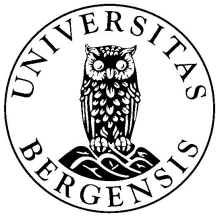
Skadebeløpsmodellering i skadeforsikring

Hege Kristine Kristvik Holmedal

Masteroppgave i statistikk
Finansteori og forsikringsmatematikk



Matematisk institutt
Universitetet i Bergen
1. juni 2009

Takk

Jeg ønsker å takke min veileder Jostein Paulsen for gode råd og tanker på veien.
Jeg vil takke min samboer Thomas Lone Johannessen for å ha holdt ut med meg i denne perioden, og for å ha vært hjelpsom og forståelsesfull.
Mine medstudenter på Kroepelien vil jeg takke for lange lunsjpauser med mange og varierte diskusjoner.
Ellers ønsker jeg å takke familien min for støtte gjennom hele studietiden.

Bergen, 1. juni 2009
Hege Holmedal

Innhold

1 Introduksjon **1**

2 Beskrivelse av problem og data **3**
2.1 Valg av modell 4

3 Sannsynlighetsmaksimeringsestimator **6**
3.1 Generell teori 6
3.2 Lognormal fordeling 9
3.3 Paretofordelingen 11
3.4 Sammenslått lognormal-Pareto fordeling 14
3.4.1 Fordelingen 14
3.4.2 Estimering 22
3.4.3 Tilpasse sammenslått lognormal-Pareto fordeling . . . 24
3.5 Sammenligning av de tre modellene 25
3.6 Egenandel 27
3.6.1 Lognormal fordeling 27
3.6.2 Paretofordelingen 30
3.6.3 Sammenslått lognormal-Pareto fordeling 31

4 Regresjonsmodeller **37**
4.1 AIC-kriteriet 37
4.2 Lineær regresjon 38
4.3 Generaliserte lineære modeller 38
4.3.1 Den eksponensielle familien 39
4.3.2 Modellen 41
4.3.3 Scorefunksjon 41

4.3.4 Informasjonsmatrisen 42
4.3.5 Newton-Raphson iterasjonsprosedyre 43
4.3.6 Iterative Weighted Least Squares (IWLS) 44
4.4 Estimering 49
4.4.1 Lognormalfordelingen 49
4.4.2 Paretofordelingen 51
4.5 Sammenligning av modellene 53
4.6 Sammenslått lognormal-Pareto fordeling 55
4.7 Egenandel 58
4.7.1 Lognormal fordeling 59
4.7.2 Paretofordeling 63
4.7.3 Sammenslått lognormal-Pareto fordeling 64

5 Oppsummering **67**

Tabeller

3.1 Resultater for estimater for alle skip i lognormal fordeling . . . 10
3.2 Resultater for estimater for tankskip i lognormal fordeling . . 10
3.3 Forventningsverdier og varianser i lognormal fordeling 11
3.4 Resultater for estimater for Paretofordelingen 13
3.5 Resultater for sammenslått lognormal-Paretofordeling 24
3.6 Resultater for estimater i lognormal fordeling for alle skip med egenandeler 28
3.7 Resultater for estimater i lognormal fordeling for tankskip med egenandeler 28
3.8 Gjennomsnittlig forventningsverdi og varians i lognormalfordelingen med egenandel 30
3.9 Resultater for estimater for Paretofordelingen med egenandel . 30
3.10 Resultater for sammenslått lognormal-Paretofordeling med egenandeler 36

4.1 Regresjonsmodell for lognormal fordeling 51
4.2 Forventningsverdi og varians for lognormalfordelingen 51
4.3 GLM for Paretofordeling 53
4.4 Regresjonsmodell for lognormal fordeling med venstretrunkerte data 62
4.5 Gjennomsnittlig forventningsverdi og varians i lognormalfordelingen med egenandel 63
4.6 GLM for Paretofordeling med egenandel 64

Figurer

2.1 Tetthet og histogram for skadedata 5

3.1 QQ-plott for lognormalfordelingen 11
3.2 QQ-plott for Paretofordelingen 13
3.3 QQ-plott for sammenslått lognormal-Paretofordeling 25
3.4 Tettheter for skadegrad på alle skip. 26
3.5 Tettheter for skadegrad på tankskip. 26

4.1 Tetthet til logaritmen av skadedata 50
4.2 QQ-plott for residualene i lognormalfordelingen 54
4.3 QQ-plott for residualene i Paretofordelingen 55

1 Introduksjon

I artikkelen Cooray & Ananda [2005], introduserer artikkelforfatterne en ny fordelingsfunksjon sammensatt at lognormalfordelingen og Paretofordelingen. Hensikten med dette er å finne en fordeling som gir bedre tilpasning til skadedata enn det lognormalfordelingen og Paretofordelingen gjør hver for seg. Ofte vil det være slik at lognormalfordelingen passer best til små, hyppige skader, mens Paretofordelingen egner seg best til å modellere de store og sjeldnere skadene. Cooray & Ananda [2005] ser derfor denne sammenslåtte fordelingen som en mulighet til å tilpasse en god modell til hele spekteret av skadeforekomster. Jeg ønsker å tilpasse denne sammenslåtte lognormal-Paretofordelingen til et datasett med skadegrader på skip. Deretter vil jeg se hvor godt denne modellen passer til mine data sammenlignet med lognormalfordelingen og Paretofordelingen hver for seg. Jeg begynner i kapittel 2 med å introdusere datasettet og problemstillingen. I kapittel 3 tilpasser jeg først en lognormalfordeling og en Paretofordeling til skadegradene med sannsynlighetsmaksimeringsestimatorer. Jeg vil videre introdusere den sammenslåtte lognormal-Paretofordelingen for så å regne ut fordelingsfunksjoner og momenter i denne fordelingen. Deretter tilpasser jeg også denne fordelingen til de samme skadegradene. I neste del av kapittel 3 vil jeg ta hensyn til at det kun er de skadegradene som overstiger egenandelsgradene som blir innrapportert til forsikringsselskapene. Jeg tilpasser derfor venstretrunkerte modeller til alle de tre sannsynlighetsfordelingene. En grundig innføring i likelihoodfunksjoner finner man i Pawitan [2001]. I kapittel 4 introduserer jeg generaliserte lineære modeller, som er annen metode for å tilpasse fordelinger til data. I GLM benyttes det observerte forklaringsvariabler for å tilpasse en modell til dataene. I tilfellet med skipene kan egenskaper ved skipene benyttes til å si noe om skadens omfang. Jeg tilpasser lognormalfordelingen og Paretofordelingen til skadegradene ved GLM, og ser på hvordan man også

kan tilpasse den sammenslåtte lognormal-Paretofordelingen. I neste del av kapittel 4 tar jeg igjen hensyn til at de minste skadene ikke er inneholdt i datasettet mitt. Jeg tilpasser venstretrunkerte modeller til dataene med lognormalfordeling og Paretofordeling, og ser på hvordan man kan tilpasse en slik modell til den sammenslåtte lognormal-Paretofordelingen. Teorien om GLM har jeg funnet i Dobson [2002], Pawitan [2001], Lindsey [1997], samt i forelesningsrekken av Heuch [2007]. Ved behov for mer generell statistisk teori har Casella & Berger [2002] vært til stor hjelp, og Klugman *et al.* [2004] har mye generell teori knyttet til forsikring. Ved statistiske analyser og utregninger har jeg benyttet programpakken R [2007] versjon 2.6.1.

2

Beskrivelse av problem og data

Skadeforsikringsselskapene plikter i henhold til forsikringsvirksomhetsloven å sette av nok penger til å dekke alle krav som oppfyller forsikringsvilkårene. Det innebærer at selskapene må ta inn nok premie til å dekke fremtidig risiko. Selskapene er derfor avhengige av å ha gode modeller for skadeutbetalinger, slik at de med stor grad av sikkerhet skal kunne forutse fremtidige krav. Samtidig vil disse modellene kunne benyttes til å fastsette forsikringspremiene for kundene. Ofte brukes det modeller som ikke passer til alle situasjonene som kan oppstå, og spesielt ytterpunkter kan feilvurderes. Det er vanlig å bruke lognormalfordelingen selv om denne har en tendens til å underestimere størrelsen på de største kravene. Dersom man benytter seg av Paretofordelingen får man ikke noe bilde av de små kravene, selv om Paretofordelingen skulle passe godt til de store kravene. Ved hjelp av et datasett med skipsdata, vil jeg først se hvordan disse fordelingene passer til observasjonene i datasettet. Deretter vil jeg se på om det finnes modeller som gjengir observasjonene på en mer tilfredsstillende måte. I datasettet er det registrert 6755 skip og ulike egenskaper ved hvert av skipene:

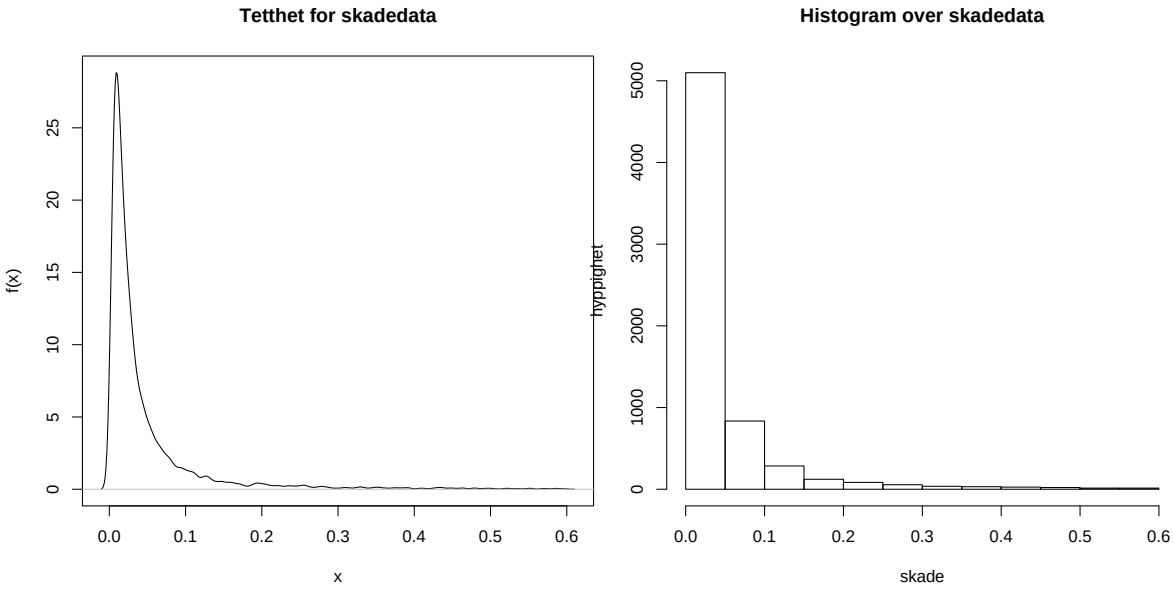
1. Skadegrad = skadens beløp dividert med forsikringsbeløp
2. Egenandelsgrad = egenandel dividert med forsikringsbeløp
3. Skipets alder
4. Bruttotonn
5. Skipets verdi

6. Antall hestekrefter
7. Indikator for maskintype. 1 for 4-takt og 0 for 2-takt
8. Forsikringsår
9. Indikator for skipstype, der det er 9 ulike typer skip
   (a) Bulkskip
   (b) Fraktskip
   (c) Kontainerskip
   (d) Annet
   (e) Passasjerskip
   (f) Tankskip
   (g) LNG/LPG skip (Gasstankskip)
   (h) Cruiseskip
   (i) Forsyningsskip

Selv om datasettet består av 6755 skip, benytter jeg kun de 6622 som har skadegrad mindre enn 0.6. Av disse er 122 ufullstendige, og jeg har derfor valgt å utelate dem fra beregningene. Jeg står dermed igjen med 6500 fullstendige observasjoner. I tillegg til å utføre beregningene for alle skipene, vil jeg også velge ut tankskipene spesielt, og gjøre de samme beregningene for disse. 1421 av alle de registrerte skipene er tankskip.

2.1 Valg av modell

Jeg begynner med å se på hvordan skadene fordeler seg i et histogram, og hvordan tettheten ser ut. Begge plottene finner jeg med kommandoene 'hist' og 'density' i R.



Figur 2.1: Tetthet og histogram for skadedata

Tettheten er høy og smal med lang hale ut til høyre. Den ser ut til å ha fasongen til lognormalfordelingen, en fordeling som ofte blir brukt på denne typen data. Histogrammet begynner med høy hyppighet på de minste skadene, mens frekvensen avtar med skadeomfanget. Utifra histogrammet kan det være naturlig å benytte en tetthet som er monotont synkende. Her er Paretofordelingen aktuell fordi denne er mye benyttet på skadedata. Jeg ønsker derfor å tilpasse begge fordelingene til datasettet.

3

Sannsynlighetsmaksimeringsestimator

3.1 Generell teori

Definisjon 1. *Likelihoodfunksjonen $L(\theta; \mathbf{x})$ er sannsynligheten for de observerte data som en funksjon av en fast men ukjent parameter θ. Dersom dataene er i.i.d skriver vi*

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f_X(x_i; \theta).$$

SME for den ukjente parameteren θ vil være den som maksimerer likelihoodfunksjonen $L(\theta; \mathbf{x})$. I mange tilfeller benyttes også loglikelihoodfunksjonen $l(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x})$ til å finne maksimum, da dette gir samme resultat. For SME gjelder invariansprinsippet:

Teorem 3.1.1. *Dersom $\hat{\theta}$ er SME for θ, er $g(\hat{\theta})$ SME for $g(\theta)$.*

Definisjon 2. *Dersom $(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n)$ er SME for $(\theta_1, \theta_2, ...\theta_n)$, er den observerte Fisherinformasjonen*

$$\mathbf{I}(\hat{\theta}) = \begin{pmatrix} I_{11} & I_{12} & \cdots & I_{1n} \\ I_{21} & I_{22} & \cdots & I_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ I_{n1} & I_{n2} & \cdots & I_{nn} \end{pmatrix},$$

hvor $I_{ij}(\hat{\theta}) = -\frac{\delta^2}{\delta\theta_i \delta\theta_j} \log(L(\theta))$.

Fisherinformasjonen kalles også ofte for informasjonsmatrisen. Det kan vises at asymptotisk vil kovariansmatrisen for $(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n)$ gå mot den inverse

Fisherinformasjonen. Det vil si at dersom

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I^{-1}}(\theta)),$$

kan $\mathbf{I}(\theta)$ erstattes med $\mathbf{I}(\hat{\theta})$, og

$$\mathbf{\Sigma}(\hat{\theta}) = \mathbf{I}^{-1}(\hat{\theta}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix},$$

med $\sigma_{ij} = \text{Cov}(\hat{\theta}_i, \hat{\theta}_j)$.

QQ-plott

Quantile-Quantile plott brukes for å se hvor godt et datasett passer til en bestemt fordeling. QQ-plott er mye brukt for å undersøke om data er normalfordelte, men kan benyttes på alle fordelinger. Dersom dataene kommer fra den valgte fordelingen, bør de ordnede dataene oppføre seg som ordningsobservatorer trukket fra fordelingen. Derfor skal skal dataene plottet mot ordningsobservatorene, gi en tilnærmet rett linje.

Definisjon 3. *Ordningsobservatoren for et tilfeldig utvalg $X_1, X_2, ..., X_n$, er verdiene av utvalget sortert i økende rekkefølge.*
Disse betegnes med $X_{(1)}, X_{(2)}, ..., X_{(n)}$.

For ordningsobservatoren er sannsynlighetstettheten som følger.

Teorem 3.1.2. *La $X_{(1)}, X_{(2)}, ..., X_{(n)}$ være ordningsobservatoren fra et tilfeldig utvalg $X_1, X_2, ..., X_n$ fra en kontinuerlig fordeling med kumulativ fordelingsfunksjon $F_X(x)$ og tetthetsfunksjon $f_X(x)$. Da er tetthetsfunksjonen for $X_{(j)}$*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1!)(n-j)!} f_X(x)[F_X(x)]^{j-1}[1 - F_X(x)]^{n-j}.$$

Det kan også vises at

$$\begin{aligned} &f_{X_{(1)},X_{(2)},...,X_{(n)}}(x_1, x_2, ..., x_n) = \\ &n! f_X(x_1) f_X(x_2) \times ... \times f_X(x_n), \qquad -\infty < x_1 < x_2 < ... < x_n < \infty \end{aligned} \tag{3.1}$$

Det blir også behov for kjennskap til uniform fordeling og betafordeling.

Definisjon 4. *Uniform fordeling definert på intervallet $a \leq y \leq b$, har tetthetsfunksjonen*

$$f(y) = \frac{1}{b-a}.$$

Forventningsverdi og varians for den uniforme fordelingen er

$$E(Y) = \frac{b+a}{2}, \quad Var(Y) = \frac{(b-a)^2}{12}.$$

Vi skriver $Y \sim U[a,b]$.

Definisjon 5. *Dersom Y er betafordelt, har Y sannsynlighetstetthet*

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}.$$

Forventningsverdi og varians i betafordelingen er

$$E(Y) = \frac{\alpha}{\alpha+\beta}, \quad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Vi skriver $Y \sim \mathcal{B}(\alpha,\beta)$.

Anta nå at observasjonene $x_1, x_2, ...x_n$ er uavhengige og identisk fordelte med kumulativ fordeling $F_{\mathbf{X}}(x)$. Jeg velger nå å definere Y som

$$y_i = F_X(x_i). \tag{3.2}$$

Utifra dette kan jeg også finne kumulativ fordelingsfunksjon for Y.

$$\begin{aligned} F_Y(y) = P(F_X(x) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y, \\ F_Y(y) = y = \int_0^y f_Y(y)dy = \int_0^y dy. \end{aligned} \tag{3.3}$$

Med ligning 3.3 og definisjon 4, kan man se at $Y \sim U[0,1]$. Siden $F_{\mathbf{X}}(x)$ er monotont økende, følger det av ligning 3.2 at

$$y_{(j)} = F_X(x_{(j)}), \quad \forall \quad j = 1, 2, ..., n. \tag{3.4}$$

Nå ønsker jeg å finne forventningsverdien for ordningsobservatoren i den uniforme fordelingen. Jeg må derfor først finne tetthetsfunksjonen ved å benytte teorem 3.1.2.

$$\begin{aligned} f_{Y_{(j)}}(y) &= \frac{n!}{(j-1)!(n-j)!} y^{j-1}(1-y)^{n-j} \\ &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} y^{j-1}(1-y)^{(n-j+1)-1}. \end{aligned} \tag{3.5}$$

Fra ligning 3.5 ser man at $Y_{(j)} \sim \mathcal{B}(j, n-j+1)$ i henhold til definisjon 5. Forventningsverdien for $Y_{(j)}$ følger derfor også av definisjonen og blir

$$E(Y_{(j)}) = \frac{j}{n+1}.$$

Derfor kan man anta at

$$y_{(j)} \approx \frac{j}{n+1}. \tag{3.6}$$

Når jeg setter ligning 3.6 inn i ligning 3.2 får jeg at

$$F_X(x_{(j)}) \approx \frac{j}{n+1}$$

Når jeg løser denne ligningen med hensyn på $x_{(j)}$, får jeg det som skal plottes i QQ-plottet; de teoretiske kvantilene mot de ordnede observasjonene i den valgte fordelingen.

$$x_{(j)} = F_X^{-1}(\frac{j}{n+1}) \tag{3.7}$$

Denne ligningen vil stemme dersom X har den fordelingen som antas, og plottet vil da være en tilnærmet rett linje.

