for NB-N modell hvor ς behandles som en parameter.

	NB-N	NB-N*	NB-N1	NB-N1*
Verdier for $\boldsymbol{\beta}$:				
Intercept	-4.78(0.98)	-4.76(0.99)	-4.78(0.98)	-4.77(0.99)
log(Age+2)	0.29(0.06)	0.35(0.06)	0.29(0.06)	0.35(0.06)
log(GT)	-0.14(0.04)	-0.14(0.04)	-0.14(0.04)	-0.14(0.04)
log(Sum.Insured)	0.19(0.06)	0.20(0.06)	0.19(0.06)	0.20(0.06)
Stroke	0.26(0.10)	0.27(0.10)	0.26(0.10)	0.27(0.10)
$\hat{\sigma}$	0.13(0.03)	0.12(0.03)	0.13(0.03)	0.12(0.03)
$\hat{\varsigma}$			1.09(0.21)	1.26(0.27)
Antall parametre	6	6	7	7
AIC	8 161.7	8 101.8	8 163.6	8 103.8
BIC	8 206.2	8 146.3	8 215.6	8 155.7

Tabell 5.6: Estimerte verdier for tankskip.

Når ς behandles som en konstant vil den hele tiden ha den estimerte verdien fra den vanlige NB modellen, dvs. 1.06 for NB-N og 1.23 for NB-N*.

For å kunne sammenligne med Poisson-log-Gamma og Poisson-Normal modellene har jeg nedenfor en tabell med de estimerte tilfeldige effektene for tankskip. Som for Poisson-Normal har jeg her oppgitt $e^{\hat{v}}$.

	$e^{\hat{v}_1}$	$e^{\hat{v}_2}$	$e^{\hat{v}_3}$	$e^{\hat{v}_4}$	$e^{\hat{v}_5}$	$e^{\hat{v}_6}$	$e^{\hat{v}_7}$	$e^{\hat{v}_8}$	$e^{\hat{v}_9}$	$e^{\hat{v}_{10}}$
NB-N	1.13	0.99	0.92	1.08	1.17	0.91	1.03	0.84	1.07	0.91
NB-N*	1.13	0.99	0.93	1.09	1.16	0.91	1.01	0.85	1.07	0.92
NB-N1	1.13	0.99	0.92	1.08	1.17	0.91	1.03	0.84	1.07	0.91
NB-N1*	1.13	0.99	0.93	1.09	1.16	0.91	1.01	0.85	1.07	0.92

Tabell 5.7: Estimerte verdier av de tilfeldige effektene for tankskip.

Utifra tabellen ser man at de tilfeldige effektene ikke blir spesielt påvirket av hvorvidt ς behandles som en parameter eller ikke. Jeg vil bruke NB-N1 og NB-N1* når jeg skal sammenligne modeller senere i masteroppgaven.

For alle typer skip fikk jeg problemer med RAM-bruk i R på samme måte som for Poisson-Normal modellen. Algoritmen jeg brukte sa stopp ved den 2 iterasjonen for alle typer skip, derfor har jeg ikke oppgitt noen estimater for det datasettet. Datasettet viste seg å være for stort til at beregningene kunne fungere.

5.4.4 Vurdering av resultatene

Jeg vil nå sammenligne resultatene fra Poisson-log-Gamma og NB-Normal modellene med tidligere modeller.
For tankskip har Poisson-Normal modellen litt mindre AIC og BIC verdier enn Poisson-log-Gamma modellen. NB-Normal modellen gir markert bedre AIC og BIC verdier enn de to Poisson modellene med tilfeldige effekter. NB-Normal modellen gir også lavere AIC verdier enn en vanlig NB modell, men BIC verdiene til NB-Normal er litt høyere enn verdiene til NB modellen. Derimot har NB-Normal høyere AIC og BIC verdier enn det ZINB modellen har. Alle sammenligningene her gjelder både når man har Days.covered som motvekt og når man både har Days.covered og $\log(\rho)$ som motvekt. For tankskip er det dermed fremdeles ZINB modellen som gir best tilpasning til datasettet.
For alle typer skip er det bare en ny modell i dette kapittelet, Poisson-log-Gamma, siden de numeriske beregningene ikke gikk bra for NB-Normal. Her får Poisson-log-Gamma høyere AIC og BIC verdier enn det Poisson-Normal får, både når bare Days.covered er motvekt og når både Days.covered og $\log(\rho)$ er motvekter. Dermed gir ZINB fremdeles best tilpasning også for alle typer skip.

6
Analyse av miksede modeller

I dette kapittelet vil jeg undersøke de miksede modellene for skadefrekvens vha. hypotesetesting og generering av datasett. Som jeg tidligere har nevnt kan man ikke bruke AIC ukritisk for miksede modeller, derfor er det interessant å bruke noen andre metoder for å vurdere hvor gode de miksede modellene er.