3.2 Lognormal fordeling

Jeg begynner med å tilpasse en lognormalfordeling til skadene med SME. En nyttig egenskap ved lognormalfordelingen er at man ved en enkel transformasjon kan få normalfordeling.

Definisjon 6. *Dersom $Z \sim N(\mu, \sigma^2)$ og $Y = e^Z$, sies Y å være lognormalfordelt. Tettheten er*

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{1}{2\sigma^2}(\log y - \mu)^2}.$$

Med kjennskap til denne transformasjonen, kan man benytte seg av de kjente egenskapene ved normalfordelingen. Forventningsverdi og varians i lognormalfordelingen er

$$\begin{aligned} E(Y) &= e^{\mu + \frac{\sigma^2}{2}}, \\ Var(Y) &= e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}. \end{aligned} \tag{3.8}$$

Fordi SME for normalfordelingen er kjent, kan jeg ved invariansprinsippet i teorem 3.1.1 finne SME for lognormalfordelingen uten å sette opp likelihood-funksjonen. SME for normalfordelingen er

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2.$$

Når jeg velger transformasjonen $Y = e^Z$ fra definisjon 6 og benytter teorem 3.1.1 får jeg at SME for lognormalfordelingen blir

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log Y_i,$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\log Y_i - \mu)^2.$$

Tabell 3.1 og 3.2 viser SME i lognormalfordelingen for henholdsvis alle skipene og tankskipene.

	Alle skip			
	Estimat	Var	Se	Cov
$\hat{\mu}$	-3.7897	0.00019	0.0137	0
$\hat{\sigma}^2$	1.2158	0.00045	0.0213	0

Tabell 3.1: Resultater for estimater for alle skip i lognormal fordeling

	Tankskip			
	Estimat	Var	Se	Cov
$\hat{\mu}$	-3.8624	0.0007	0.0273	0
$\hat{\sigma}^2$	1.0588	0.0016	0.0397	0

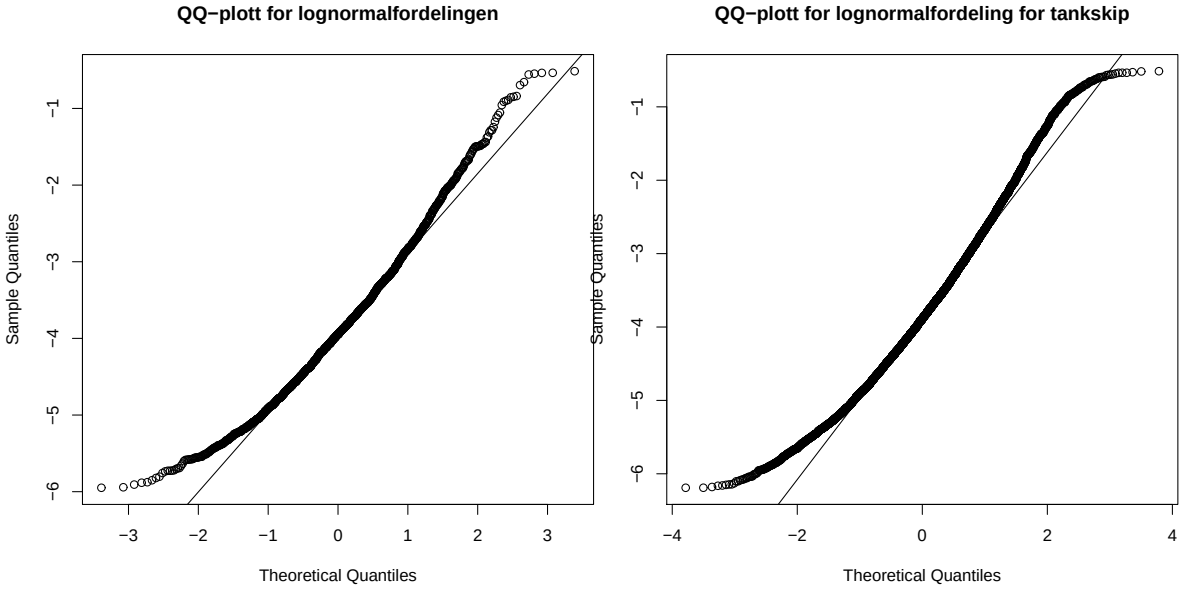
Tabell 3.2: Resultater for estimater for tankskip i lognormal fordeling

Ved å sette resultatene inn i ligning 3.8 finner jeg forventningsverdi og varians for lognormalfordelingen. Disse gjengis i tabell 3.3.

	Alle skip	Tankskip
$E(Y)$	0.0415	0.0357
$\mathrm{Var}(Y)$	0.0057	0.0035

Tabell 3.3: Forventningsverdier og varianser i lognormal fordeling

Nå kan jeg finne QQ-plottet for lognormalfordelingen for henholdsvis alle skipene og tankskipene.



Figur 3.1: QQ-plott for lognormalfordelingen

Fra QQ-plottene i figur 3.1 ser det ikke ut som om lognormalfordelingen gir en god tilpasning til skadegraden. Spesielt i halene avviker skadegraden fra antagelsen om lognormalfordeling. Som tidligere antatt underestimerer lognormalfordelingen de store skadegradene. I tillegg underestimerer den også de små skadegradene.

3.3 Paretofordelingen

Nå ønsker jeg å benytte samme metode for å tilpasse en Paretofordeling til skadene. En viktig egenskap ved Paretofordelingen som vil komme til nytte

senere, er at den greit kan transformeres til en eksponensiell fordeling. Tetthetsfunksjon, forventningsverdi og varians for den eksponensielle fordelingen er

$$f(z) = \alpha e^{-\alpha z},$$
$$E(Z) = \frac{1}{\alpha}, \qquad \mathrm{Var}(Z) = \frac{1}{\alpha^2}.$$

Jeg kan dermed definere en transformasjon slik at jeg får en Paretofordeling.

Definisjon 7. *Dersom $Z \sim exp(\alpha)$ og $Y = \theta e^Z$, sies Y å være Paretofordelt. Tettheten er*

$$f(y) = \frac{\alpha\theta^\alpha}{y^{\alpha+1}}, \quad \textit{for } \theta < y < \infty \textit{ og } \theta, \alpha > 0.$$

Forventningsverdi og varians i Paretofordelingen er

$$E(Y) = \frac{\alpha\theta}{\alpha - 1}, \qquad \text{for } \alpha \geq 1,$$
$$Var(Y) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}, \qquad \text{for } \alpha \geq 2.$$

Likelihoodfunksjonen og loglikelihoodfunksjonen for Paretofordelingen finner jeg ved å benytte definisjon 1 på tettheten for Paretofordelingen.

$$L(\alpha, \theta; \mathbf{y}) = \alpha^n \theta^{n\alpha} \prod_{i=1}^{n} \frac{1}{y_i^{\alpha+1}}, \quad \text{for } y_i \geq \theta.$$

$$l(\alpha, \theta; \mathbf{y}) = n \log \alpha + n\alpha \log \theta - (\alpha + 1) \sum_{i=1}^{n} \log y_i, \quad \text{for } y_i \geq \theta. \tag{3.9}$$

Utifra likelihoodfunksjonen ser jeg at

$$\hat{\theta} = \min_i \{y_i\}$$

er SME for θ. Ved derivasjon med hensyn på α, får jeg at SME i Paretofordelingen blir

$$\hat{\theta} = \min_i \{y_i\},$$
$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log \frac{y_i}{\hat{\theta}}}.$$

Fordi estimatet at θ er en kjent observasjon antar jeg nå at Paretofordelingen kun har en parameter. Da kan jeg finne Fisherinformasjonen og varians for estimatet for α.

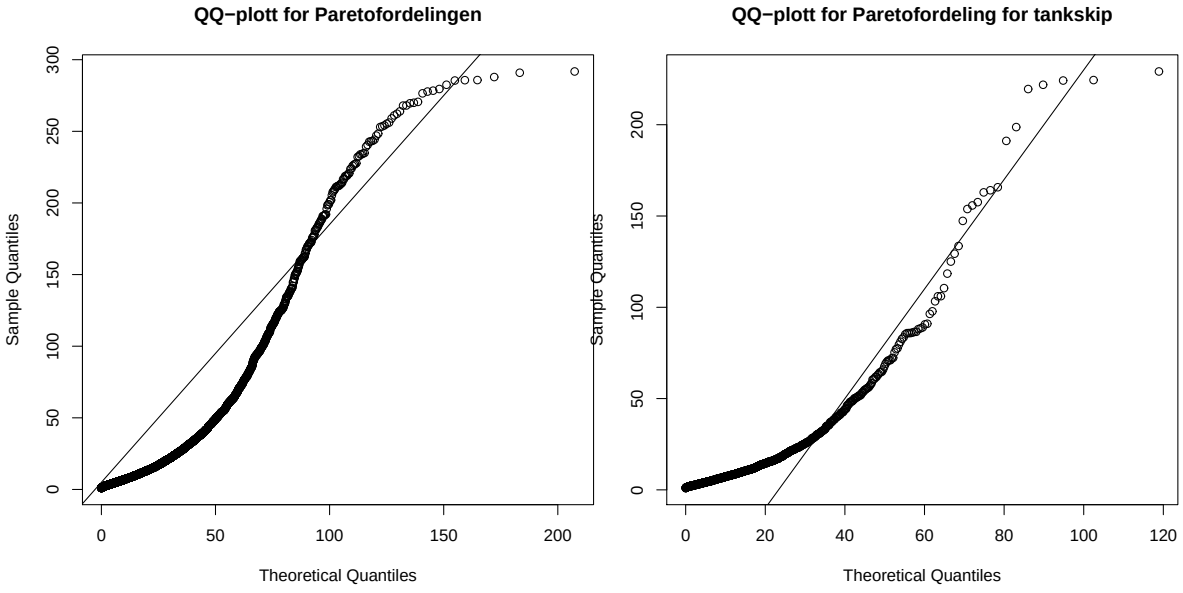
$$I(\hat{\alpha}) = \frac{n}{\hat{\alpha}^2},$$
$$\text{Var}(\hat{\alpha}) = I^{-1}(\hat{\alpha}) = \frac{\hat{\alpha}^2}{n}.$$

Resultatene fra estimeringen er gjengitt i tabell 3.4.

	Alle skip			Tankskip		
	Estimat	Var	se	Estimat	Var	se
$\hat{\alpha}$	0.4166	0	0.0052	0.4793	0.0052	0.0720
$\hat{\theta}$	0.0020	-	-	0.0026	-	-

Tabell 3.4: Resultater for estimater for Paretofordelingen

Jeg tilpasser QQ-plott til begge situasjonene for å undersøke hvordan antagelsen om Paretofordelte skader passer. Ved tilpasningen av QQ-plottene, har jeg benyttet transformasjonen til eksponensiell fordeling fra definisjon 7.



Figur 3.2: QQ-plott for Paretofordelingen

Utifra plottet kan man konkludere med at skadegradene i dette tilfellet ikke er Paretofordelte, hverken for alle skipene eller for tankskipene. Her passer ikke Paretofordelingen til de små skadegradene, men heller ikke til de store.

3.4 Sammenslått lognormal-Pareto fordeling

Nå har jeg tilpasset to ulike fordelinger til datasettet. Selv om ingen av fordelingene passer veldig godt, passer lognormalfordelingen bedre enn Paretofordelingen. Ofte vil lognormalfordelingen passe best for små skadegrader, mens Paretofordelingen passer best for de store. I artikkelen Cooray & Ananda [2005] utvikles derfor en ny fordeling som benytter lognormalfordelingen for de minste dataene og Paretofordelingen for de større dataene. Jeg ønsker nå å se hvordan denne sammenslåtte fordelingen passer til mine skadegrader. Jeg vil derfor først finne fordelingsfunksjoner og momenter i denne fordelingen før jeg tilpasser fordelingen til skadegradene. I artikkelen blir fremgangsmåten for å komme frem til formler og størrelser lite kommentert. Jeg vil her gå igjennom utledningen av uttrykkene med kommentarer.

3.4.1 Fordelingen

Den nye fordelingen skal lages ved å skjøte sammen en lognormalfordeling og en Paretofordeling. Siden estimatoren for θ i Paretofordelingen er

$$\hat{\theta} = \min_i \{y_i\},$$

bør skiftet mellom fordelingene skje i et ukjent punkt y_0 som skal tilsvare parameter θ i Paretofordelingen. Derfor må tettheten for den nye fordelingen være på formen

$$f(y) = \begin{cases} c_1 f_1(y) = c_1 \dfrac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{1}{\sigma^2}(\log y - \mu)^2}, & \text{for } 0 < y \leq y_0, \\ c_2 f_2(y) = c_2 \dfrac{\alpha y_0^\alpha}{y^{\alpha+1}}, & \text{for } y_0 < y \leq \infty. \end{cases} \tag{3.10}$$

I artikkelen Cooray & Ananda [2005], diskuteres ikke muligheten for å velge

$$c_1 \neq c_2$$

i de to tetthetsfunksjonene. Istedet settes modellen uten begrunnelse opp med

$$c = c_1 = c_2.$$

Ved å velge ulike c ville kan kunne oppnå større fleksibilitet i modellen med en ekstra parameter, slik at man kunne fått en enda bedre tilpasning til dataene. Jeg har i denne oppgaven tatt utgangspunkt i artikkelen, og kun sett på alternativet med samme konstant c. For å komme frem til den sammenslåtte lognormal-Paretofordelingen vil jeg gjøre følgende.

1. Innføre krav om glatthet i tettheten
2. Redusere antall parametere til et minimum
3. Finne konstanten c slik at integralet av sannsynlighetstettheten blir 1
4. Sette alle resultatene inn i tetthetsfunksjonen i ligning 3.10
5. Finne kumulativ fordelingsfunksjon
6. Finne forventningsverdi og varians

Glatthet i tettheten

Begge tetthetsfunksjonene som skal skjøtes sammen er kontinuerlige og deriverbare. Derfor er det naturlig at den sammenslåtte lognormal-Paretofordelingen også er kontinuerlig og deriverbar over hele definisjonsområdet, også i punktet y_0.
For å oppnå dette må to ligninger oppfylles.

$$f_1(y_0) = f_2(y_0), \tag{3.11}$$

$$\frac{df_1(y_0)}{dy_0} = \frac{df_2(y_0)}{dy_0}. \tag{3.12}$$

Her er y_0 den observasjonen der skiftet mellom fordelingene foregår. Jeg begynner med å derivere begge tetthetsfunksjonene.

$$\frac{df_1(y)}{dx} = -f_1(y)\frac{1}{y}\left(1 + \frac{1}{\sigma^2}(\log y - \mu)\right),$$

$$\frac{df_2(y)}{dy} = -\frac{\alpha(\alpha+1)y_0^{\alpha}}{y^{\alpha+2}}.$$

Jeg ser da at 3.11 kan settes inn i ligning 3.12 i punktet y_0, slik at den deriverte av $f_1(y_0)$ blir en funksjon av $f_2(y_0)$.

$$\frac{df_1(y_0)}{dx_0} = -f_2(y_0)\frac{1}{y_0}(1 + \frac{1}{\sigma^2}(\log y_0 - \mu)).$$

Nå kan jeg bruke ligning 3.12, og sette de to deriverte lik hverandre i punktet y_0.

$$\begin{aligned} \frac{df_1(y_0)}{dx_0} &= \frac{df_2(y_0)}{dy_0}, \\ f_2(y_0)\frac{1}{y_0}(1 + \frac{1}{\sigma^2}(\log y_0 - \mu)) &= \frac{\alpha(\alpha+1)}{y_0^2}. \end{aligned}$$

Når jeg setter inn $f_2(y_0)$ i ligningen gir det

$$\log y_0 - \mu = \alpha\sigma^2. \tag{3.13}$$

Dette resultatet kan settes inn i ligning 3.11:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\alpha^2\sigma^2} &= \alpha, \\ \alpha\sigma\sqrt{2\pi} &= e^{-\frac{1}{2}\alpha^2\sigma^2}. \end{aligned}$$

Dersom man setter $k = \alpha\sigma$ står man igjen med ligningen

$$k\sqrt{2\pi} = e^{-\frac{1}{2}k^2}. \tag{3.14}$$

Da er det tydelig at k må være en konstant. Cooray & Ananda [2005] har regnet ut at $k = 0.3722$.

Reduksjon av antall parametre

Sannsynlighetsfordelingen har nå fire parametre. Ved å benytte ligning 3.13 og at $k = \alpha\sigma$, kan $f_1(y)$ omformes til en funksjon av α og y_0.

$$f_1(y) = \frac{\alpha}{\sqrt{2\pi}ky}e^{-\frac{\alpha^2}{2k^2}(\log\frac{y}{y_0} - \frac{k^2}{\alpha})^2}$$

Ved å sette inn for k ved hjelp av ligning 3.14 får jeg at

$$f_1(y) = \frac{\alpha y_0^\alpha}{y^{\alpha+1}}e^{-\frac{\alpha^2}{2k^2}(\log\frac{y}{y_0})^2}.$$

Her ser man at μ og σ forsvinner helt ut av fordelingen, slik at den nye fordelingen kun inneholder to parametre.