6.1 Hypotesetest av miksede modeller

Jeg vil nå bruke en hypotesetest for å undersøke om det er meningsfyllt å ha tilfeldige effekter i modeller for skadefrekvens. Det ble forholdsvis liten forbedring ved å innføre tilfeldige effekter i modellene for skadefrekvens tidligere i oppgaven, noe AIC og BIC verdiene viser. Derfor er det verdt å undersøke om de tilfeldige effektene er relevante for datasettet. Referansen for denne hypotesetestingen er Demidenko (2004).
Her vil jeg først utvikle en hypotesetest for variansparametrene i en lineært mikset modell, for så å generalisere denne testen til ikke-lineære modeller.
Jeg lar den lineært miksede modellen være gitt ved formel 3.2 på side 23 hvor kovariansmatrisen til de tilfeldige effektene kan skrives som $\boldsymbol{B} = \sigma^2 \boldsymbol{D}$. For denne modellen tilsvarer q antall tilfeldige effekter i modellen. Den aktuelle

hypotesetesten er på formen

$$H_0 : \ \boldsymbol{D} = \mathbf{0}_q, \tag{6.1}$$

hvor $\mathbf{0}_q$ er en matrise av dimensjon $q \times q$ bestående av verdien 0. Formel 6.1 tilhører ikke standard hypotesetesting siden $\boldsymbol{D} = \mathbf{0}_q$ er ytterpunktet for parameterområdet til $\boldsymbol{D}$.

La p være antall kovariater i $\boldsymbol{X}$, n antall observasjoner og N antall grupper de tilfeldige effektene deles inn i. For hypotesetesten brukes matrisen

$$\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{Z}^\star] = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{Z}_1 & \mathbf{0} & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \boldsymbol{X}_N & \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_N \end{bmatrix},$$

som har dimensjon $n \times (p + Nq)$. Indeksene i $\boldsymbol{W}$ skiller de ulike gruppene fra hverandre, for eksempel betyr $\boldsymbol{X}_1$ kovariatmatrisen for gruppe 1(tegningsår 1995). I $\boldsymbol{W}$ består diagonalen til $\boldsymbol{Z}^\star$ av kovariatmatrisen $\boldsymbol{Z}$ oppdelt gruppevis, alle andre elementer i $\boldsymbol{Z}^\star$ er lik 0.
Ideen bak testen er at når $\boldsymbol{D} = \mathbf{0}_q$ så bør forskjellen mellom kvadratsummen for minste kvadraters estimator(MKE) med tilfeldige effekter, S_{min}, og kvadratsummen for MKE uten tilfeldige effekter, S_{OLS}, være liten. Først beregner man kvadratsummen under forutsetningen om ingen tilfeldige effekter

$$S_{ols} = ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ols}||^2,$$

hvor $||\boldsymbol{y}||$ er den euklidske normen til vektoren $\boldsymbol{y}$. Deretter beregner man kvadratsummen når tilfeldige effekter er tilstede

$$S_{min} = \min_{\boldsymbol{\iota}} ||\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\iota}||^2,$$

hvor $\boldsymbol{\iota} = (\boldsymbol{\beta}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)$. Testen tar utgangspunkt i følgende teorem:

Teorem 6.1.1

La $r = R(\boldsymbol{W}) = \text{rang}(\boldsymbol{W})$. Under formel 6.1 vil forholdet mellom de to kvadratsummene S_{ols} og S_{min} være F-fordelt på formen

$$\frac{(S_{ols} - S_{min})/(r-p)}{S_{min}/(n-r)} \sim F(r-p, n-r). \tag{6.2}$$

Når $\boldsymbol{D} \neq \mathbf{0}_q$ i den lineært miksede modellen bør S_{min} være forholdsvis liten slik at testobservatoren i formel 6.2 på forrige side blir stor. Derfor forkaster man $H_0 : \boldsymbol{D} = \mathbf{0}_q$ hvis venstre side av formel 6.2 på forrige side er stor.
F-testen foregår på følgende måte: La $1 - \alpha$ være et valgt signifikansnivå og la $\Lambda_{1-\alpha}$ være kvantilen til F-fordelingen med $r - p$ og $n - r$ frihetsgrader. Man forkaster H_0 hvis

$$\frac{(S_{ols} - S_{min})/(r - m)}{S_{min}/(n - r)} > \Lambda_{1-\alpha} \, .$$