Integralet av tettheten

For at $f(y)$ skal være en sannsynlighetstetthet, må det eksistere en c slik at

$$\int_0^\infty f(y)dy = c\int_0^{y_0} f_1(y)dy + c\int_{y_0}^\infty f_2(y)dy = 1.$$

Jeg ønsker derfor å løse de to integralene, slik at jeg kan finne en c som oppfyller dette.

$$\begin{aligned}
&c\int_0^{y_0} f_1(y)dy + c\int_{y_0}^\infty f_2(y)dy \\
&= c\left[P(Y \leq y_0) + 1\right] \\
&= c\left[P\left(\frac{\log Y - \mu}{\sigma} \leq \frac{\log y_0 - \mu}{\sigma}\right) + 1\right] \\
&= c\left[\Phi\left(\frac{\log y_0 - \mu}{\sigma}\right) + 1\right] \\
&= c\left[\Phi(\alpha\sigma) + 1\right] \\
&= c[\Phi(k) + 1],
\end{aligned}$$

der $\Phi(k)$ er den kumulative standardnormalfordelingen av k. Nå kan ligningen ovenfor løses med hensyn på c, slik at jeg får at

$$c = \frac{1}{1 + \Phi(k)}. \tag{3.15}$$

Den endelige tetthetsfunksjonen

Nå kan alle resultatene settes inn i ligning 3.10 slik at jeg får sannsynlighetstettheten for den sammenslåtte lognormal-Paretofordelingen.

$$f(y) = \begin{cases} \dfrac{\alpha y_0^\alpha e^{-\frac{\alpha^2}{2k^2}(\log\frac{y}{y_0})^2}}{(1+\Phi(k))y^{\alpha+1}}, & \text{for } 0 < y \leq y_0, \\ \dfrac{\alpha y_0^\alpha}{(1+\Phi(k))y^{\alpha+1}}, & \text{for } y_0 \leq y < \infty\ , \end{cases} \tag{3.16}$$

med $\alpha > 0$.

Den kumulative fordelingsfunksjonen

Når jeg kjenner sannsynlighetstettheten til fordelingen, kan jeg finne den kumulative fordelingsfunksjonen.

$$F(y) = \begin{cases} \displaystyle\int_0^y f(z)dz, & \text{for } 0 < z \leq y_0, \\ \displaystyle 1 - \int_y^\infty f(z)dz, & \text{for } y_0 \leq z < \infty. \end{cases}$$

Jeg begynner med å løse integralet for situasjonen der $0 < z \leq y_0$.

$$\begin{aligned} P(Y \leq y) &= \frac{1}{1+\Phi(k)} P\left(\frac{\log Y - \mu}{\sigma} \leq \frac{\log y - \mu}{\sigma}\right) \\ &= \frac{1}{1+\Phi(k)} \Phi\left(\frac{\log y - \mu}{\sigma}\right) \\ &= \frac{1}{1+\Phi(k)} \Phi\left(\frac{\log y - \log y_0 + k\sigma}{\sigma}\right) \\ &= \frac{1}{1+\Phi(k)} \Phi\left(\frac{\log \frac{y}{y_0}}{\sigma}\right) \\ &= \frac{1}{1+\Phi(k)} \Phi\left(\frac{\alpha}{k} \log \frac{y}{y_0} + k\right). \end{aligned}$$

Her har jeg brukt resultatet fra ligning 3.13 om at $\mu = \log y_0 - \alpha\sigma^2$, samt at $\sigma = \frac{k}{\alpha}$. Jeg regner så ut den kumulative fordelingsfunksjonen når $y_0 \leq z < \infty$.

$$\begin{aligned} P(Y \leq y) &= 1 - P(Y \geq y) \\ &= 1 - \frac{\alpha y_0^\alpha}{1+\Phi(k)} \int_y^\infty \frac{1}{z^{\alpha+1}} dz \\ &= 1 - \frac{y_0^\alpha}{(1+\Phi(k))y^\alpha}. \end{aligned}$$

Tilsammen blir den kumulative fordelingsfunksjonen for den sammenslåtte lognormal-Paretofordelingen som i ligning 3.17 nedenfor.

$$F(y) = \begin{cases} \dfrac{1}{1+\Phi(k)}\Phi\left(\dfrac{\alpha}{k}\log\dfrac{y}{y_0}+k\right), & \text{for } 0 < y \leq y_0, \\ 1 - \dfrac{y_0^\alpha}{(1+\Phi(k))y^\alpha}, & \text{for } y_0 \leq y < \infty. \end{cases} \tag{3.17}$$

Forventingsverdi og varians

For å finne forventingsverdi og varians har jeg behov for momentene $E(Y)$ og $E(Y^2)$. Jeg finner derfor generelt n'te moment, med $n = 1, 2, \ldots$. Har at

$$E(Y^n) = \int_0^{y_0} z^n f_1(z)dz + \int_{y_0}^{\infty} z^n f_2(z)dz.$$

Jeg begynner med å løse det første integralet.

$$\begin{aligned} \int_0^{y_0} y^n f_1(y)dy &= \frac{1}{(1+\Phi(k))\sqrt{2\pi}\sigma}\int_0^{y_0} y^{n-1}e^{-\frac{1}{2\sigma^2}(\log y-\mu)^2}dy \\ &= \frac{e^{n\mu}}{(1+\Phi(k))\sqrt{2\pi}}\int_{-\infty}^{\frac{\log y_0-\mu}{\sigma}} e^{n\sigma z-\frac{z^2}{2}}dz \\ &= \frac{e^{n\mu+\frac{1}{2}n^2\sigma^2}}{(1+\Phi(k))\sqrt{2\pi}}\int_{-\infty}^{\frac{\log y_0-\mu}{\sigma}} e^{-\frac{1}{2}(z-n\sigma)^2}dz \\ &= \frac{e^{n\mu+\frac{1}{2}n^2\sigma^2}}{(1+\Phi(k))\sqrt{2\pi}}\int_{-\infty}^{\frac{\log y_0-\mu-n\sigma^2}{\sigma}} e^{-\frac{1}{2}v^2}dv \\ &= \frac{\Phi(\frac{\log y_0-\mu-n\sigma^2}{\sigma})}{1+\Phi(k)}e^{n\mu+\frac{1}{2}n^2\sigma^2} \\ &= \frac{y_0^n\Phi(k-\frac{nk}{\alpha})}{1+\Phi(k)}e^{\frac{nk}{\alpha}(\frac{n}{2}-\alpha)}. \end{aligned}$$

I den siste overgangen benyttet jeg ligningene 3.13 og 3.14 som beskriver forholdene mellom parameterene i lognormalfordelingen og Paretofordelingen. Løsningen av det andre integralet blir

$$\begin{aligned} \int_{y_0}^{\infty} y^n f_2(y)dy &= \frac{\alpha y_0^\alpha}{1+\Phi(k)}\int_{y_0}^{\infty} z^{n-\alpha-1}dz \\ &= \frac{\alpha y_0^n}{(1+\Phi(k))(\alpha-n)}, \qquad \text{for } n < \alpha. \end{aligned}$$

Dersom $n \geq \alpha$ blir løsningen av det siste integralet ∞. Samlet blir n'te moment for fordelingen

$$E(Y^n) = \frac{y_0^n}{1+\Phi(k)}\left(\frac{\alpha}{\alpha-n} + \Phi(k-\frac{nk}{\alpha})e^{\frac{nk}{\alpha}(\frac{n}{2}-\alpha)}\right), \quad \text{for } n < \alpha. \tag{3.18}$$

Jeg får forventingsverdien ved å sette inn $n = 1$ i formel 3.18 ovenfor.

$$E(Y) = \frac{y_0}{1+\Phi(k)}\left(\frac{\alpha}{\alpha-1} + \Phi(k-\frac{k}{\alpha})e^{\frac{k}{\alpha}(\frac{1}{2}-\alpha)}\right), \quad \text{for } \alpha > 1. \tag{3.19}$$

Jeg gjør det samme for å finne andremomentet.

$$E(Y^2) = \frac{y_0^2}{1+\Phi(k)}\left(\frac{\alpha}{\alpha-2} + \Phi(k-\frac{2k}{\alpha})e^{\frac{2k}{\alpha}(1-\alpha)}\right), \quad \text{for } \alpha > 2.$$

Nå har jeg alt jeg trenger for å finne variansen. Jeg benytter det kjente resultatet

$$Var(Y) = E(Y^2) - (E(Y))^2.$$

Da får jeg at variansen for den sammenslåtte lognormal-Paretofordelingen er

$$\begin{aligned} Var(Y) = &\frac{y_0^2}{1+\Phi(k)}\left(\frac{\alpha}{2-\alpha} + \Phi(k-\frac{2k}{\alpha})e^{\frac{2k}{\alpha}(1-\alpha)}\right) \\ &- \frac{y_0^2}{(1+\Phi(k))^2}\left[\frac{\alpha}{1-\alpha} + \Phi(k-\frac{k}{\alpha})e^{\frac{k}{\alpha}(\frac{1}{2}-\alpha)}\right]^2, \quad \text{for } \alpha > 2. \end{aligned} \tag{3.20}$$

En nyttig transformasjon

Det vil ved flere anledninger være nyttig å transformere den sammenslåtte lognormal-Paretofordelingen, på samme måte som for både lognormalfordelingen og Paretofordelingen hver for seg. Transformasjonene jeg har brukt for disse to fordelingene har vært på samme form, slik at det vil være naturlig å benytte den samme typen transformasjon også i det sammenslåtte tilfellet.

Jeg velger nå å benytte transformasjonen $Z = \log \frac{Y}{y_0}$ på begge tetthetsfunksjonene, som er den samme transformasjonen jeg benyttet på Paretofordelingen. Jeg får da en tetthetsfunksjon på formen

$$f_Z(z) = \begin{cases} f_1(z), & \text{for } -\infty < z \leq 0, \\ f_2(z), & \text{for } 0 \leq z < \infty. \end{cases}$$

Ved transformasjonen blir tetthetsfunksjonen for Z som følger.

$$f_Z(z) = \begin{cases} \dfrac{\alpha}{1+\Phi(k)} e^{-(\alpha z + \frac{\alpha^2}{2k^2} z^2)}, & \text{for } -\infty < z \leq 0, \\ \dfrac{\alpha e^{-\alpha z}}{1+\Phi(k)}, & \text{for } 0 \leq z < \infty. \end{cases} \tag{3.21}$$

Jeg vil også ha behov for den kumulative fordelingsfunksjonen. Denne finner jeg ved å integrere tetthetsfunksjonene i ligning 3.21 som nedenunder.

$$F_Z(z) = \begin{cases} \displaystyle\int_{-\infty}^{z} f_1(z)dz, & \text{for } -\infty < z \leq 0, \\ 1 - \displaystyle\int_{z}^{\infty} f_2(z)dz, & \text{for } 0 \leq z < \infty. \end{cases}$$

Integreringen gir følgende kumulative fordelingsfunksjon.

$$F_z(z) = \begin{cases} \dfrac{\Phi(\frac{\alpha z + k^2}{k})}{1+\Phi(k)}, & \text{for } -\infty < z \leq 0, \\ 1 - \dfrac{e^{-\alpha z}}{1+\Phi(k)}, & \text{for } 0 \leq z < \infty. \end{cases} \tag{3.22}$$

For å finne forventningsverdien i fordelingen, må jeg løse ligning 3.23.

$$E(Z) = \frac{\alpha}{1+\Phi(k)} \left[\int_{-\infty}^{0} z f_1(z)dz + \int_{0}^{\infty} z f_2(z)dz \right]. \tag{3.23}$$

Jeg starter med å finne det første integralet.

$$\begin{aligned}
&\frac{\alpha}{1+\Phi(k)}\int_{-\infty}^{0} ze^{-\alpha z-\frac{\alpha^2}{2k^2}z^2}dz \\
&= \frac{\alpha}{k\sqrt{2\pi}(1+\Phi(k))}\int_{-\infty}^{0} ze^{-\frac{\alpha^2}{2k^2}(z+\frac{k^2}{\alpha})^2}dz \\
&= \frac{1}{\sqrt{2\pi}(1+\Phi(k))}\int_{-\infty}^{k}\left(\frac{kv}{\alpha}-\frac{k^2}{\alpha}\right)e^{-\frac{v^2}{2}}dv \\
&= -\frac{k^2e^{-\frac{k^2}{2}}}{\alpha\sqrt{2\pi}(1+\Phi(k))}-\frac{k^2\Phi(k)}{\alpha(1+\Phi(k))} \\
&= -\frac{k^2}{\alpha(1+\Phi(k))}-\frac{k^2}{\alpha(1+\Phi(k))} \\
&= -\frac{k^2}{\alpha}.
\end{aligned}$$

I den nest siste overgangen har jeg benyttet ligning 3.14 som sier at

$$k\sqrt{2\pi} = e^{-\frac{k^2}{2}}.$$

Løsningen av det andre integralet blir som tidligere

$$\frac{\alpha}{1+\Phi(k)}\int_{0}^{\infty} ze^{-\alpha z}dz = \frac{1}{\alpha(1+\Phi(k)).}$$

Jeg setter løsningene av de to integralene inn i ligning 3.23 og finner at forventningsverdien for Z blir

$$E(Z) = \frac{1-k^2(1+\Phi(k))}{\alpha(1+\Phi(k))}. \tag{3.24}$$

3.4.2 Estimering

Nå vil jeg tilpasse den sammenslåtte lognormal-Paretofordelingen til dataene ved å finne SME $\hat{y}_0$ og $\hat{\alpha}$. Maksimering med hensyn på y_0 og α skjer numerisk i R. $\hat{y}_0$ er punktet der skiftet mellom lognormalfordelingen og Paretofordelingen vil foregå. Det innebærer at antall observasjoner med lognormal fordeling og Paretofordeling er ukjent. Det fremkommer av ligning 3.16 at

$$f_1(y) = f_2(y)e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y}{y_0}}.$$

Siden jeg vil bruke lognormal fordeling på de minste observasjonene og Paretofordelingen på de største, må jeg sortere skadegradene i økende rekkefølge. Da blir likelihoodfunksjonen

$$L_m(y_0, \alpha; \mathbf{y}) = \prod_{i=1}^{n} \frac{\alpha y_0^{\alpha}}{(1+\Phi(k))y_{(i)}^{\alpha+1}} \prod_{j=1}^{m} e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y_{(j)}}{y_0}}, \quad \text{med } m \in \{1,2,...,n\}.$$

Her er m antall observasjoner som tilordnes til lognormalfordelingen. Siden m er ukjent, blir det en likelihoodfunksjon for hver $m, medm \in \{1,2,...,n\}$. Det betyr at det blir like mange alternative $\hat{\alpha}$ og $\hat{y}_0$ som det er observasjoner. Til gitt m vil SME for y_0 oppfylle $y_0 \in [y_{(m)}, y_{(m+1)}]$. Likelihoodfunksjonen er monotont økende i y_0. I intervallet vil det derfor være den øvre grensen $\hat{y}_0 = y_{(m+1)}$ som maksimerer likelihoodfunksjonen. Dersom jeg velger $\hat{y}_0 = y_{(m+1)}$ for alle m, er $\hat{y}_0$ en kjent observasjon. Når jeg setter dette inn i likelihoodfunksjonen, får jeg derfor at det nå er m som må estimeres, og ikke y_0 som tidligere. Dermed kan jeg skrive likelihoodfunksjonen ovenfor på følgende måte.

$$L(m, \alpha; \mathbf{y}) = \frac{\alpha^n y_{(m+1)}^{n\alpha}}{(\prod_{i=1}^{n} y_{(i)})(1+\Phi(k))^n} e^{-\frac{\alpha^2}{2k^2}\sum_{j=1}^{m}\log^2\frac{y_{(j)}}{y_{(m+1)}}} \prod_{i=1}^{n} \frac{1}{y_{(i)}^{\alpha}}.$$

Ved å ta logaritmen av likelihoodfunksjonen, finner jeg loglikelihoodfunksjonen som skal maksimeres:

$$\begin{aligned} l(m, \alpha; \mathbf{y}) = {} & n\log\alpha + n\alpha\log y_{(m+1)} - \sum_{i=1}^{n}\log y_{(i)} - n\log(1+\Phi(k)) \\ & - \alpha\sum_{i=1}^{n}\log y_{(i)} - \frac{\alpha^2}{2k^2}\sum_{j=1}^{m}\log^2\frac{y_{(j)}}{y_{(m+1)}}. \end{aligned}$$

Nå kan jeg finne den $\hat{m}$ og den $\hat{\alpha}_{(m)}$ som maksimerer likelihoodfunksjonen. Her er $n = 6500$. Numerisk finner jeg for alle skipene at $\hat{m} = 2177$. Dermed blir SME for y_0 og α lik $(\hat{y}_0, \hat{\alpha}) = (0.013, 0.692)$. Det innebærer at overgangen mellom lognormal fordeling og Paretofordeling kommer ved observasjon nr 2177 når alle skipene er tatt med. Dersom $\hat{m}$ hadde kommet på en ende av datasettet, hadde det vært mest fornuftig å kun benytte seg av den ene fordelingen. Hadde den falt som en av observasjonene lengst til høyre, kunne lognormalfordelingen ha blitt brukt, mens en estimator som falt langt til venstre her ville betydd at en Paretofordeling ville være den beste av disse modellene. Siden mine beregninger baserer seg på at det er 6500 skip, er det tydelig at min estimator ligger langt inne i datasettet fra begge sider. Foreløpig ser det ut til at man skal kunne bruke den nye fordelingen. Når jeg

bare ser på tankskipene, skjer overgangen ved observasjon $\hat{m} = 475$. Antall tankskip er 1421, så også her er det langt ut til endene av datasettet. Med $\hat{y}_0 = 0.012$, blir $\hat{\alpha} = 0.755$. Fordi $\hat{\alpha} < 1$ i begge tilfellene, kan man finne verken forveningsverdi eller varians i modellen. Jeg oppsummerer resultatene for modellen i tabell 3.5.