F-testen ovenfor blir i Demidenko (2004) generalisert til å gjelde for ikke-lineære modeller ved å bytte ut kvadratsummen S med $-2l$, hvor l er log likelihood for den aktuelle modellen. Denne testen gjelder når man har modeller med multivariate tilfeldige effekter, dvs. når $q \geq 2$. Jeg lar l_0 være maksimum til standard log likelihood for en modell uten tilfeldige effekter. Så lar jeg l_{maks} være maksimum til log likelihood for den samme modellen med tilfeldige effekter. Etter transformasjonen får jeg følgende testobservator

$$\frac{(l_0 - l_{maks})/(r - p)}{l_{maks}/(n - r)} \sim F(r - p, n - r).$$

Som før forkastes formel 6.1 på forrige side hvis testobservatoren er større enn $\Lambda_{1-\alpha}$.
I modellene jeg har undersøkt i masteroppgaven er $q = 1$, dvs. at det ikke er multivariate tilfeldige effekter i modellene. Hypotesetesten blir da redusert til $H_0 : \psi^2 = 0$. Endringen i testobservatoren for modeller der $q = 1$ i forhold til $q > 1$ er liten, Demidenko (2004) gir følgende testobservator for en Poisson-Normal modell hvor Intercept er det eneste leddet i $\boldsymbol{Z}$:

$$\frac{(l_0 - l_{maks})/(N - p)}{l_{maks}/(n - N)} \sim F(N - p, n - N). \tag{6.3}$$

Nå blir H_0 forkastet hvis testobservatoren er større enn $\Lambda_{1-\alpha}$ med $N - p$ og $n - N$ frihetsgrader.

6.1.1 Testing av modeller

Nå vil jeg bruke F-testen for $H_0 : \psi^2 = 0$ på de miksede modellene i kapittel 4 og 5. For mitt datasett er $N = 10$, p vil være 5 for tankskip og 6 for alle typer skip, og n vil være 12 318 for tankskip og 53 588 for alle typer skip.

Jeg lar $\alpha = 0.05$. Denne hypotesetesten blir i Demidenko (2004) bare brukt for Poisson-Normal modellen hvor Intercept er eneste element i $\boldsymbol{Z}$. Jeg har også brukt testen på de andre modellene med tilfeldige effekter i kapittel 4 og kapittel 5.

I tabell 6.1 blir Poisson-Normal modellen og Poisson-log-Gamma modellen målt opp mot den vanlige Poisson modellen. Negativt Binomisk modell med normalfordelte tilfeldige effekter(NB-Normal) blir målt mot en vanlig Negativt Binomisk modell, men bare for tankskip. For NB-Normal bruker jeg modellen der ς behandles som en parameter. For modellene i tabell 6.1 har jeg bare brukt Days.covered som motvekt.

Modell	Skipstype	Testobservator	Λ	Forkastning av H_0
Poisson-Normal	tankskip	2.61	2.21	Ja
Poisson-Normal	alle skip	2.94	2.37	Ja
Poisson-log-Gamma	tankskip	1.98	2.21	Nei
Poisson-log-Gamma	alle skip	1.89	2.37	Nei
NB-Normal	tankskip	2.02	2.21	Nei

Tabell 6.1: Oversikt over hypotesetesting av modeller.

Tabellen viser at jeg kan forkaste hypotesen i formel 6.1 på side 69 for Poisson-Normal modellen, men ikke for Poisson-log-Gamma og NB-Normal. Verdiene til testobservatorene i tabellen er enten litt over Λ eller litt under Λ, det er ingen stor forskjell.
For tankskip får Poisson-Normal forkastning, mens NB-Normal ikke får det. Dette kan virke overraskende siden NB-Normal har betydelig mindre AIC og BIC verdier enn Poisson-Normal. Men dette er forståelig siden den vanlige NB modellen har betydelig lavere AIC og BIC verdi enn den vanlige Poisson modellen, og NB-Normal blir målt opp mot den vanlige NB modellen.

Hvis jeg bruker $\log(\rho)$ som motvekt i tillegg til Days.covered så endres testobservatorene. For tankskip minker testobservatorene, men endringen er forholdsvis liten og resultatene mhp. hypotesetesten endrer seg ikke. For alle typer skip blir derimot resultatene litt annerledes, noe tabellen nedenfor viser.

Modell	Skipstype	Testobservator	Λ	Forkastning av H_0
Poisson-Normal	alle skip	5.85	2.37	Ja
Poisson-log-Gamma	alle skip	4.25	2.37	Ja

Tabell 6.2: Hypotesetesting av modeller når både $\log(\rho)$ og Days.covered er motvekter.

Testobservatorene øker altså for alle typer skip. For Poisson-log-Gamma øker testobservatoren så mye at jeg nå får forkastning av hypotesen i formel 6.1 på side 69.
Hypotesetestingen viser at det er hensiktsmessig å bruke normalfordelte tilfeldige effekter for Poisson modellen, både med og uten $\log(\rho)$ som motvekt. De tilfeldige effektene for Poisson-log-Gamma modellen er derimot ikke signifikante ifølge hypotesetestingen, med unntak for alle typer skip med $\log(\rho)$ som motvekt. Hypotesetestingen for NB-Normal modellen tyder på at de normalfordelte tilfeldige effektene ikke er signifikante for NB modellen til tankskip.

6.2 Generering av datasett

Tidligere har jeg brukt AIC og BIC til å sammenligne hvilke modeller som passer til datasettet. Nå vil jeg bruke en alternativ metode til å evaluere modellene. Denne metoden går ut på å generere flere datasett fra en modell for så å sammenligne datasettene med de opprinnelige observasjonene.

De fleste av de miksede modellene har jeg beregnet parametre til på egenhånd, mens parametre for de andre modellene er beregnet vha. funksjoner i R. Derfor genererer jeg hovedsakelig datasett for de miksede modellene, det er disse modellene det er knyttet mest usikkerhet til. Jeg vil dermed undersøke hvor gode de miksede modellene er ved å generere observasjoner fra de miksede modellene mhp. parameterestimatene beregnet i kapittel 5, og så sammenligne de genererte observasjonene med de virkelige observasjonene.

For å utføre datagenereringen til de miksede modellene bruker jeg følgende fremgangsmåte:

1 Ta utgangspunkt i de estimerte parametrene for fordelingen til de tilfeldige effektene og generer data fra fordelingen vha. disse parameterestimatene.

2 Ta utgangspunkt i de estimerte parametrene for fordelingen til responsvariablene. Bruk disse parameterverdiene, og dataene fra punkt 1, og generer data fra fordelingen til responsvariablene.

3 Lagre de genererte dataene fra punkt 2. Utfør 1-2 mange ganger. Sammenlign så de genererte datasettene med de opprinnelige observasjonene for skadefrekvens.

For å kunne sammenligne de genererte datasettene og de opprinnelige observasjonene bruker jeg Pearsons kji-kvadrat test. Denne testen er nærmere forklart i tillegg A. En av fordelene med denne testen er at den er enkel å bruke. En negativ ting med testen er at den er veldig streng. Store datasett fører vanligvis til forkastning. For å kunne bruke denne testen deler jeg observasjonene og genererte datasett for skadefrekvens opp i tabeller på følgende form:

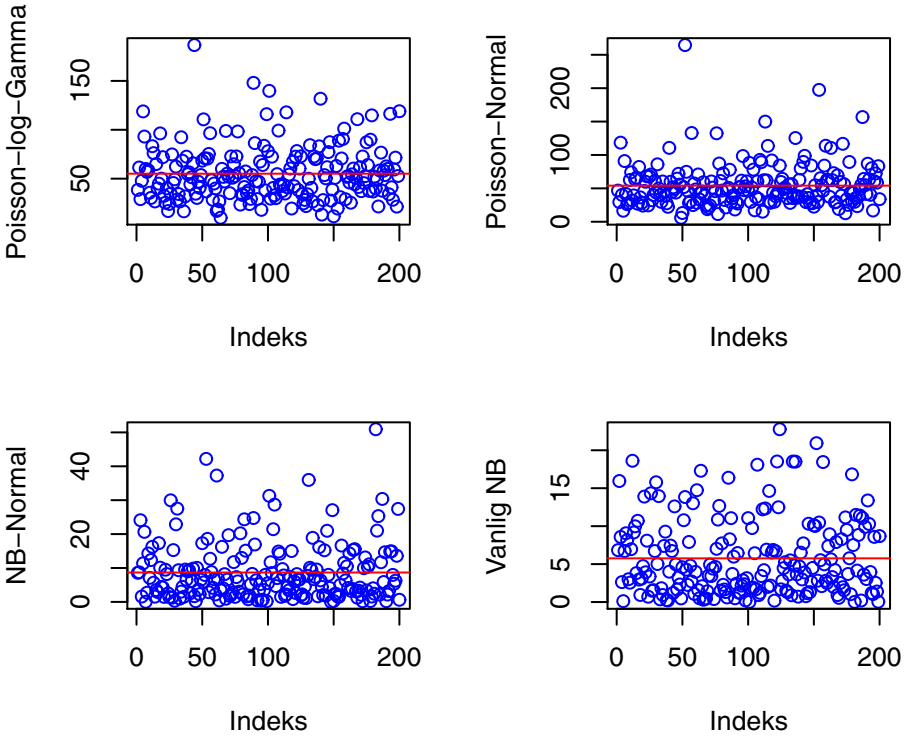
Skadeantall	0	1	2-5
Frekvens for skadeantall	O_0	O_1	O_{2-5}

Tabell 6.3: Formen til tabellene.