	Alle skip	Tankskip
$\hat{m}$	2177	475
$\hat{\alpha}$	0.692	0.755
$\hat{y}_0$	0.013	0.012
n	6500	1421

Tabell 3.5: Resultater for sammenslått lognormal-Paretofordeling

3.4.3 Tilpasse sammenslått lognormal-Pareto fordeling

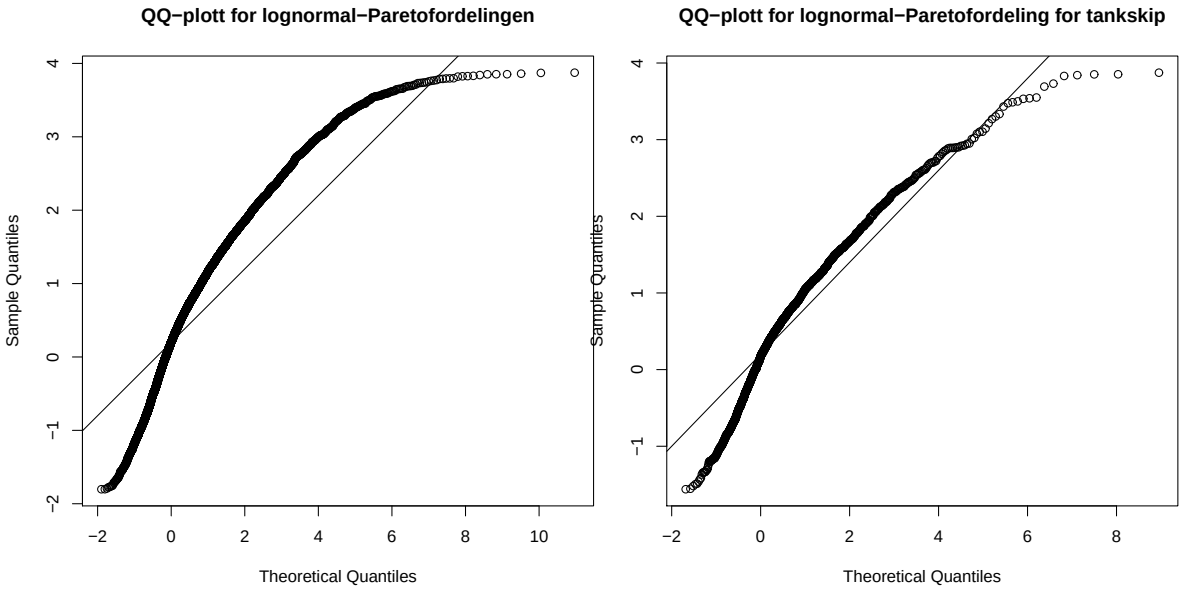
Jeg ønsker nå å tilpasse et QQ-plott for å se hvordan fordelingen passer til observasjonene. Jeg benytter transformasjonen av den sammenslåtte lognormal-Paretofordelingen fra avsnitt 3.4.1, og benytter den kumulative fordelingsfunksjonen for Z fra ligning 3.22.

$$F_Z(z) = \begin{cases} \dfrac{\Phi(\frac{\alpha z + k^2}{k})}{1 + \Phi(k)}, & \text{for } -\infty < z \leq 0, \\ \dfrac{1 - e^{-\alpha z}}{1 + \Phi(k)}, & \text{for } 0 \leq z < \infty. \end{cases}$$

For å lage QQ-plottet, må jeg finne den inverse kumulative fordelingsfunksjonen, som beskrevet i avsnitt 3.1. Jeg velger å sette $t = F_Z(z)$ og inverterer den kumulative fordelingsfunksjonen som i ligning 3.7, slik at jeg får at

$$z = \begin{cases} \dfrac{k}{\alpha}\Phi^{-1}[t(1 + \Phi(k))] - \dfrac{k^2}{\alpha}, & \text{for } 0 < t \leq \frac{\Phi(k)}{1+\Phi(k)}, \\ -\dfrac{\log[(1-t)(1+\Phi(k))]}{\alpha}, & \text{for } \frac{\Phi(k)}{1+\Phi(k)} \leq t < \infty. \end{cases}$$

Jeg setter deretter inn de n kvantilene for t og plotter funksjonen mot observasjonene, som også er ordnet i økende rekkefølge. Jeg får da følgende to QQ-plott for henholdsvis alle skipene og tankskipene.

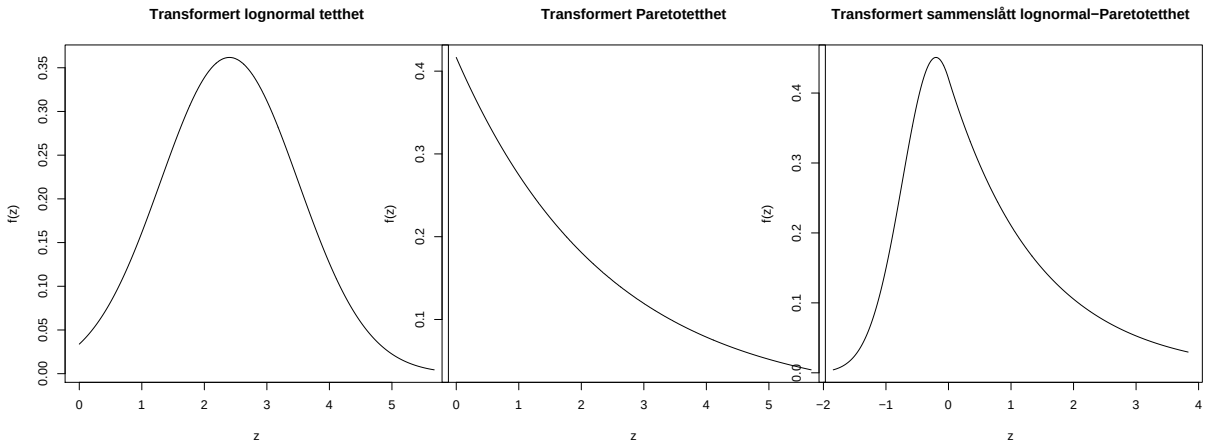


Figur 3.3: QQ-plott for sammenslått lognormal-Paretofordeling

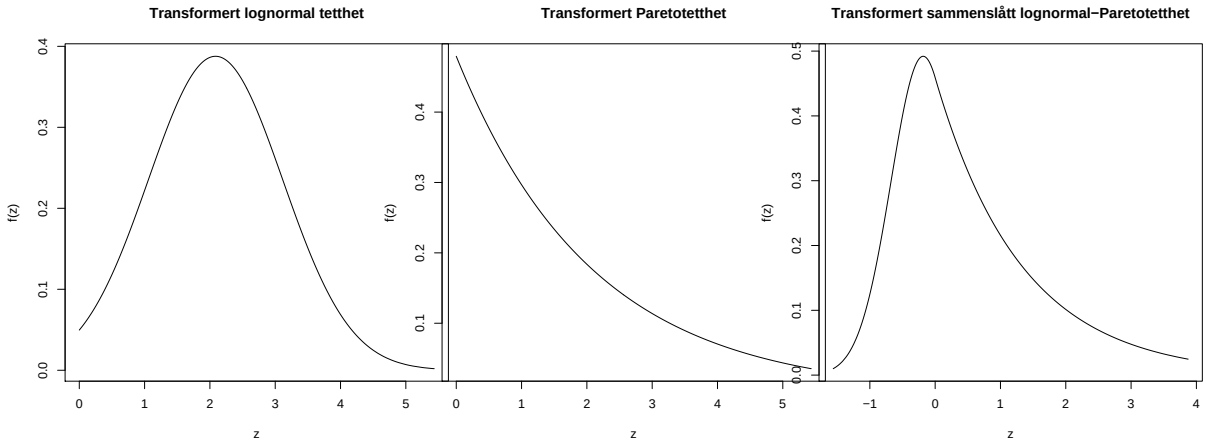
Det kan se ut som om denne tettheten passer bedre for tankskipene enn for alle skipene. For alle skipene ser det ut som om man kan benytte lognormalfordelingen og få et bedre resultat. QQ-plottet for tankskipene er heller ikke bedre her enn for lognormalfordelingen alene. I dette tilfellet ser det ikke ut til at den sammenslåtte lognormal-Paretofordelingen gir ønsker resultat.

3.5 Sammenligning av de tre modellene

Jeg ønsker nå å plotte de tre tetthetsfunksjonene mot hverandre for å undersøke hvordan disse ser ut i forhold til hverandre. For å få alle over på samme skala, har jeg plottet tetthetene med transformasjon $Z = \log \frac{Y}{y_0}$. For lognormalfordelingen og Paretofordelingen hver for seg, har vi at $\theta = y_0$.



Figur 3.4: Tettheter for skadegrad på alle skip.



Figur 3.5: Tettheter for skadegrad på tankskip.

Her ser man tydelig at den nye tetthetsfunksjonen er en kombinasjon av de to andre. Plottene er ganske like både for alle skipene og for tankskipene alene.

3.6 Egenandel

I modellene i forrige kapittel er det ikke tatt hensyn til egenandel ved forsikringsutbetalinger. Skader der skadegraden er mindre enn egenandelsgraden blir ikke rapportert til forsikringsselskapene. Derfor er ikke disse skadene registrert i datasettet, som dermed er ufullstendig. La Y være skadegraden og d egenandelsgraden. Da vi har observert Y betinget at $Y > d$ blir den kumulative fordelingsfunksjonen til den observerte Y

$$F_d(y) = P(Y \leq y \mid Y \geq d) = \frac{F(y)}{P(Y > d)} = \frac{F(y)}{1 - F(d)}, \quad \text{for } y \geq d.$$

Tettheten for den observerte Y blir

$$f_d(y) = \frac{d}{dy} F_d(y) = \frac{f(y)}{1 - F(d)}, \quad \text{for } y \geq d. \tag{3.25}$$

Dersom det ikke er noen egenandel, har vi $d = 0$ slik at $P(Y > 0) = 1$, og vi får samme tetthetsfunksjon som tidligere. Jeg skal nå se på hva denne endringen i antagelsene vil gjøre med de modellene som ble presentert i forrige kapittel.

3.6.1 Lognormal fordeling

For lognormalfordelingen har vi at

$$F(d) = P(Y < d) = \Phi(\frac{\log d - \mu}{\sigma}).$$

Jeg benytter igjen transformasjon til normalfordeling fra definisjon 6 ved å la $Z = \log Y$, slik at ligning 3.25 for Z blir

$$f_d(z) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(z-\mu)^2}}{1 - \Phi(\frac{\log d - \mu}{\sigma})}. \tag{3.26}$$

Her er både μ og σ ukjente parametere. Siden de samme parametrene forekommer i både teller og nevner, må disse estimeres ved hjelp av hele formelen. Dermed får man følgende likelihoodfunksjon.

$$L(\mu, \sigma^2; \mathbf{z}, \mathbf{d}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2} \prod_{i=1}^{n} \frac{1}{(1 - \Phi(\frac{\log d_i - \mu}{\sigma}))}.$$

Fra denne får man loglikelihoodfunksjonen som skal maksimeres med hensyn på μ og σ^2,

$$l(\mu, \sigma^2; \mathbf{z}, \mathbf{d}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i - \mu)^2 - \sum_{i=1}^{n}\log\left(1 - \Phi\left(\frac{\log d_i - \mu}{\sigma}\right)\right).$$

Resultatene for henholdsvis alle skipene og tankskipene fremkommer i tabellene 3.6 og 3.7.

	Alle skip			
	Estimat	Var	Se	Cov
$\hat{\mu}$	-4.9266	0.0030	0.0547	-0.0012
$\hat{\sigma}^2$	2.2264	0.0006	0.0253	-0.0012

Tabell 3.6: Resultater for estimater i lognormal fordeling for alle skip med egenandeler

	Tankskip			
	Estimat	Var	Se	Cov
$\hat{\mu}$	-4.7156	0.0069	0.0828	-0.0027
$\hat{\sigma}^2$	1.7205	0.0018	0.0421	-0.0027

Tabell 3.7: Resultater for estimater i lognormal fordeling for tankskip med egenandeler

Her ser vi at $\hat{\mu}$ blir mye lavere her enn for situasjonen uten egenandel, mens vi får et høyere estimat for σ^2. Forventningsverdi og varians for den venstretrunkerte modellen blir ikke de samme som for lognormalfordelingen generelt. Jeg må derfor beregne disse på nytt. Jeg beregner de n første mo-

mentene som følger.

$$\begin{aligned} E(Y^n) &= \frac{1}{1-\Phi(\frac{\log d-\mu}{\sigma})}\int_d^\infty y^n f(y)dy \\ &= \frac{1}{(1-\Phi(\frac{\log d-\mu}{\sigma}))\sqrt{2\pi}\sigma}\int_d^\infty y^{n-1}e^{-\frac{1}{2\sigma^2}(\log y-\mu)^2}dy \\ &= \frac{e^{n\mu+\frac{1}{2}n^2\sigma^2}}{(1-\Phi(\frac{\log d-\mu}{\sigma}))\sqrt{2\pi}}\int_{\frac{\log d-\mu}{\sigma}}^\infty e^{-\frac{1}{2}(v-n\sigma)^2}dv \\ &= \frac{1-\Phi(\frac{\log d-\mu-n\sigma^2}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}e^{n\mu+\frac{1}{2}n^2\sigma^2}. \end{aligned} \tag{3.27}$$

Nå kan jeg finne forventning og varians i den venstretrunkerte lognormalfordelingen ved å sette inn for n i ligning 3.27.

$$E(Y) = \frac{1-\Phi(\frac{\log d-\mu-\sigma^2}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}e^{\mu+\frac{1}{2}\sigma^2}. \tag{3.28}$$

$$E(Y^2) = \frac{1-\Phi(\frac{\log d-\mu-2\sigma^2}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}e^{2\mu+2\sigma^2}.$$

Jeg benytter igjen at

$$Var(Y) = E(Y^2) - (E(Y))^2.$$

Dermed får jeg at variansen i den venstretrunkerte lognormalfordelingen blir

$$\begin{aligned} Var(Y) =& \frac{e^{2\mu+\sigma^2}}{1-\Phi(\frac{\log d-\mu}{\sigma})}\times \\ & \left[(1-\Phi(\frac{\log d-\mu-2\sigma^2}{\sigma}))e^{\sigma^2} - \frac{(1-\Phi(\frac{\log d-\mu-\sigma^2}{\sigma}))^2}{1-\Phi(\frac{\log d-\mu}{\sigma})}\right]. \end{aligned} \tag{3.29}$$

Fordi det er ulike egenandeler på skadene, får jeg her en forventningsverdi og en varians for hver enkelt skade. Resultatene i tabell 3.8 er derfor gjennomsnittlig forventningsverdi og varians.

	Alle skip	Tankskip
$\overline{E(Y)}$	0.0446	0.0372
$\overline{\text{Var}(Y)}$	0.0069	0.0031

Tabell 3.8: Gjennomsnittlig forventningsverdi og varians i lognormalfordelingen med egenandel

Her ser man at både forventningsverdien og variansen ligger i nærheten av resultatene for lognormalfordelingen uten egenandel, men at begge deler er litt større i tilfellet med egenandel. Dette er å forvente, da resultatet uten egenandel baserer seg på at datasettet er fullstendig.

3.6.2 Paretofordelingen

Paretofordelingen har den kumulative fordelingsfunksjonen

$$P(Y < y) = 1 - \frac{\theta^\alpha}{y^\alpha}, \quad \text{for } y > \theta.$$

Når jeg setter dette inn i ligning 3.25, får jeg følgende tetthet for skadene.

$$f_d(y) = \frac{f(y)}{1 - F(d)} = \frac{\alpha\theta^\alpha d^\alpha}{y^{\alpha+1}\theta^\alpha} = \frac{\alpha d^\alpha}{y^{\alpha+1}} \quad \text{når } d > \theta$$

Siden θ forsvinner, blir dette den samme tettheten som for Paretofordelingen, hvor θ er byttet ut med d. Her er det ulik egenandel for hver observasjon. Jeg får følgende likelihoodfunksjon.

$$L(\alpha; \mathbf{y}, \mathbf{d}) = \alpha^n \prod_{i=1}^{n} \frac{d_i^\alpha}{y_i^{\alpha+1}}.$$

Loglikelihoodfunksjonen blir dermed

$$l(\alpha; \mathbf{y}, \mathbf{d}) = n \log \alpha + \alpha \sum_{i=1}^{n} \log d_i - (\alpha + 1) \sum_{i=1}^{n} \log y_i.$$

Resultatene fra estimeringen kan oppsummeres i tabell 3.9.

	Alle skip			Tankskip		
	Estimat	Var	se	Estimat	Var	se
$\hat{\alpha}$	0.7338674	0.000083	0.0091	0.7681377	0.00042	0.01229

Tabell 3.9: Resultater for estimater for Paretofordelingen med egenandel

Her ser vi at $\hat{\alpha}$ er mye større enn i tilfellet uten egenandel. Det betyr at tettheten i dette tilfellet begynner høyere oppe, og synker brattere enn Paretotettheten uten egenandel. Som i kapittelet uten egenandel, har vi at

$$E(Y) = \frac{\alpha d}{\alpha - 1}, \qquad \text{for } \alpha > 1,$$
$$Var(Y) = \frac{\alpha d^2}{(\alpha-1)^2(\alpha-2)}, \qquad \text{for } \alpha > 2.$$

Derfor finner man heller ikke her forventingsverdi eller varians.

3.6.3 Sammenslått lognormal-Pareto fordeling

Kumulativ fordelingsfunksjon

Fra ligning 3.25 finner jeg nå tettheten for den sammenslåtte lognormal-Paretofordelingen når det er tatt hensyn til at dataene er venstretrunkerte. Jeg må begynne med å finne $\bar{F}(d) = 1 - F(d) = P(Y \geq d)$. Denne avhenger av om d er større eller mindre enn y_0, slik at jeg må finne

$$\bar{F}(d) = \begin{cases} \bar{F}_1(d), & \text{for } d < y_0, \\ \bar{F}_2(d), & \text{for } d \geq y_0. \end{cases}$$

I ligning 3.17 har jeg funnet den kumulative fordelingsfunksjonen for den sammenslåtte lognormal-Pareto fordelingen. Når jeg setter inn i denne ligningen får jeg at

$$\bar{F}(d) = \begin{cases} \dfrac{1}{1+\Phi(k)}\left(1 + \Phi(k) - \Phi\left(\dfrac{\alpha}{k}\log\dfrac{d}{y_0} + k\right)\right), & \text{for } d < y_0, \\ \dfrac{y_0^\alpha}{(1+\Phi(k))d^\alpha}, & \text{for } d \geq y_0. \end{cases}$$

Jeg definerer nå Z som

$$z = \begin{cases} 1, & d < y_0, \\ 0, & d \geq y_0, \end{cases} \tag{3.30}$$

og får en enkelt funksjon for $\bar{F}(d)$.

$$\bar{F}(d) = \frac{y_0^{\alpha(1-z)}}{d^{\alpha(1-z)}} \frac{(1 + \Phi(k) - \Phi(\frac{\alpha}{k}\log\frac{d}{y_0} + k))^z}{1+\Phi(k)}. \tag{3.31}$$

Resultatet settes inn i ligning 3.25 for å få den nye tetthetsfunksjonen.