For tankskip og alle typer skip slår jeg sammen observasjoner med skadefrekvens større eller lik 2 til en kategori. Dette er fordi en så høy skadefrekvens er forholdsvis sjelden, det tilsvarer mindre enn 1 % av datasettene. Dette forhindrer også problemer med at den estimerte frekvensen for skadeantall av og til ikke har noen skadeantall større enn 2.

For mine modeller er det problematisk å bestemme frihetsgraden, og dermed p-verdier, til kji-kvadrat testen. For tankskip og alle typer skip vil dimensjonen til parametervektoren minst være 5, mens antall observerte frekvenser i tabellen ovenfor er 3. Derfor har jeg valgt å heller undersøke testobservatorene til de ulike modellene. En lav testobservator er et godt tegn på at den aktuelle modellen passer til observasjonene, mens en høy testobservator tyder på det motsatte. For å ha noe å sammenligne testobservatorene med, og dermed bestemme hva som er høye og lave testobservatorer, har jeg også generert datasett for en vanlig NB modell, både for tankskip og alle typer skip.

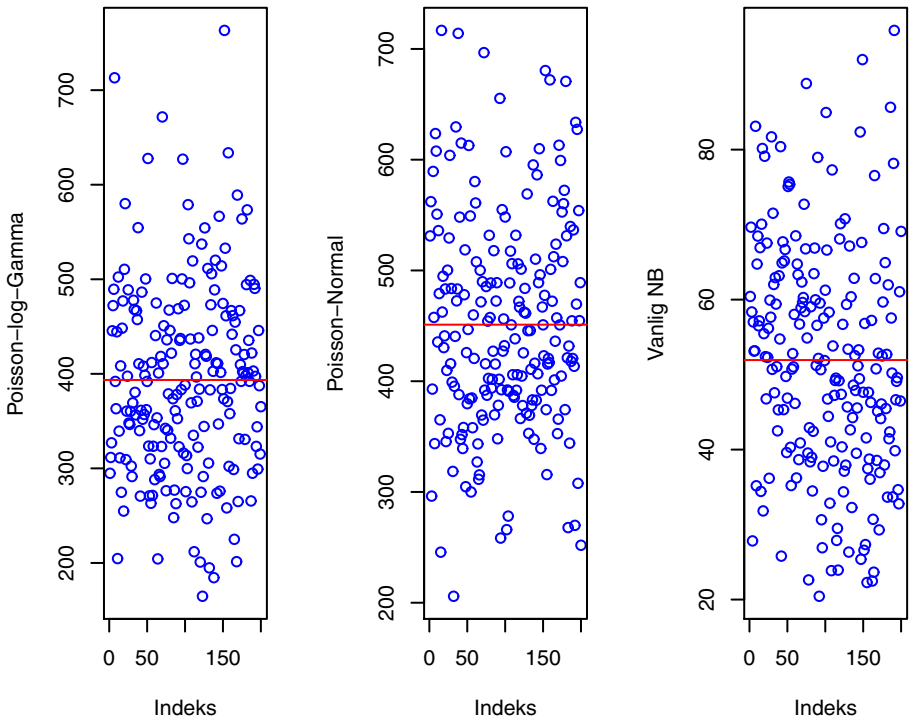
Jeg har generert 200 datasett for hver av de miksede modellene og NB modellen. Her undersøker jeg bare modellene hvor Days.covered er eneste motvekt. For NB-Normal bruker jeg modellen hvor ς behandles som en vanlig parameter.



Figur 6.1: Testobservatorer for modellene til tankskip, der bare Days.covered er motvekt. Den rette linjen gjennom hvert plott viser gjennomsnittet til testobservatorene for hver modell. Gjennomsnittet for testobservatorene til NB-Normal er 8.66 mens det er 5.75 for den vanlige NB modellen.

Figur 6.1 viser at testobservatorene til NB-Normal og NB modellene er betydelig mindre enn testobservatorene til Poisson-Normal og Poisson-log-Gamma modellene. Testobservatorene for den vanlige NB modellen er også litt lavere enn testobservatorene til NB-Normal modellen. NB modellen har også en mindre parameter enn NB-Normal modellen, dermed tyder datagenereringen på at NB modellen passer bedre til observasjonene enn NB-Normal. Datagenereringen tyder også på at NB modellene passer bedre til observasjonene enn Poisson modeller med tilfeldige effekter.

For alle typer skip har jeg bare 2 modeller med tilfeldige effekter, siden estimeringen for NB-Normal modellen ikke gikk bra.



Figur 6.2: Testobservatorer for modellene til alle typer skip, der bare Days.covered er motvekt. Den rette linjen gjennom hvert plott viser gjennomsnittet til testobservatorene for hver modell.

Det er naturlig at testobservatorene fra kji-kvadrat testen i figur 6.2 er en del større her, siden man nå ser på testobservatorer til et mye større datasett enn i figur 6.1 på forrige side. Figur 6.2 viser at den vanlige NB modellen gir betydelig lavere testobservatorer enn Poisson modellene med tilfeldige effekter.

En periode trodde jeg at hvor lave testobservatorene til de miksede modellene er avhenger av hvor like de genererte tilfeldige effektene er de estimerte tilfeldige effektene. Men når jeg sammenlignet testobservatorene, de genererte tilfeldige effektene og de estimerte tilfeldige effektene så fant jeg ikke noen klar sammenheng.

Datagenereringen viser at Poisson modellene med tilfeldige effekter produserer datasett som gir høye testobservatorer, både for tankskip og alle typer skip. For å undersøke om dette gjelder generelt for Poisson modeller genererte jeg også 200 datasett for den vanlige Poisson modellen, både for tankskip og alle typer skip. Testobservatorene for den vanlige Poisson modellen ble mar-

kert større enn testobservatorene til Poisson-log-Gamma og Poisson-Normal modellene.

Datagenereringen tyder på at de tilfeldige effektene i Poisson-Normal og Poisson-log-Gamma bidrar til at modellene får lavere testobservatorer sammenliget med en vanlig Poisson modell. Men sammenlignet med den vanlige NB modellen er ikke forbedringen spesielt stor. Både datagenereringen og hypotesetestingen for tankskip tyder på at den vanlige NB modellen er såpass godt tilpasset til de opprinnelige observasjonene at det ikke er nødvendige å innføre tilfeldige effekter for denne modellen. Disse resultatene stemmer overens med de tidligere modellsammenligningene i masteroppgaven basert på AIC og BIC verdier.

7
Sammendrag

Denne masteroppgaven har tatt for seg sammenligning av modeller for skadefrekvens i skipsforsikringsdata. For alle modellene jeg har prøvd har innføringen av motvekten $\log(\rho)$, som ble laget i avsnitt 3.3, ført til en markert reduksjon i AIC og BIC. Alle modellsammenligningene viser at modeller basert på den Negativt Binomiske fordelingen gir bedre tilpasning enn modeller basert på Poisson fordelingen. De modellene som har fått lavest AIC og BIC verdier er nullforhøyde modeller, spesielt ZINB modeller.

Innføringen av tilfeldige effekter for Poisson og NB modellene har ført til litt lavere AIC og BIC verdier for disse modellene. Men kapittel 6 viser at dette ikke nødvendigvis er en signifikant forbedring. Og selv om forbedringen for noen av modellene er signifikant så får de miksede modellene høyere AIC og BIC verdier enn de nullforhøyde modellene. For mitt datasett mener jeg det er unødvendig å bruke miksede modeller for å modellere skadefrekvensen.

H-likelihood har vist seg å være en metode det er vanskelig å forstå. Beregningene er ikke spesielt vanskelige, men teorien beregningene bygger på er ofte uklar og uoversiktelig. Hvis jeg senere får bruk for å beregne parametre til miksede modeller så vil jeg heller ta utgangspunkt i en annen metode.

Forslag til videre arbeid

Det hadde vært interessant å prøve flere metoder som kan beregne parametre til miksede modeller. McCulloch og Searle (2001) foreslår blant annet EM algoritmen og MCMC metoder. Disse metodene er i utgangspunktet mer kompliserte enn H-likelihood, men de er ofte brukt i artikler om miksede modeller.

Å bruke tegningsår som tilfeldig effekt virker som en god ide. For eksempel kan man se på hvordan miksede modeller fungerer på datasett fra bilforsikring. For slike datasett kan man forvente mer overdispersjon enn man har i datasettet jeg tok utgangspunkt i for skipsforsikring.

En interessant modell jeg ikke fikk tid til å undersøke er en nullforhøyd modell med tilfeldige effekter, der det nullforhøyde leddet avhenger av kovariater. En slik modell er omhandlet i Hall (2000), men ikke for skadeforsikringsdata. Men for å beregne en slik modell for store datasett bør man nok bruke andre programmeringsspråk enn R, siden R ikke fungerer bra for store datasett.