Ny tetthetsfunksjon

Nå har jeg det jeg trenger for å finne den nye tetthetsfunksjonen. Jeg har fra ligning 3.16 at tetthetsfunksjonen er på formen

$$f(y) = \begin{cases} f_1(y), & \text{for } 0 < y \leq y_0, \\ f_2(y), & \text{for } y_0 \leq y < \infty. \end{cases} \tag{3.32}$$

For å få den nye tetthetsfunksjonen setter jeg ligningene 3.31 og 3.16 inn i 3.32.

$$f_d(y) = \begin{cases} \dfrac{\alpha y_0^{z\alpha} d^{\alpha(1-z)}}{y^{\alpha+1}} \dfrac{e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y}{y_0}}}{(1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}, & \text{for } 0 < y \leq y_0, \\ \dfrac{\alpha y_0^{z\alpha} d^{\alpha(1-z)}}{y^{\alpha+1}(1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}, & \text{for } y_0 \leq y < \infty. \end{cases} \tag{3.33}$$

Forventningsverdi og varians

Siden jeg her regner med den venstretrunkerte tettheten for den sammenslåtte lognormal-Paretofordeling, må jeg finne forventningsverdi og varians på nytt. Jeg må dermed løse følgende for å finne de n første momentene.

$$E(Y^n) = \frac{\alpha y_0^{\alpha z} d^{\alpha(1-z)}}{(1-\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}\left[\int_d^{y_0} \frac{y^n}{y^{\alpha+1}} e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y}{y_0}} dy + \int_d^{\infty} \frac{y^n}{y^{\alpha+1}} dy\right]. \tag{3.34}$$

Fra utledningen av tetthetsfunksjonen er det kjent at

$$\frac{e^{-\frac{\alpha^2}{k^2}\log^2\frac{y}{y_0}}}{y^{\alpha+1}} = \frac{1}{\sqrt{2\pi}\sigma y \alpha y_0^{\alpha}} e^{-\frac{1}{2\sigma^2}(\log y-\mu)^2}.$$

Dette benytter jeg til å løse det første integralet i ligning 3.34.

$$\begin{aligned}
\int_d^{y_0} y^{n-\alpha-1} e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y}{y_0}} dy &= \frac{1}{\sqrt{2\pi}\sigma\alpha y_0^\alpha} \int_d^\infty y^{n-1} e^{-\frac{1}{2\sigma^2}(\log y-\mu)^2} dy \\
&= \frac{e^{n\mu+\frac{n^2\sigma^2}{2}}}{\sqrt{2\pi}\alpha y_0^\alpha} \int_{\frac{\log d-\mu}{\sigma}}^{\frac{\log y_0-\mu}{\sigma}} e^{-\frac{1}{2}(v-n\sigma)^2} dv \\
&= \frac{e^{n\mu+\frac{n^2\sigma^2}{2}}}{\alpha y_0^\alpha} \times \\
&\quad \left[\Phi(\frac{\log y_0-\mu-n\sigma^2}{\sigma}) - \Phi(\frac{\log d-\mu-n\sigma^2}{\sigma})\right] \\
&= \frac{y_0^n e^{\frac{nk}{\alpha}}}{\alpha y_0^\alpha} \left[\Phi(k-\frac{nk}{\alpha}) - \Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{nk}{\alpha})\right].
\end{aligned}$$

Her har jeg benyttet ligning 3.13 og 3.14 fra kapittel 3.4 som beskriver forholdet mellom parametrene μ,σ,α og y_0, samt konstanten k. Da får jeg som tidligere kun to parametere igjen. Det andre integralet i ligning 3.34, blir som tidligere.

$$\int_d^\infty y^{n-\alpha-1} dy = \frac{d^{\alpha-n}}{n-\alpha}, \qquad \text{for } n < \alpha.$$

Jeg kan nå sette resultatene inn i ligning 3.34. Da blir n'te moment for den sammenslåtte lognormal-Paretofordelingen

$$\begin{aligned}
E(Y^n) = &\frac{\alpha}{(1-\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z} \times \\
&\left(y_0^{z+n-1} d^{\alpha(1-z)} e^{\frac{nk}{\alpha}(\frac{nk}{2\alpha}-1)} \left[\Phi(k-\frac{nk}{\alpha}) - \Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{nk}{\alpha})\right] \right. \\
&\left. + \frac{y_0^{z\alpha} d^{n-z\alpha}}{\alpha-n} \right).
\end{aligned} \tag{3.35}$$

Nå kan jeg finne forventningsverdi og varians i fordelingen ved å sette inn for n i ligning 3.35.

$$E(Y) = \frac{\alpha}{(1-\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}\times \left(y_0^z d^{\alpha(1-z)} e^{\frac{k}{\alpha}(\frac{k}{2\alpha-1})}\left[\Phi(k-\frac{k}{\alpha})-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{k}{\alpha})\right] + \frac{y_0^{z\alpha}d^{1-z\alpha}}{\alpha-1}\right). \tag{3.36}$$

$$E(Y^2) = \frac{\alpha}{(1-\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}\times \left(y_0^{z+1} d^{\alpha(1-z)} e^{\frac{2k}{\alpha}(\frac{k}{\alpha}-1)}\left[\Phi(k-\frac{2k}{\alpha})-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{2k}{\alpha})\right] + x_0^{z\alpha}d^{2-z\alpha}\right). \tag{3.37}$$

Variansen finner man som tidligere ved å sette inn ligning 3.36 og 3.37 i

$$Var(Y) = E(Y^2) - E(Y)^2.$$

Da får jeg at variansen blir

$$Var(Y) = \frac{\alpha}{(1-\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^z}\times \left(y_0^{z+1} d^{\alpha(1-z)} e^{\frac{2k}{\alpha}(\frac{k}{\alpha}-1)}\left[\Phi(k-\frac{2k}{\alpha})-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{2k}{\alpha})\right] + y_0^{z\alpha}d^{2-z\alpha}\right) - \frac{\alpha^2}{(1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k))^{2z}}\times \left(y_0^z d^{\alpha(1-z)} e^{\frac{k}{\alpha}(\frac{k}{2\alpha-1})}\left[\Phi(k-\frac{k}{\alpha})-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k-\frac{k}{\alpha})\right] + \frac{y_0^{\alpha z}d^{1-\alpha z}}{\alpha-1}\right)^2. \tag{3.38}$$

Likelihoodfunksjon

Når jeg har funnet tettheten, kan jeg også finne likelihoodfunksjonen. Her er $(y_{(1)}, y_{(2)}, ..., y_{(n)})$ ordningsobservatoren for skadebeløp $1, 2, ..., n$. Tilhørende verdier for egenandel og indikatorfunksjon kaller jeg for $(d_{(1)}, d_{(2)}, ..., d_{(n)})$ og $(z_{(1)}, z_{(2)}, ..., z_{(n)})$.

$$L_m(y_0, \alpha; \mathbf{z}, \mathbf{y}, \mathbf{d}) = \alpha^n \prod_{i=1}^{n} \frac{y_0^{z_{(i)}\alpha} d_{(i)}^{\alpha(1-z_{(i)})}}{y_{(i)}^{\alpha+1}(1 + \Phi(k) - \Phi(\frac{\alpha}{k}\log\frac{d_{(i)}}{y_0} + k))^{z_{(i)}}} \times \prod_{j=1}^{m} e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y_{(j)}}{y_0}}.$$

Som i tilfellet uten egenandel, ønsker vi å finne den y_0 som faller i mellom observasjon m og $m+1$, altså $y_0 \in [y_{(m)}, y_{(m+1)}]$. Her er likelihoodfunksjonen monotont økende i y_0, slik at $\hat{y}_0 = y_{(m+1)}$ vil maksimere likelihoodfunksjonen i det gitte intervallet. Det betyr at y_0 for alle ulike m vil ha en kjent verdi. Derfor vil det være m som skal estimeres istedetfor y_0. Det gir følgende likelihoodfunksjon

$$L(m, \alpha, z; y, d) = \alpha^n \prod_{i=1}^{n} \frac{y_{(m+1)}^{z_{(i)}\alpha} d_{(i)}^{\alpha(1-z_{(i)})}}{y_{(i)}^{\alpha+1}(1 + \Phi(k) - \Phi(\frac{\alpha}{k}\log\frac{d_{(i)}}{y_{(m+1)}} + k))^{z_{(i)}}} \times \prod_{j=1}^{m} e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y_{(j)}}{y_{(m+1)}}}.$$

Ved å ta logaritmen finner jeg loglikelihoodfunksjonen som jeg kan maksimere for å finne SME.

$$\begin{aligned} l(m, \alpha, z; y, d) = {} & n\log\alpha + \alpha\log y_{(m+1)}\sum_{i=1}^{n} z_i \\ & + \alpha\sum_{i=1}^{n}(1 - z_i)\log d_i - (\alpha+1)\sum_{i=1}^{n}\log y_i \\ & - \sum_{i=1}^{n} z_i \log(1 + \Phi(k) - \Phi(\frac{\alpha}{k}\log\frac{d_i}{y_{(m+1)}} + k)) \\ & - \frac{\alpha^2}{2k^2}\sum_{j=1}^{m}\log^2\frac{y_j}{y_{(m+1)}}. \end{aligned}$$

SME finner jeg numerisk i R. Resultatene er oppsummert i tabell 3.10.

	Alle skip	Tankskip
$\hat{m}$	962	446
$\hat{\alpha}$	0.050	0.610
$\hat{y_0}$	0.007	0.012
n	6500	1421

Tabell 3.10: Resultater for sammenslått lognormal-Paretofordeling med egenandeler

I situasjonen uten egenandel, lå $\hat{\alpha}$ i samme størrelsesorden både for alle skipene og tankskipene. I dette tilfellet ligger $\hat{\alpha}$ for tankskipene i det samme området, mens denne for alle skipene er veldig liten. Når jeg undersøker verdiene for hver enkelt $\hat{\alpha}_{(m)}$, finner jeg at den største av disse er 0.05, slik at i jeg ville ha fått en liten verdi for $\hat{\alpha}$ uavhengig av ved hvilken observasjon jeg lot skiftet mellom fordelingene foregå. Her vil $\hat{y_0}$ være mindre enn tidligere fordi skiftet mellom fordelingene foregår mye tidligere enn det gjorde uten egenandeler. Jeg ser at også for tankskipene forekommer skiftet mellom fordelingene tidligere enn i tilfellet uten egenandeler, men her er det ikke så stor forandring som for alle skipene. Også her er $\hat{\alpha}$ lavere. Ved gjennomgang av $\hat{\alpha}_{(m)}$ for alle iterasjonene, ser jeg at de fleste verdiene også her er små, men at verdien ved den observasjonen som maksimerer likelihoodfunksjonen er en av de største. I denne situasjonen har $\hat{y_0}$ samme verdi som tidligere.

4 Regresjonsmodeller

Definisjon 8. *I en statistisk regresjonsmodell benyttes en eller flere forklaringsvariable* $\mathbf{X}$ *til å tilpasse en modell for responsvariablene* $\mathbf{Y}$*. Modellene er på formen*

$$\mathbf{Y} = f(\mathbf{X}, \beta),$$

hvor β *er parameterene som skal estimeres.*

4.1 AIC-kriteriet

Det finnes mange metoder for å finne ut hvor mange og hvilke responsvariabler som skal benyttes i regresjonsmodeller. Jeg velger å benytte AIC-kriteriet. AIC står for Akaike Information Criterion, og ble foreslått som en metode for valg av parametere av den japanske statistikeren Akaike i 1974. AIC-kriteriet settes opp som følger.

$$AIC = -2\log(L(\hat{\theta})) + 2p, \tag{4.1}$$

hvor $\hat{\theta}$ er sannsynlighetsmaksimeringsestimator for θ, $l(\hat{\theta})$ er likelihoodfunksjonen for $\hat{\theta}$, og p er antall frie parametere i modellen. Den beste modellen er den som minimerer AIC. En modell vil være mer nøyaktig desto mer informasjon den inneholder. I virkeligheten kan man ha svært lite informasjon å bygge på. Derfor bør modellen være så enkel som mulig, med få parametere. Tanken bak AIC-kriteriet er at det bare er de forklaringsvariablene som gir mye ny informasjon som tas med, og at man dermed straffes for å ta med mange parametere.

4.2 Lineær regresjon

En lineær regresjonsmodell er en modell som kan skrives på formen

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{4.2}$$

hvor

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

Y_1,Y_2,...,Y_n er uavhengige og $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
Her er $\mathbf{Y}$ responsvariablene som vi ønsker å lage en modell for. $\mathbf{X}$ er designmatrisen, som inneholder alle de aktuelle forklaringsvariablene for $\mathbf{Y}$. β er en vektor med parametere, mens ε er feilleddet. Ved regresjonsanalysen ønsker en å finne de verdiene for $\beta_j, j = 1, 2, ..., p$ som forklarer responsvariablene best. β er dermed en vektor som bestemmer hvordan de ulike forklaringsvariablene skal vektes. Siden $E(\epsilon) = \mathbf{0}$, får vi at

$$E(\mathbf{Y}) = \mu = \mathbf{X}\beta.$$

Variansen i regresjonsmodellen blir

$$\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\varepsilon) = \sigma^2.$$

Variansen i regresjonsmodellen kommer altså fra feilleddet, slik at modellen uten feilledd ville vært deterministisk. Antagelsen om normalfordeling av feilleddet gir at residualene også må være normalfordelte med

$$r_i = y_i - \hat{\mu}_i \sim \mathcal{N}(0, \sigma^2).$$

Fordi $\mu_i = \mathbf{x_i^T}\beta$, må også Y_i være normalfordelt med

$$Y_i \sim \mathcal{N}(\mathbf{x_i^T}\beta, \sigma^2) \quad \text{når} \quad \hat{\mu}_i \xrightarrow{p} \mu_i.$$

4.3 Generaliserte lineære modeller

Lineær regresjon er et spesialtilfelle av generaliserte lineære modeller. GLM kan benyttes på langt flere sannsynlighetsfordelinger enn normalfordelingen, men normalfordelingen er det enkleste tilfellet. For at den generelle teorien for GLM skal kunne benyttes på en sannsynlighetsfordeling, må noen kriterier oppfylles.

- Fordelingen må tilhøre den eksponensielle familien
- Fordelingen må være på kanonisk form

Generaliserte lineære modeller utnytter muligheten til å finne en transformasjon i sannsynlighetstetthetene slik at man kan benytte teorien for lineær regresjon også for data som ikke er normalfordelte. Transformasjonen foregår ved hjelp av en linkfunksjon. Anta at $E(Y_i) = \gamma_i$. En kanonisk linkfunksjon η_i blir da en funksjon som oppfyller

$$\eta_i = g(\gamma_i) = \mathbf{x_i^T}\beta. \tag{4.3}$$

4.3.1 Den eksponensielle familien

Definisjon 9. *En sannsynlighetstetthet tilhører den eksponensielle fordelingsfamilien dersom den kan skrives på formen*

$$f(y;\theta) = e^{a(y)b(\theta)+c(\theta)+d(y)}.$$

Dersom $a(y) = y$, sier vi at fordelingen er på kanonisk form.

Teorem 4.3.1. *For fordelinger i den eksponensielle familien gjelder*

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}.$$

Bevis. Generelt gjelder

$$\int_{-\infty}^{\infty} f(y;\theta)dy = 1.$$

Det gir at

$$0 = \frac{d}{d\theta}\int_{-\infty}^{\infty} f(y;\theta)dy = \int_{-\infty}^{\infty} \frac{d}{d\theta} f(y;\theta)dy.$$

For den eksponensielle familien får vi at den deriverte tettheten blir på følgende form.

$$\begin{aligned}\frac{df(y;\theta)}{d\theta} &= a(y)b'(\theta)e^{a(y)b(\theta)+c(\theta)+d(y)} + c'(\theta)e^{a(y)b(\theta)+c(\theta)+d(y)} \\ &= [a(y)b'(\theta) + c'(\theta)]f(y;\theta).\end{aligned}$$

Setter man denne deriverte inn i integralet ovenfor, blir det derfor som nedenunder.

$$\int_{-\infty}^{\infty}[a(y)b'(\theta)+c'(\theta)]f(y;\theta)dy=0,$$
$$b'(\theta)\int_{-\infty}^{\infty}a(y)f(y;\theta)dy+c'(\theta)\int_{-\infty}^{\infty}f(y;\theta)dy=0,$$
$$b'(\theta)E[a(Y)]+c'(\theta)=0.$$

Når man løser for $E(a[Y])$ finner man at

$$E[a(Y)]=-\frac{c'(\theta)}{b'(\theta)}.$$

□

Teorem 4.3.2. *For fordelinger i den eksponensielle familien gjelder*

$$Var[a(Y)]=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{(b'(\theta))^3}.$$

Bevis. Har generelt at

$$\int_{-\infty}^{\infty}\frac{d^2}{d\theta^2}f(y;\theta)dy=0.$$

For den eksponensielle familien har vi at den annenderiverte av tetthetsfunksjonen med hensyn på parameteren får følgende generelle form.

$$\frac{df^2(y,\theta)}{d\theta^2}=[a(y)b''(\theta)+c''(\theta)]f(y;\theta)+[a(y)b'(\theta)+c'(\theta)]^2f(y;\theta)=0.$$

Nå jeg setter den annenderiverte inn i integralet ovenfor, finner jeg følgende.

$$b''(\theta)\int_{-\infty}^{\infty}a(y)f(y;\theta)dy$$
$$+c''(\theta)\int_{-\infty}^{\infty}f(y;\theta)dy+b'(\theta)^2\int_{-\infty}^{\infty}[a(y)+\frac{c'(\theta)}{b'(\theta)}]^2f(y;\theta)dy=0.$$

Jeg setter nå inn resultatet fra teorem 4.3.1 ovenfor, slik at ligningen ser ut som følger.

$$b''(\theta)[-\frac{c'(\theta)}{b'(\theta)}]+c''(\theta)+b'(\theta)^2Var[a(Y)]=0.$$

Løsning med hensyn på $\text{Var}[a(Y)]$ gir det ønskede,

$$\text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3}.$$

□

4.3.2 Modellen

For å lage en generalisert lineær modell trenger jeg følgende elementer.