A

Egenskaper brukt i utledninger og mellomregninger

Brukt i beregning av standardfeil til ZINB modell:

Teorem A.0.1: Delta-metoden

La $\hat{\nu}$ være et estimat av ν basert på et utvalg av størrelse n. Anta at

$$\sqrt{n}(\hat{\nu} - \nu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

La $h(\cdot)$ være en funksjon som er differensierbar omkring ν og $h'(\nu) \neq 0$. Da vil

$$\sqrt{n}(h(\hat{\nu}) - h(\nu)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(h'(\nu))^2).$$

Brukt i forklaringen av PQL:

Definisjon A.0.2

Hvis $\boldsymbol{A}$ og $\boldsymbol{B}$ begge er $n \times n$ matriser så er

1 $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$,

2 $|A'| = |A|$,

3 A er invertibel hvis og bare hvis $|A| \neq 0$ og i så fall er $|A^{-1}| = |A|^{-1}$.

Definisjon A.0.3: Laplace approksimasjon for multivariate tettheter

La $D \subseteq \mathbb{R}^m$.
Anta at $g(.)$ er en glatt funksjon som avbilder D til $\mathbb{R}$ og at $g(.)$ har et unikt minimum i $\boldsymbol{x_0}$ som ligger i området til D. Da har man følgende approksimasjon

$$\int_D e^{-g(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \simeq \frac{(2\pi)^{m/2}}{|g''(\boldsymbol{x_0})|^{1/2}} e^{-g(\boldsymbol{x_0})}. \tag{A.1}$$

Her er $g''(\boldsymbol{x_0})$ en hessisk matrise med dimensjon $m \times m$ innsatt $\boldsymbol{x_0}$.

Brukt i kapittel 6:

Definisjon A.0.4: Rang til matrise

La $\boldsymbol{A}$ være en matrise med dimensjon $m \times n$. Rangen til $\boldsymbol{A}$, $R(\boldsymbol{A})$, defineres som det høyeste antall lineært uavhengige rekker til $\boldsymbol{A}$.
Antall lineært uavhengige rekker til $\boldsymbol{A}$ er lik antall lineært uavhengige søyler. Da en rekke i $\boldsymbol{A}$ er en søyle i $\boldsymbol{A}'$ betyr det at $R(\boldsymbol{A}) = R(\boldsymbol{A}')$.

Definisjon A.0.5: Pearson kji-kvadrat test

La O være den observerte frekvensen til et utfall i et utvalg. La så E være den korresponderende forventede frekvensen til en aktuell modell. Testobservatoren X^2 er definert ved

$$X^2 = \sum_{i=1}^{m} r_i^2.$$

Her er r_i Pearson residualen som er definert ved

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}},\ i = 1, \ldots, m\ ,$$

hvor m er antall observerte frekvenser. La p være antallet parametre som har blitt estimert til den aktuelle modellen vha. de observerte frekvensene. Hvis modellen er "rett"vil X^2 approksimalt være kji-kvadrat fordelt med $m - p - 1$ frihetsgrader.

Når det er snakk om kontinuerlige tilfeldige variabler vil frekvensene bestå av verdiområder. For andre tilfeldige variabler er det vanlig å kombinere omkringliggende sjeldne verdier til en enkelt kategori. Dette er fordi kji-kvadrat approksimasjonen ikke fungerer hvis det er for mange små forventede frekvenser.

A.1 Teoretisk begrunnelse for beregning av motvekt

La uttrykkene være definert slik avsnitt 3.3 oppgir. La så

$$\begin{aligned} \rho &= \Pr(\tilde{Y} > d; \boldsymbol{\theta}) = 1 - F(d; \boldsymbol{\theta}), \\ N|\tilde{N} = h &\sim \text{Binomisk}(\rho, h), \\ \tilde{N}_i &\sim \text{Poisson}(\mu). \end{aligned}$$

Utifra dette kan man beregne fordelingen til N. Siden N er registrert skadefrekvens og $\tilde{N}$ er virkelig skadefrekvens så vil alltid $\tilde{N} \geq N$.

$$\begin{aligned} \Pr(N = r) &= \sum_{h=r}^{\infty} \Pr(N = r, \tilde{N} = h) \\ &= \sum_{h=r}^{\infty} \Pr(N = r|\tilde{N} = h) \Pr(\tilde{N} = h) \\ &= \sum_{h=r}^{\infty} \binom{h}{r} \rho^r (1-\rho)^{h-r} \frac{\mu^h}{h!} e^{-\mu} \\ &= \sum_{h=r}^{\infty} \frac{h!}{r!(h-r)!} \rho^r \mu^r ((1-\rho)\mu)^{h-r} \frac{e^{-\mu}}{h!} \\ &= \frac{(\rho\mu)^r}{r!} e^{-\mu} \sum_{k=0}^{\infty} \frac{((1-\rho)\mu)^k}{k!} \\ &= \frac{(\rho\mu)^r}{r!} e^{-\mu} e^{(1-\rho)\mu} \\ &= \frac{(\rho\mu)^r}{r!} e^{-\rho\mu}. \end{aligned}$$

Utifra dette kan ser man at $N_i \sim \text{Poisson}(\mu_i \rho_i),\ i = 1, \ldots, n$.

Litteratur

Akaike H. (1974). «A new look at the statistical model identification». *IEEE Trans. Automatic Control*, **volum AC-19**, side 716–723. ISSN 0018-9286. System identification and time-series analysis. Referert til på side 14.

Breslow N.E. og Clayton D.G. (1993). «Approximate inference in generalized linear mixed models». *Journal of the American Statistical Association*, **volum 88**, nummer 421, side 9–25. ISSN 01621459. URL: `http://www.jstor.org/stable/2290687`. Referert til på side 41.

Cameron A. og Trivedi P. (1998). *Regression analysis of count data*. Cambridge University Press. Referert til på side 10 og 11.

Demidenko E. (2004). *Mixed models*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons]. ISBN 0-471-60161-6. Theory and applications. Referert til på side 16, 68, 70 og 71.

Fox J. og Monette G. (1992). «Generalized collinearity diagnostics». *Journal of the American Statistical Association*, **volum 87**, nummer 417, side 178–183. ISSN 01621459. URL: `http://www.jstor.org/stable/2290467`. Referert til på side 16.

Hall D.B. (2000). «Zero-inflated poisson and binomial regression with random effects: A case study». *Biometrics*, **volum 56**, nummer 4, side 1030–1039. ISSN 0006341X. URL: `http://www.jstor.org/stable/2677034`. Referert til på side 11 og 78.

Lambert D. (1992). «Zero-inflated poisson regression, with an application to defects in manufacturing». *Technometrics*, **volum 34**, nummer 1, side 1–14. ISSN 00401706. URL: `http://www.jstor.org/stable/1269547`. Referert til på side 11.

Lawless J.F. (1987). «Negative binomial and mixed Poisson regression». *Canad. J. Statist.*, **volum 15**, nummer 3, side 209–225. ISSN 0319-5724. Referert til på side 9.

Lee Y. og Nelder J.A. (2009). «Likelihood inference for models with unobservables: another view». Referert til på side 52.

Lee Y., Nelder J.A. og Pawitan Y. (2006). *Generalized linear models with random effects*, volum 106 av *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL. ISBN 978-1-58488-631-0; 1-58488-631-5. Unified analysis via *H*-likelihood, With 1 CD-ROM (Windows). Referert til på side 49, 50, 51, 54, 55, 60 og 64.

Marquardt D.W. (1970). «Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation». *Technometrics*, **volum 12**, nummer 3, side 591–612. ISSN 00401706.
URL: `http://www.jstor.org/stable/1267205`. Referert til på side 17.

McCullagh P. og Nelder J. (1989). *Generalized linear models*. Chapman & Hall/CRC. Referert til på side 6.

McCullagh P. og Nelder J.A. (1983). *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall. ISBN 0-412-23850-0. Referert til på side 4.

McCulloch C.E. (2003). *Generalized linear mixed models*. NSF-CBMS Regional Conference Series in Probability and Statistics, 7. Institute of Mathematical Statistics. ISBN 0-940600-54-4. Referert til på side 42.

McCulloch C.E. og Searle S.R. (2001). *Generalized, linear, and mixed models*. Wiley Series in Probability and Statistics: Texts, References, and Pocketbooks Section. Wiley-Interscience [John Wiley & Sons]. ISBN 0-471-19364-X. Referert til på side 21, 35, 37, 41, 42, 44 og 78.

Nelder J. og Wedderburn R. (1972). «Generalized linear models». *Journal of the Royal Statistical Society. Series A (General)*, side 370–384. Referert til på side 4.

Noh M., Lee Y. og Pawitan Y. (2005). «Robust ascertainment-adjusted parameter estimation». *Genetic Epidemiology*, **volum 29**, nummer 1, side 68–75. ISSN 0741-0395. DOI: 10.1002/gepi.20078. Referert til på side 62.

Paulsen J., Lunde A. og Skaug H.J. (2008). «Fitting mixed-effects models when data are left truncated». *Insurance: Mathematics and Economics*, **volum 43**, nummer 1, side 121–133.
URL: `http://ideas.repec.org/a/eee/insuma/v43y2008i1p121-133.html`. Referert til på side 31.

Pawitan Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press. Referert til på side 12 og 15.

Venables W. og Ripley B. (2002). *Modern applied statistics with S*. Springer. Referert til på side 9.

Wedderburn R.W.M. (1974). «Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method». *Biometrika*, **volum 61**, side 439–447. ISSN 0006-3444. Referert til på side 6.

Zeileis A., Kleiber C. og Jackman S. (2008). «Regression models for count data in R». *Journal of Statistical Software*, **volum 27**, nummer 8, side 1–25.
URL: `http://www.jstatsoft.org/v27/i08/`. Referert til på side 8, 10, 11 og 18.