1. Scorefunksjonene $S_j = \frac{\delta l}{\delta \beta_j}$
2. Informasjonsmatrisen med elementer I_{jk}
3. En iterasjonsprosedyre for å løse $S_j = 0$, slik at man kan finne $\hat{\beta}$

Jeg går igjennom elementene punkt for punkt i avsnittene nedenfor.

4.3.3 Scorefunksjon

Scorefunksjonen er definert som

$$S(\theta) = \frac{\delta l(\theta)}{\delta \theta}. \tag{4.4}$$

Siden teorien for GLM forutsetter både at fordelingen er med i den eksponensielle familien og at den er på kanonisk form, kan vi sette opp en generell likelihoodfunksjon. Det er fullt mulig å lage generaliserte lineære modeller av tetthetsfunksjoner som ikke er på kanonisk form. Da kan man imidlertid ikke sette direkte inn i de generelle formlene som utledes nedenfor, men må foreta utregningene selv. Det samme gjelder dersom man ønsker å bruke en linkfunksjon som ikke er kanonisk. Nå ønsker jeg imidlertid å finne den generelle scorefunksjonen for den eksponensielle familien på kanonisk form. Først finner jeg den generelle formen på likelihoodfunksjonen.

$$L = \prod_{i=1}^{n} e^{y_i b(\theta_i) + c(\theta_i) + d(y_i)}.$$

Fra dette finner vi at loglikelihoodfunksjon som må deriveres for å finne scorefunksjonen er følgende.

$$l = \sum_{i=1}^{n} \Big(y_i b(\theta_i) + c(\theta_i) + d(y_i) \Big) = \sum_{i=1}^{n} l_i(\theta_i).$$

Anta at

$$\begin{aligned} \gamma_i &= E(Y_i), \\ \eta_i &= g(\gamma_i) = \mathbf{x_i^T}\beta. \end{aligned}$$

Da blir scorefunksjonen

$$S_j = \frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{\delta l_i}{\delta \theta_i} \frac{\delta \theta_i}{\delta \gamma_i} \frac{\delta \gamma_i}{\delta \beta_j}. \tag{4.5}$$

Nå utfører jeg derivasjonene hver for seg, og benytter samtidig teoremene 4.3.1 og 4.3.2.

$$\begin{aligned} \frac{\delta l_i}{\delta \theta_i} = Y_i b'(\theta_i) + c'(\theta_i) &= b'(\theta_i)(Y_i - \gamma_i), \\ \frac{\delta \theta_i}{\delta \gamma_i} = \frac{1}{\frac{\delta \gamma_i}{\delta \theta_i}} &= \frac{1}{b'(\theta_i)\mathrm{Var}(Y_i)}, \\ \frac{\delta \gamma_i}{\delta \beta_j} = \frac{\delta \gamma_i}{\delta \eta_i} \frac{\delta \eta_i}{\delta \beta_j} &= \frac{\delta \gamma_i}{\delta \eta_i} x_{ij}. \end{aligned}$$

Med det kan jeg sette opp den generelle scorefunksjonen.

$$S_j = \frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \gamma_i}{\mathrm{Var}(Y_i)} \frac{\delta \gamma_i}{\delta \eta_i} x_{ij}. \tag{4.6}$$

4.3.4 Informasjonsmatrisen

I modellen har jeg også behov for den observerte informasjonsmatrisen **I** fra definisjon 2. Asymptotisk vil denne også være den inverse kovariansmatrisen

for estimatene av β_j. Her behøver jeg den videre for å utføre iterasjonsprosedyren i neste avsnitt. Nå finner jeg element I_{jk} i informasjonsmatrisen.

$$I_{jk} = Cov(S_j, S_k) = E\left(\sum_{i=1}^{n} \frac{Y_i - \gamma_i}{\mathrm{Var}(Y_i)} x_{ij} \frac{\delta\gamma_i}{\delta\eta_i} \sum_{l=1}^{n} \frac{Y_l - \gamma_l}{\mathrm{Var}(Y_l)} x_{lk} \frac{\delta\gamma_l}{\delta\eta_l} \right)$$
$$= \sum_{i=1}^{n} \frac{E(Y_i - \gamma_i)^2}{\mathrm{Var}(Y_i)^2} x_{ij} x_{ik} (\frac{\delta\gamma_i}{\delta\eta_i})^2.$$

Forkortning gir at element I_{jk} i informasjonsmatrisen blir

$$I_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\delta\gamma_i}{\delta\eta_i} \right)^2. \tag{4.7}$$

4.3.5 Newton-Raphson iterasjonsprosedyre

For å finne SME i modellen, er det vanlig å finne toppunktet på kurven ved å velge $S_j = 0$ og løse for β_j. I GLM er det ofte vanskelig å løse uttrykkene for β_j, derfor er det behov for en numerisk løsningsmetode for å komme frem til estimatorene. Det er vanlig å bruke iterative vektede minste kvadraters estimatorer, som man kan komme frem til ved Newton-Raphson iterasjon. Jeg begynner med å utlede metoden med en parameter. Ønsker å løse

$$S(\theta) = 0.$$

Da kan man linearisere uttrykket slik at man får

$$S(\theta) \approx S(\theta^{(0)}) + S'(\theta^{(0)})(\theta - \theta^{(0)}) = 0,$$
$$\theta^{(1)} = \theta^{(0)} - \frac{S(\theta^{(0)})}{S'(\theta^{(0)})}.$$

Fra tidligere er det kjent at

$$I(\theta) = -\frac{\delta^2 l}{\delta\theta^2} = S'(\theta).$$

Dermed blir uttrykket ovenfor

$$\theta^{(1)} = \theta^{(0)} + \frac{S(\theta^{(0)})}{I(\theta^{(0)})}.$$

Dette kan gjøres mange ganger, helt til $\theta^{(m)}$ konvergerer mot en verdi. Løsningen kan overføres til situasjonen i GLM der man har mange parametere. Da får man følgende resultat.

$$\begin{aligned} \beta^{(\mathbf{m})} &= \beta^{(\mathbf{m-1})} + (\mathbf{I}^{(\mathbf{m-1})})^{-1}\mathbf{S}^{(\mathbf{m-1})}, \\ \mathbf{I}^{(\mathbf{m-1})}\beta^{(\mathbf{m})} &= \mathbf{I}^{(\mathbf{m-1})}\beta^{(\mathbf{m-1})} + \mathbf{S}^{(\mathbf{m-1})}. \end{aligned} \tag{4.8}$$

4.3.6 Iterative Weighted Least Squares (IWLS)

Man kan vise at Newton-Raphson iterasjonen gir samme løsning som en vektet minste kvadraters estimator. Jeg vet fra ligning 4.7 at

$$I_{jk} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\delta\gamma_i}{\delta\eta_i}\right)^2.$$

La

$$q_{ii} = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\delta\gamma_i}{\delta\eta_i}\right)^2, \tag{4.9}$$

og la $\mathbf{Q}$ være diagonalmatrisen

$$\mathbf{Q} = \begin{pmatrix} q_{11} & 0 & \cdots & 0 \\ 0 & q_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{nn} \end{pmatrix}. \tag{4.10}$$

Nå kan ligning 4.7 skrives om slik at vi får følgende formel for elementene i informasjonsmatrisen

$$I_{jk} = \sum_{i=1}^{n} x_{ij}x_{ik}q_{ii}.$$

Den observerte Fisherinformasjonen kan dermed skrives på matriseform som

$$\mathbf{I} = \mathbf{X}^T\mathbf{Q}\mathbf{X}. \tag{4.11}$$

Nå går jeg tilbake og ser på resultatet av Newton-Raphson iterasjonen. Fra ligning 4.8 vet vi at

$$\mathbf{I}^{(\mathbf{m-1})}\hat{\beta}^{(\mathbf{m})} = \mathbf{I}^{(\mathbf{m-1})}\hat{\beta}^{(\mathbf{m-1})} + \mathbf{S}^{(\mathbf{m-1})}.$$

Høyresiden av ligningen kan også skrives som summer. Da får vi

$$\begin{aligned}
\mathbf{I^{(m-1)}}\hat{\beta}^{\mathbf{(m)}} &= \sum_{k=1}^{p}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)}\left(\frac{\delta\gamma_i}{\delta\eta_i}\right)^2\hat{\beta}_k^{(m-1)} + \sum_{i=1}^{n}\frac{(Y_i-\gamma_i)x_{ij}}{\mathrm{Var}(Y_i)}\frac{\delta\gamma_i}{\delta\eta_i} \\
&= \sum_{k=1}^{p}\sum_{i=1}^{n}x_{ij}x_{ik}q_{ii}\hat{\beta}_k^{(m-1)} + \sum_{i=1}^{n}(Y_i-\gamma_i)x_{ij}q_{ii}\frac{\delta\eta_i}{\delta\gamma_i} \\
&= \sum_{i=1}^{n}q_{ii}x_{ij}\left[\sum_{k=1}^{p}x_{ik}\hat{\beta}_k^{(m-1)} + (Y_i-\gamma_i)\frac{\delta\eta_i}{\delta\gamma_i}\right] \\
&= \sum_{i=1}^{n}q_{ii}x_{ij}w_i,
\end{aligned}$$

hvor w_i er definert som

$$w_i = \sum_{k=1}^{p}x_{ik}\hat{\beta}_k^{(m-1)} + (Y_i-\gamma_i)\frac{\delta\eta_i}{\delta\gamma_i}. \tag{4.12}$$

Nå har jeg at

$$\mathbf{I^{(m-1)}}\hat{\beta}^{\mathbf{(m)}} = \mathbf{X}^T\mathbf{Q}\mathbf{w}.$$

Ved å bruke ligning 4.11 får jeg følgende uttrykk.

$$\mathbf{X^T}\mathbf{Q}\mathbf{X}\hat{\beta}^{\mathbf{(m)}} = \mathbf{X^T}\mathbf{Q}\mathbf{w}.$$

Løsning med hensyn på β gir

$$\hat{\beta}^{\mathbf{(m)}} = (\mathbf{X^T}\mathbf{Q}\mathbf{X})^{-1}\mathbf{X^T}\mathbf{Q}\mathbf{w}. \tag{4.13}$$

Fordi både $\mathbf{Q}$ og $\mathbf{w}$ i de fleste tilfeller vil inneholde $\hat{\beta}$, må ligningen løses iterativt. Det kan gjøres ved å benytte $\hat{\beta}^{\mathbf{(m-1)}}$ på høyre side av ligningen. Dette er den samme ligningen man får ved å benytte minste kvadraters estimatorer for $\hat{\beta}$.

Eksempel 4.3.3. Normalfordelingen
Nå vil jeg vise at prinsippene ovenfor stemmer godt for normalfordelingen. For normalfordelingen kan tetthetsfunksjonen skrives om slik at vi får

$$f(y) = e^{-\frac{1}{2\sigma^2}(y-\mu)^2-\frac{1}{2}\log 2\pi\sigma^2} = e^{\frac{\mu y}{\sigma^2}-\frac{y^2}{2\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log 2\pi\sigma^2}.$$

Nå er det greit å finne ut om normalfordelingen er med i den eksponensielle familien. I denne sammenhengen er det μ som er interessant, mens σ^2 oppfattes som en støyparameter.

$$a(y) = y, \qquad b(\mu) = \frac{\mu}{\sigma^2},$$
$$d(y) = -\frac{y^2}{2\sigma^2}, \qquad c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2.$$

Med dette ser vi at normalfordelingen er med i den eksponensielle familien. Siden $a(y) = y$ er den også på kanonisk form. Det betyr at den generelle teorien for GLM kan benyttes. Vi kan nå vise at teoremene 4.3.1 og 4.3.2 stemmer i denne situasjonen.

$$E(Y) = -\frac{c'(\mu)}{b'(\mu)} = \frac{\frac{\mu}{\sigma^2}}{\frac{1}{\sigma^2}} = \mu,$$

$$\mathrm{Var}(Y) = \frac{b''(\mu)c'(\mu) - c''(\mu)b'(\mu)}{(b'(\mu))^3} = \frac{\sigma^6}{\sigma^4} = \sigma^2.$$

Dette er kjente resultater for normalfordelingen. Ser da at linkfunksjonen blir

$$\eta_i = g(\mu_i) = \mu_i = \mathbf{x_i^T}\beta.$$

Derfor er det det ikke behov for transformasjon, og teorien for lineær regresjon kan benyttes direkte. Nå kan vi finne scorefunksjonen S_j ved å maksimere likelihoodfunksjonen med hensyn på β_j.

$$S_j = \frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{\delta l_i}{\delta \mu_i}\frac{\delta \mu_i}{\delta \beta_j}.$$

Ved derivasjon finner man greit at

$$S_j = \frac{\delta l}{\delta \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \gamma_i}{\mathrm{Var}(Y_i)}\frac{\delta \gamma_i}{\delta \eta_i}x_{ij} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)x_{ij}$$

På vektorform blir det

$$\mathbf{S} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{Y} - \frac{1}{\sigma^2}\mathbf{X}^T\mu = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\beta).$$

Metoden med SME innebærer at scorefunksjonen skal settes lik null for at likelihoodfunksjonen skal være størst mulig. Når jeg gjør dette, kan jeg i dette

tilfellet løse ligningen spesifikt på β, slik at jeg ikke behøver å benytte noen iterasjonsprosedyre for å finne estimatoren.

$$\begin{aligned} \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\beta) &= \mathbf{0} \\ \mathbf{X}^T\mathbf{Y} &= \mathbf{X}^T\mathbf{X}\beta \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \end{aligned} \tag{4.14}$$

Jeg vil nå vise at iterasjonsløsningen gir samme løsning som ligning 4.14. Jeg finner først informasjonsmatrisen I,

$$\begin{aligned} I_{jk} &= E(S_j S_k) \\ &= E\left[\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)x_{ij}\right)\left(\frac{1}{\sigma^2}\sum_{l=1}^{n}(y_l - \mu_l)x_{kl}\right)\right] \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{n} E(y_i - \mu)^2 x_{ij}x_{ik} \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}. \end{aligned}$$

På matriseform gir det

$$\mathbf{I} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}.$$

Dermed finner jeg at kovariansmatrisen for $\hat{\beta}$ asymptotisk er

$$\mathbf{\Sigma} = \mathbf{I}^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Vanligvis oppfattes σ^2 som en støyparameter i denne modellen. Denne er det også behov for å estimere dersom man ønsker å finne Fisherinformasjonen eller kovariansmatrisen. Jeg benytter da kjente resultater fra χ^2-fordelingen. Kvadrasummen i normalfordelingen blir

$$T = \sum_{i=1}^{n}(Y_i - \mu_i)^2.$$

Det er da kjent fra χ^2-fordelingen at

$$\frac{T}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2 \sim \chi^2(n).$$

I modellen estimerer jeg μ_i, slik at jeg må erstatte μ_i med $\hat{\mu}_i$ når jeg skal estimere σ^2. Jeg vil ikke vise utledningen av estimatet for σ^2. Siden $\hat{\mu}_i = \mathbf{x_i}^T\hat{\beta}$, benytter jeg meg av de p parametrene i $\hat{\beta}$ for å estimere σ^2. Derfor er det bare $n-p$ frie parametre igjen. Vi har at

$$\frac{\hat{T}}{\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 \sim \chi^2(n-p).$$

Dersom $Z \sim \chi^2(m)$, har vi forventningsverdi og varians som følger.

$$E(Z) = m, \quad \mathrm{Var}(Z) = 2m.$$

Med kjennskap til dette kan jeg velge en forventningsrett estimator for σ^2 som følger,

$$\hat{\sigma}^2 = \frac{\hat{T}}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2.$$

Ved å sette inn linkfunksjonen og skrive svaret på vektorform, får jeg at

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{x_i}^T\hat{\beta})^T(\mathbf{Y} - \mathbf{x_i}^T\hat{\beta}).$$

I det normalfordelte tilfellet er det mulig å løse ligning 4.14 direkte med hensyn på β, noe jeg har gjort ovenfor. Nå har jeg imidlertid nok informasjon til å benytte den iterative løsningen og vise at resultatet blir det samme. Har fra ligning 4.13 at

$$\hat{\beta}^{(\mathbf{m})} = (\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}\mathbf{w}.$$

For normalfordelingen får vi fra ligning 4.9 og 4.12 at

$$\begin{aligned} q_{ii} &= \frac{1}{\sigma^2}, \\ w_i &= \mathbf{x_i^T}\beta + (Y_i - \mathbf{x_i^T}\beta) = Y_i. \end{aligned}$$

Dermed kan vi sette inn i ligning 4.13 og finne løsningen

$$\beta^{(m)} = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

som er samme løsning som jeg fikk ved å løse scorefunksjonen direkte med hensyn på β i ligning 4.14. Det er tydelig at de oppsatte ligningene for generaliserte lineære modeller passer godt for normalfordelingen.

4.4 Estimering

Jeg vil nå benytte teorien om generaliserte lineære modeller til å tilpasse modeller for mine skadedata. Jeg vil som i forrige kapittel tilpasse lognormalfordelingen og Paretofordelingen til skadegradene. Jeg vil også se på hvordan den sammenslåtte lognormal-Paretofordelingen passer inn i rammeverket for GLM.

4.4.1 Lognormalfordelingen

Jeg vil begynne med å tilpasse en lognormalfordeling til dataene. Tettheten for lognormalfordeling kan skrives som

$$f(y) = e^{-\frac{\log^2 y}{2\sigma^2} \frac{\mu \log y}{\sigma^2} - \frac{\log^2 y}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2 - \log y}.$$

Ser da at lognormalfordelingen tilhører den eksponensielle familien med

$$\begin{aligned} a(y) &= \log y, & b(\mu) &= \frac{\mu}{\sigma^2}, \\ d(y) &= -\frac{\log^2 y}{2\sigma^2} - \log y, & c(\mu) &= -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2. \end{aligned}$$

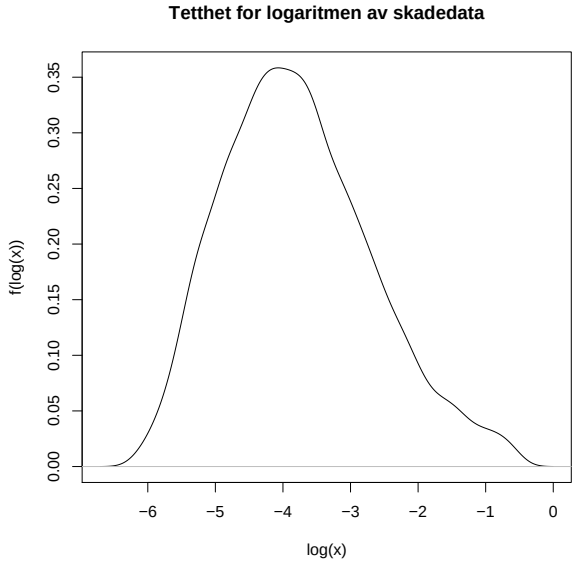
Her regnes σ^2 som en støyparameter, men denne kan som vist i eksemplet om normalfordeling også estimeres. Her kan man imidlertid se at lognormalfordelingen ikke er på kanonisk form, slik at den må transformeres til en tetthet på kanonisk form for å kunne benytte den generelle teorien om GLM. Jeg benytter definisjon 6 som sier at dersom $Y \sim \text{lognormal}(\mu, \sigma^2)$ og $Z = \log Y$ så er $Z \sim \mathcal{N}(\mu, \sigma^2)$. Fordi vi her har normalfordeling, kan vi nå benytte resultatene fra eksempel 4.3.3 og lage en regresjonsmodell med

$$\eta_i = E(Z_i) = \gamma_i = \mu_i = \mathbf{x_i^T}\beta.$$

Fra eksemplet er det også kjent at

$$\begin{aligned} S_j &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Z_i - \mu_i)x_{ij}, \\ \hat{\beta} &= (\mathbf{X^T X})^{-1}\mathbf{X^T Z}, \\ \mathbf{I} &= \frac{1}{\sigma^2}\mathbf{X^T X}. \end{aligned}$$

Figur 4.1 viser hvordan tettheten til logaritmen av skadegradene vil se ut. For å lage figuren har jeg benyttet programmet 'density' i R.



Figur 4.1: Tetthet til logaritmen av skadedata

Denne tettheten ønsker jeg å tilpasse ved hjelp av GLM. Jeg velger derfor de variablene som kan ha innvirkning på skadene, og står igjen med følgende forklaringsvariabler.

- β_0 = Skjæring med y-aksen
- β_1 =Logaritmen av skipets alder
- β_2 =Logaritmen av bruttotonn
- β_3 =Logaritmen av skipets verdi
- β_4 =Logaritmen av skipets hestekrefter
- β_5 =Indikator for type maskin

Jeg bruker AIC-kriteriet fra ligning 4.1 for å bestemme hvilke av forklaringsvariablene jeg skal ta med i modellen. For modellen med alle skipene, resulterer det at jeg bruker logaritmen av skipets alder, bruttotonn og verdi som forklaringsvariabler, mens for tankskipene benytter jeg logaritmen av skipets verdi og hestekrefter. Resultatene er oppsummert i tabell 4.1. Standardavviket i tabellen er diagonalen i den inverse informasjonsmatrisen. I

denne og alle de tilsvarende tabellene der AIC-kriteriet er benyttet, indikerer de tomme feltene at parameteren er fjernet fra modellen ved AIC-kriteriet.

	Alle skip		Tankskip	
	Estimat	Standardavvik	Estimat	Standardavvik
β_0	6.65607	0.26672	6.75468	0.48982
β_1	0.20388	0.02009		
β_2	0.18211	0.01236		
β_3	-0.77614	0.01764	-0.91188	0.03455
β_4			0.48456	0.04279
β_5				
N	6500		1421	
$N-p$	6496		1418	
R^2	0.3857		0.3338	

Tabell 4.1: Regresjonsmodell for lognormal fordeling

Felles for begge modellene er at R^2 ikke er særlig høy, altså at modellen ikke forklarer responsvariablene på en spesielt god måte. Jeg kan nå finne $\hat{\mu}_i$ og $\hat{\sigma}^2$. Fordi det er n ulike $\hat{\mu}_i$ har jeg her kun oppgitt gjennomsnittet av alle. Ved beregningen av $\mathrm{E}(Y)$ og $\mathrm{Var}(Y)$ har jeg imidlertid benyttet hver enkelt $\hat{\mu}_i$, slik at jeg også oppgir gjennomsnittlig forventningsverdi og varians for alle skadegradene. Ved hjelp av disse kan jeg regne ut forventningsverdi og varians. Resultatene finnes i tabell 4.2.

	Alle skipene	Tankskipene
$\hat{\mu}$	-3.7897	-3.8624
$\hat{\sigma}^2$	0.7473	0.7063
$\overline{E(Y)}$	0.0413	0.0363
$\overline{\mathrm{Var}(Y)}$	0.0029	0.0021

Tabell 4.2: Forventningsverdi og varians for lognormalfordelingen

Her kan man se at forventningsverdien og variansen ligger tett opp til estimatene i tabell 3.3 i kapittel 3, hvor jeg har benyttet SME.

4.4.2 Paretofordelingen

Nå vil jeg tilpasse en Paretofordelingen med en regresjonsmodell. Det er de samme forklaringsvariablene som er aktuelle som for lognormalfordelingen i

forrige avsnitt. Ser at Paretofordelingen er med i den eksponensielle familien siden

$$f(y) = \frac{\alpha\theta^\alpha}{y^{\alpha+1}} = e^{-\alpha\log y+\alpha\log\theta-\log y+\log\alpha},$$

$$\begin{aligned} \text{med} \quad a(y) &= \log y, \\ b(\alpha) &= -\alpha, \\ c(\alpha) &= \log\alpha - \alpha\log\theta, \\ d(y) &= -\log y. \end{aligned}$$

Her ser jeg fra ligningen at Paretofordelingen ikke er på kanonisk form. I kapittel 3 benyttet jeg en transformasjon til eksponensiell fordeling i definisjon 7. Jeg ønsker å benytte denne igjen, da den eksponensielle fordelingen er på kanonisk form med

$$\begin{aligned} f(z) &= \alpha e^{-\alpha z} = e^{-\alpha z+\log\alpha}, \\ a(z) &= z, \\ b(\alpha) &= -\alpha, \\ c(\alpha) &= \log\alpha, \\ d(z) &= 0. \end{aligned}$$

Fordi estimatet for θ i Paretofordelingen er den minste skadegraden og kan leses direkte ut fra dataene, kan jeg tilpasse en generalisert lineær modell til en eksponensiell fordeling med én parameter. I en GLM må som tidligere formelen for linkfunksjonen oppfylle

$$\eta_i = g(\gamma_i) = \mathbf{x_i^T}\beta.$$

Den kanoniske linkfunksjonen for denne tettheten vil være den inverse linkfunksjonen, som jeg velger å benytte.

$$\eta_i = E(Z_i) = \gamma_i = \frac{1}{\alpha_i} = \mathbf{x_i^T}\beta.$$

Nå kan jeg finne scorefunksjonen og informasjonsmatrisen.

$$S_j = \sum_{i=1}^{n} \frac{Z_i-\gamma_i}{\mathrm{Var}(Z_i)}\frac{\delta\gamma_i}{\delta\eta_i}x_{ij} = \sum_{i=1}^{n}\alpha_i(\alpha_i Z_i - 1)x_{ij}.$$

$$I_{jk} = \sum_{i=1}^{n} \frac{E(z_i-\gamma_i)^2}{\mathrm{Var}(Z_i)^2}\left(\frac{\delta\gamma_i}{\delta\eta_i}\right)^2 x_{ij}x_{ik} = \sum_{i=1}^{n}\alpha_i^2 x_{ij}x_{ik}.$$

For den eksponensielle fordelingen finner jeg at

$$q_{ii} = \frac{1}{\mathrm{Var}(Z_i)}(\frac{\delta\gamma_i}{\delta\eta_i})^2 \qquad = \alpha_i^2,$$
$$w_i = \mathbf{x_i^T}\beta + (Z_i - \mathbf{x_i^T}\beta) \quad = Z_i.$$

Nå kan jeg sette dette inn for $\hat{\beta}^{(\mathbf{m})}$ i ligning 4.13 slik at jeg finner en iterasjonsløsning.

$$\beta^{(\mathbf{m})} = (\mathbf{X^T Q X})^{-1}\mathbf{X^T Q Z}.$$

Her kan ikke $\mathbf{Q}$ fortkortes bort slik som tidligere for normalfordelingen. β er også inneholdt i $\mathbf{Q}$ slik at det er behov for å benytte iterasjonsprosedyren. Resultatene for alle skipene og tankskipene oppsummeres i tabell 4.3.

	Alle skip		Tankskip	
	Estimat	Standardavvik	Estimat	Standardavvik
β_0	-1.203	0.129	-1.673	0.357
β_1	-0.045	0.011	-0.032	0.028
β_2	-0.033	0.008	-0.047	0.015
β_3	0.120	0.009	0.165	0.023
β_4	0.012	0.011		
β_5				
n	6500		1421	
$n - p_0$	6499		1420	
$n - p_1$	6495		1417	
AIC	17321		3743.3	

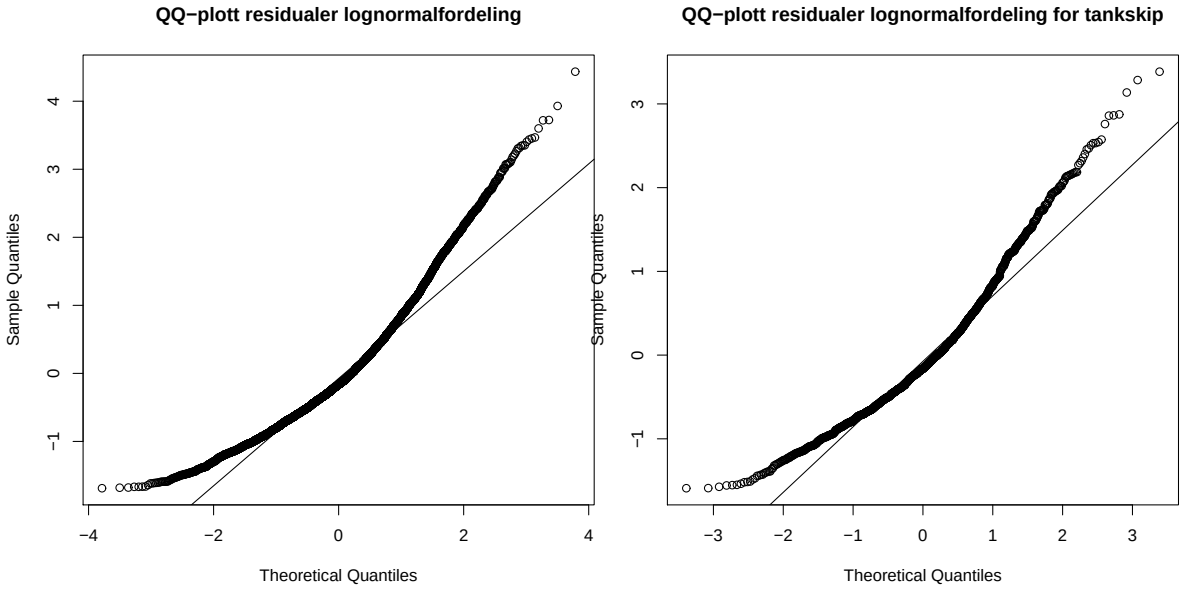
Tabell 4.3: GLM for Paretofordeling

De gjennomsnittlige verdiene jeg får for $\hat{\alpha}$ for henholdsvis alle skipene og tankskipene blir $(0.450, 0.513)$. Disse er litt større enn verdiene jeg fikk i kapittel 3 med beregning med SME. Her har jeg igjen brukt AIC-kriteriet fra ligning 4.1 for å velge ut forklaringsvariablene. Minste AIC får jeg dersom jeg fjerner maskintype og loghestekrefter som forklaringsvariabler for tankskipene, og kun maskintype for alle skipene.

4.5 Sammenligning av modellene

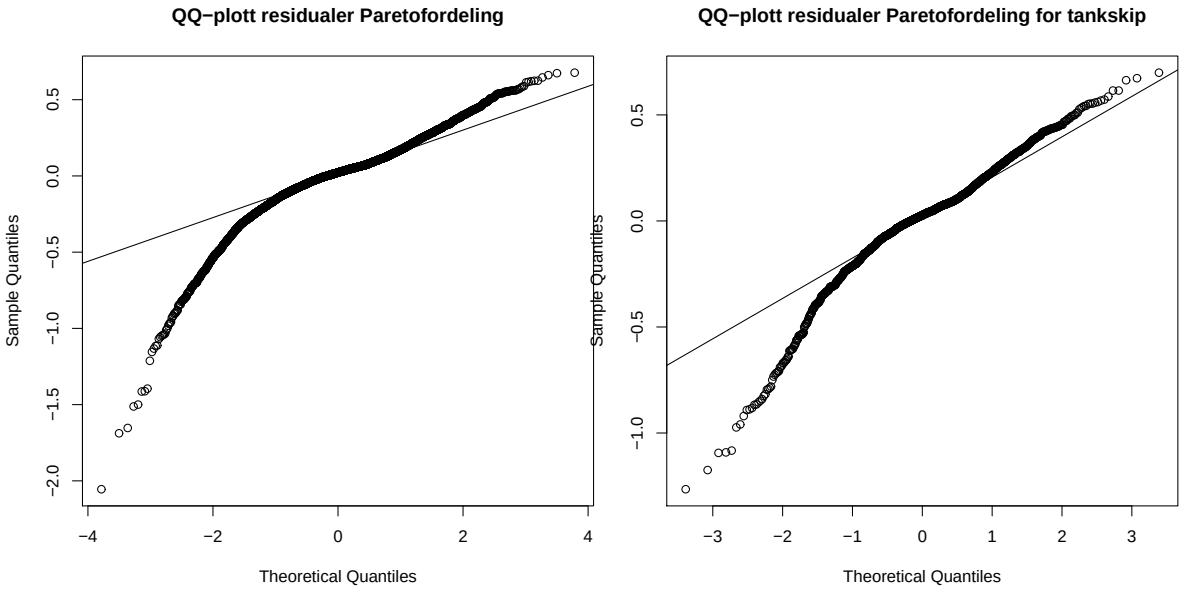
Asymptotisk vil residualene for generaliserte lineære modeller være normalfordelte med $r_i = y_i - \hat{\mu}_i \sim \mathcal{N}(0, \hat{\sigma}^2)$. Her kaller jeg $E(Y) = \mu$ og $\mathrm{Var}(Y) = \sigma^2$.

Dersom modellantagelsene om lognormalfordeling og Paretofordeling stemmer, bør derfor QQ-plottet for residualene danne en rett linje. Jeg ser på QQ-plottene for residualene til de tilpassede modellene.



Figur 4.2: QQ-plott for residualene i lognormalfordelingen

Her er det spesielt tydelig at lognormalfordelingen underestimerer de store skadegradene, både for alle skipene samlet, og for tankskipene alene. Det kan også se ut som om også de minste skadegradene underestimeres i begge tilfeller.



Figur 4.3: QQ-plott for residualene i Paretofordelingen

Fra QQ-plottet ser man at Paretofordelingen overestimerer de små skadene både for alle skipene og tankskipene alene. I tillegg underestimerer også Paretofordelingen de største skadene, men ikke på langt nær så mye som lognormalfordelingen. Når jeg har tilpasset GLM til skadegradene med både lognormalfordeling og Paretofordeling, er det mer tydelig enn i kapittel 3 at lognormalfordelingen passer best for de små skadegradene, og at Paretofordelingen passer best til de store. Det kan derfor være interessant å se på den sammenslåtte lognormal-Paretofordelingen i lys av teorien for GLM. Det er allikevel tydelig at ingen av fordelingen passer spesielt godt til skadedataene.

4.6 Sammenslått lognormal-Pareto fordeling

Tettheten for denne fordelingen er ikke på kanonisk form. Jeg ønsker å finne en transformasjon som gjør at jeg kan tilpasse en generalisert lineær modell til dataene. Fra ligning 3.16 er det kjent at tettheten i den sammenslåtte

lognormal-Paretofordelingen er

$$f(y) = \begin{cases} \dfrac{\alpha y_0^\alpha e^{-\frac{\alpha^2}{2k^2}(\log\frac{y}{y_0})^2}}{(1+\Phi(k))y^{\alpha+1}}, & \text{for } 0 < y \leq y_0, \\ \dfrac{\alpha y_0^\alpha}{(1+\Phi(k))y^{\alpha+1}}, & \text{for } y_0 \leq y < \infty, \end{cases}$$

med $\alpha > 0$. Jeg benytter transformasjonen $Z = \log\frac{Y}{y_0}$ fra definisjon 6 i kapittel 3, og ønsker å tilpasse skadegradene til fordelingen jeg fant i ligning 3.21.

$$f(z) = \begin{cases} \dfrac{\alpha}{1+\Phi(k)} e^{-(\alpha z + \frac{\alpha^2}{2k^2}z^2)}, & \text{for } -\infty < z \leq 0, \\ \dfrac{\alpha e^{-\alpha z}}{(1+\Phi(k))}, & \text{for } 0 \leq z < \infty. \end{cases}$$

Denne fordelingen har følgende kumulative fordelingsfunksjon, kjent fra ligning 3.22.

$$F_z(z) = \begin{cases} \dfrac{\Phi(\frac{\alpha z + k^2}{k})}{1+\Phi(k)}, & \text{for } -\infty < z \leq 0, \\ 1 - \dfrac{e^{-\alpha z}}{1+\Phi(k)}, & \text{for } 0 \leq z < \infty. \end{cases}$$

Denne tetthetsfunskjonen er heller ikke på kanonisk form. Derfor må man utføre alle beregningene for å komme frem til scorefunksjonen uten å benytte formelen som jeg frem i ligning 4.6.

$$S_j = \sum_{i=1}^{n} \frac{\delta l_i}{\delta \theta_i} \frac{\delta \theta_i}{\delta \gamma_i} \frac{\delta \gamma_i}{\delta \beta_j}.$$

I dette tilfellet innebærer det at man må finne likelihoodfunksjon og forventningsverdi for hver enkelt observasjon. Da må ordningsobservatoren fra definisjon 3.1.2 benyttes, og man må finne en sannsylighetstetthet for hver enkelt observasjon $Z_{(i)}$. Fra ligning 3.1 er det kjent at simultantettheten for alle $Z_{(i)}$ er

$$\begin{aligned} &f_{Z_{(1)},Z_{(2)},...,Z_{(n)}}(z_1, z_2, ..., z_n) = \\ &n! f_Z(z_1) f_Z(z_2) \times ... \times f_Z(z_n), \qquad -\infty < z_1 < z_2 < ... < z_n < \infty. \end{aligned}$$

Dette kan jeg nå benytte for å finne likelihoodfunksjonen for $Z_{(i)}$. Som tidligere for likelihoodfunksjonene for sammenslått lognormal-Paretofordeling, vil overgangen mellom de to tetthetene foregå ved ukjent observasjon m når observasjonene er ordnet i stigende rekkefølge. Også her benytter jeg at

$y_0 = y_{(m+1)}$ for å finne overgangen. Overgangen vil dermed foregå ved den m som gir størst likelihood.

$$\begin{aligned} L(m,\alpha;z) &= \frac{n!}{(1+\Phi(k))^m}\prod_{i=1}^{m}\alpha_{(i)}e^{-(\alpha_{(i)}z_{(i)}+\frac{\alpha_{(i)}^2}{2k^2}z_{(i)}^2)}\times \\ &\quad \frac{1}{(1+\Phi(k))^{n-m}}\prod_{l=m^n}\alpha_{(i)}e^{-\alpha_{(i)}z_{(i)}} \\ &= \frac{1}{(1+\Phi(k))^n}\prod_{i=1}^{n}\alpha_{(i)}e^{-\alpha_{(i)}z_{(i)}}\prod_{l=1}^{m}e^{-\frac{\alpha_{(i)}^2}{2k^2}z_{(i)}^2}. \end{aligned}$$

Med dette kan jeg også finne loglikelihoodfunksjonen som skal maksimeres.

$$\begin{aligned} l(m,\alpha;z) &= \log n! + n\log(1+\Phi(k)) + \sum_{i=1}^{n}\log\alpha_{(i)} \\ &\quad -\sum_{i=1}^{n}\alpha_{(i)}z_{(i)} - \frac{1}{2k^2}\sum_{l=1}^{m}\alpha_{(l)}^2 z_{(l)}^2. \end{aligned}$$

Da får vi at loglikelihoodfunksjonen som skal maksimeres for hver enkelt observasjon blir

$$l_{(i)} = \begin{cases} \log(1+\Phi(k)) + \log\alpha_{(i)} - \alpha_{(i)}z_{(i)} - \dfrac{\alpha_{(i)}^2}{2k^2}z_{(i)}^2, & \text{for } i \le m, \\ \log(1+\Phi(k)) + \log\alpha_{(i)} - \alpha_{(i)}z_{(i)}, & \text{for } i > m. \end{cases}$$

Derivasjon gir at

$$\frac{\delta l_{(i)}}{\delta\alpha_{(i)}} = \begin{cases} \dfrac{1}{\alpha_{(i)}} - z_{(i)} - \dfrac{\alpha_{(i)}}{k^2}z_{(i)}^2, & \text{for } i \le m \\ \dfrac{1}{\alpha_{(i)}} - z_{(i)}, & \text{for } i > m. \end{cases} \tag{4.15}$$

Tetthetsfunksjonen for ordningsobservatoren er kjent fra teorem 3.1.2.

$$f_{Z_{(i)}}(z) = \frac{n!}{(i-1!)(n-i)!}f_Z(z)[F_Z(z)]^{i-1}[1-F_Z(z)]^{n-i}.$$

Denne har man behov for dersom man ønsker å finne forventningsverdien for $Z_{(i)}$. Når jeg setter inn tetthetsfunksjonen og den kumulative fordelingsfunksjonen for Z i ligningen fra teorem 3.1.2 får jeg følgende tetthetsfunksjon for

$Z_{(i)}$.

$$f_{Z_{(i)}}(z) = \begin{cases} \dfrac{\Gamma(n+1)\alpha_{(i)}[1+\Phi(k)-\Phi(\frac{\alpha z+k^2}{k})]^{i-1}e^{-(\alpha_{(i)}z_{(i)}+\frac{\alpha_{(i)}^2}{2k^2}z_{(i)}^2)}}{\Gamma(i)\Gamma(n-i+1)(1+\Phi(k))^n}, & \text{for } i \leq m, \\ \dfrac{\Gamma(n+1)\alpha_{(i)}e^{-\alpha_{(i)}z_{(i)}(n-i+1)}[1+\Phi(k)-e^{-\alpha_{(i)}z_{(i)}}]^{i-1}}{\Gamma(i)\Gamma(n-i+1)(1+\Phi(k))^n}, & \text{for } i > m. \end{cases} \tag{4.16}$$

Forventningsverdien finner man da ved å løse integralene

$$EZ_(i) = \gamma_{(i)} =$$

$$\begin{cases} \dfrac{\Gamma(n+1)[1+\Phi(k)-\Phi(\frac{\alpha z+k^2}{k})]^{i-1}\alpha_{(i)}}{\Gamma(i)\Gamma(n-i+1)(1+\Phi(k))^n}\times & \\ \displaystyle\int_{-\infty}^{0} z_{(i)}e^{-(\alpha_{(i)}z_{(i)}+\frac{\alpha_{(i)}^2}{2k^2}z_{(i)}^2)}dz_{(i)}, & \text{for } i \leq m, \\ \dfrac{\Gamma(n+1)\alpha_{(i)}}{\Gamma(i)\Gamma(n-i+1)(1+\Phi(k))^n}\times & \\ \displaystyle\int_{0}^{\infty} z_{(i)}e^{-\alpha_{(i)}z_{(i)}(n-i+1)}[1+\Phi(k)-e^{-\alpha_{(i)}z_{(i)}}]^{i-1}dz_{(i)}, & \text{for } i > m. \end{cases} \tag{4.17}$$

Deretter må man finne en passende linkfunksjon og utførende de resterende derivasjonene i scorefunksjonen. Så settes ligning 4.15 og resultatet av utregningene i ligning 4.17 inn i ligning 4.6. Dette vil være et stort uttrykk, slik at man ikke vil kunne løse ligningen direkte med henyn på β. Man må dermed også finne informasjonsmatrisen, og sette alt inn i iterasjonsprosedyren fra avsnitt 4.3.6.

4.7 Egenandel

Som i kapittel 3 har jeg igjen det problemet at kun de skadene som er mer kostbare enn egenandelen blir innrapportert til forsikringsselskapene. Derfor ønsker jeg også her å lage modellene på nytt, men endre litt på fordelingene slik at de tar hensyn til dette. Jeg tilpasser derfor GLM til lognormalfordelingen og Paretofordelingen. Tilslutt ser jeg på hvordan man kan tilpasse den sammenslåtte lognormal-Paretofordelingen med egenandeler.

4.7.1 Lognormal fordeling

Fra ligning 3.26 er det kjent at tettheten for lognormalfordelingen justert for egenandel blir

$$f_d(y) = \frac{1}{\sqrt{2\pi}\sigma y}\frac{e^{-\frac{1}{2\sigma^2}(\log y-\mu)^2}}{1-\Phi(\frac{\log d-\mu}{\sigma})}.$$

Jeg benytter samme transformasjon som tidligere med $Y = e^Z$ fra definisjon 6, slik at jeg istedet får normalfordeling.

$$\begin{aligned} f_d(z) &= \frac{1}{\sqrt{2\pi}\sigma}\frac{e^{-\frac{1}{2\sigma^2}(z-\mu)^2}}{1-\Phi(\frac{\log d-\mu}{\sigma})} \\ &= e^{-\log\sqrt{2\pi}\sigma-\frac{z^2}{2\sigma^2}+\frac{z\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\log(1-\Phi(\frac{\log d-\mu}{\sigma^2}))}. \end{aligned}$$

Jeg ser at også denne normalfordelingen er med i den eksponensielle familien, og at den er på kanonisk form med

$$\begin{aligned} a(z) &= z, \\ b(\mu) &= \frac{\mu}{\sigma^2}, \\ c(\mu) &= -\log\sqrt{2\pi}\sigma - \frac{\mu^2}{2\sigma^2} - \log(1-\Phi(\frac{\log d-\mu}{\sigma})), \\ d(z) &= -\frac{z^2}{2\sigma^2}. \end{aligned}$$

Nå kan jeg finne forventningsverdi og varians i fordelingen. Antar videre at $\phi(\cdot)$ er tetthetsfunksjonen av standardnormalfordelingen. Jeg benytter teorien for GLM for å finne forventningsverdi og varians. Da har vi at

$$\begin{aligned} b'(\mu) &= \frac{1}{\sigma^2}, \\ b''(\mu) &= 0, \\ c'(\mu) &= -\frac{\mu}{\sigma^2} - \frac{\phi(\frac{\log d-\mu}{\sigma})}{\sigma(1-\Phi(\frac{\log d-\mu}{\sigma}))}, \\ c''(\mu) &= -\frac{1}{\sigma^2}\left(1+(\frac{\log d-\mu}{\sigma})\frac{\phi(\frac{\log d-\mu}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})} - \frac{\phi^2(\frac{\log d-\mu}{\sigma})}{(1-\Phi(\frac{\log d-\mu}{\sigma}))^2}\right). \end{aligned} \tag{4.18}$$

Med disse opplysningene finner jeg forventningsverdi og varians ved å sette de deriverte fra ligning 4.18 inn i teoremene 4.3.1 og 4.3.2.

$$E(Z) = -\frac{c'(\mu)}{b'(\mu)}, \quad \mathrm{Var}(Z) = \frac{b''(\mu)c'(\mu) - c''(\mu)b'(\mu)}{[b'(\mu)]^3}.$$

$$
\begin{aligned}
E(Z) &= \mu + \frac{\phi(\frac{\log d-\mu}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}\sigma, \\
\mathrm{Var}(Z) &= \sigma^2\left(1+(\frac{\log d-\mu}{\sigma})\frac{\phi(\frac{\log d-\mu}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})} - \frac{\phi^2(\frac{\log d-\mu}{\sigma})}{(1-\Phi(\frac{\log d-\mu}{\sigma}))^2}\right).
\end{aligned} \tag{4.19}
$$

Når jeg skal tilpasse en generalisert lineær modell, benytter jeg derfor at

$$E(Z_i) = \gamma_i = \mu + \frac{\phi(\frac{\log d-\mu}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}\sigma.$$

Jeg velger linkfunksjonen på samme måte som tidligere for normalfordelingen, slik at

$$\eta_i = \mu_i = \mathbf{x_i}^T\beta.$$

Dette er ikke den kanoniske linkfunksjonen i dette tilfellet, men ved å velge denne linkfunksjonen blir scorefunksjonen enklere enn ved den kanoniske linkfunksjonen. Loglikelihoodfunksjonen for Z etter transformasjonen blir

$$
\begin{aligned}
l(\mu,\sigma^2;\mathbf{z},\mathbf{d}) = &-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i-\mu_i)^2 \\
&- \sum_{i=1}^{n}\log\left(1-\Phi(\frac{\log d_i-\mu_i}{\sigma})\right).
\end{aligned}
$$

Før jeg kan finne scorefunksjonen trenger jeg kjennskap til $\frac{\delta\gamma}{\delta\eta} = \frac{\delta\gamma}{\delta\mu}$.

$$
\begin{aligned}
\frac{\delta\gamma}{\delta\eta} &= 1+\sigma\left[\frac{\frac{\delta}{\delta\mu}(\phi(\frac{\log d-\mu}{\sigma}))(1-\Phi(\frac{\log d-\mu}{\sigma})) - \phi(\frac{\log d-\mu}{\sigma})\frac{\delta}{\delta\mu}(1-\Phi(\frac{\log d-\mu}{\sigma}))}{(1-\Phi(\frac{\log d-\mu}{\sigma}))^2}\right] \\
&= 1+\frac{\log d-\mu}{\sigma}\frac{\phi(\frac{\log d-\mu}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})} - \frac{\phi^2(\frac{\log d-\mu}{\sigma})}{(1-\Phi(\frac{\log d-\mu}{\sigma}))^2} \\
&= \frac{1}{\sigma^2}\mathrm{Var}(Z).
\end{aligned}
$$

Nå kan jeg finne scorefunksjonen og informasjonsmatrisen.

$$
\begin{aligned}
S_j &= \sum_{i=1}^{n} \frac{z_i - \gamma_i}{\mathrm{Var}(Z_i)} \frac{\delta\gamma_i}{\delta\eta_i} x_{ij} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (z_i - \mu_i - \sigma \frac{\phi(\frac{\log d_i - \mu_i}{\sigma})}{1 - \Phi(\frac{\log d_i - \mu_i}{\sigma})}) x_{ij}, \\
I_{jk} &= E\left[\sum_{i=1}^{n} \frac{(z_i - \mu)^2}{\mathrm{Var}(Z_i)^2} (\frac{\delta\gamma_i}{\delta\eta_i})^2 x_{ij} x_{ik} \right] \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik} \left[1 + \frac{\log d_i - \mu_i}{\sigma} \frac{\phi(\frac{\log d_i - \mu_i}{\sigma_i})}{1 - \Phi(\frac{\log d_i - \mu_i}{\sigma})} - \frac{\phi^2(\frac{\log d_i - \mu_i}{\sigma})}{(1 - \Phi(\frac{\log d_i - \mu_i}{\sigma_i}))^2} \right].
\end{aligned}
$$

Jeg ønsker nå å finne iterasjonsløsningen for $\hat{\beta}^{(m)}$. I denne fordelingen finner jeg ved hjelp av ligning 4.9 og 4.12 at

$$
\begin{aligned}
q_{ii} &= \frac{1}{\mathrm{Var}(Z_i)} (\frac{\delta\gamma_i}{\delta\eta_i})^2 \\
&= \frac{1}{\sigma^2} \left[1 + \frac{\log d_i - \mu_i}{\sigma} \frac{\phi(\frac{\log d_i - \mu_i}{\sigma_i})}{1 - \Phi(\frac{\log d_i - \mu_i}{\sigma})} - \frac{\phi^2(\frac{\log d_i - \mu_i}{\sigma})}{(1 - \Phi(\frac{\log d_i - \mu_i}{\sigma_i}))^2} \right], \\
w_i &= \sum_{k=1}^{p} x_{ik} \hat{\beta}_k^{(m-1)} + (z_i - \gamma_i) \frac{\delta\eta_i}{\delta\gamma_i} \\
&= \sum_{k=1}^{p} x_{ik} \hat{\beta}_k^{(m-1)} + (z_i - \mu_i - \frac{\sigma\phi(\frac{\log d_i - \mu_i}{\sigma})}{1 - \Phi(\frac{\log d_i - \mu_i}{\sigma})}) \frac{\sigma^2}{\mathrm{Var}(Z)}.
\end{aligned}
$$

Her er $\hat{\beta}_j$ inneholdt i den kumulative normalfordelingsfunksjonen, slik at man må benytte iterasjonsløsningen og finne estimatene numerisk. Til dette har jeg brukt funksjonen 'truncreg' i pakken ved samme navn i R som tar for seg regresjonsmodeller med trunkerte data. Løsningen oppsummeres i tabell 4.4.

	Alle skip		Tankskip	
	Estimat	Standardavvik	Estimat	Standardavvik
β_0	7.026	0.4439	6.743	0.804
β_1	0.1468	0.0324		
β_2	0.156	0.0199		
β_3	-0.7978	0.0295	-0.898	0.057
β_4			0.4195	0.071
β_5				
σ	1.0685	0.0150	1.0561	0.0329
N	6500		1421	
$N-p$	6496		1418	

Tabell 4.4: Regresjonsmodell for lognormal fordeling med venstretrunkerte data

Her kan vi se at det blir de samme forklaringsvariablene som benyttes i modellen med egenandeler, som i den uten. Verdiene for de ulike $\hat{\beta}_j$ skiller seg heller ikke mye ut fra estimatene for skadegradene i situasjonen uten egenandel. Forventningsverdi og varians i den venstretrunkerte lognormalfordelingen er kjent fra kapittel 3 med ligning 3.28 og 3.29 som

$$E(Y) = \frac{1-\Phi(\frac{\log d-\mu-\sigma^2}{\sigma})}{1-\Phi(\frac{\log d-\mu}{\sigma})}e^{\mu+\frac{1}{2}\sigma^2},$$

$$Var(Y) = \frac{e^{2\mu+\sigma^2}}{1-\Phi(\frac{\log d-\mu}{\sigma})}\times \left[(1-\Phi(\frac{\log d-\mu-2\sigma^2}{\sigma}))e^{\sigma^2} - \frac{(1-\Phi(\frac{\log d-\mu-\sigma^2}{\sigma}))^2}{1-\Phi(\frac{\log d-\mu}{\sigma})}\right].$$

Jeg kan nå finne $\hat{\mu_i}$, og $\hat{\sigma}$ er kjent fra regresjonen. Resultatene for lognormalfordelingen med egenandeler oppsummeres i tabell 4.5. Som tidligere har jeg også her valgt å oppgi gjennomsnittene. Jeg har imidlertid benyttet alle $\hat{\mu}_i$ ved beregning av forventningsverdi og varians.

	Alle skip	Tankskip
$\overline{\hat{\mu}}$	-4.171	-4.260
$\hat{\sigma}^2$	1.142	1.115
$\overline{E(Y)}$	0.0457	0.0377
$\overline{\mathrm{Var}(Y)}$	0.0046	0.0033

Tabell 4.5: Gjennomsnittlig forventningsverdi og varians i lognormalfordelingen med egenandel

Disse verdiene stemmer godt overens med verdiene i tabell 3.8 i kapittel 3. Ser også her at både forventningsverdiene og variansene er litt større enn i situasjonen uten egenandel.

4.7.2 Paretofordeling

Som i avsnitt 4.4.2 må jeg også her transformere Paretofordelingen slik at jeg får en eksponensiell fordeling. Som i avsnitt 3.6.2 resulterer det i at

$$\theta = \min_i \{y_i\}$$

erstattes med egenandelen for hver enkelt skade. For å benytte eksponensiell fordeling, nå jeg bruke transformasjonen i ligning 4.20 på alle observasjonene.

$$Z_i = \log \frac{Y_i}{D_i}, \quad \text{for } i = 1, 2, ..., n. \tag{4.20}$$

Deretter kan jeg tilpasse en generalisert lineær modell på samme måte som i avsnittet 4.4.2 om Paretofordeling. Som tidligere har jeg da at

$$\begin{aligned} S_j &= \sum_{i=1}^{n} \alpha_i^2 x_{ij} z_i, \\ I_{jk} &= \sum_{i=1}^{n} \alpha_i^2 x_{ij} x_{ik}, \\ \hat{\beta}_j^{(m)} &= \frac{1}{n} \sum_{i=1}^{n} \frac{z_i}{x_{ij}}. \end{aligned}$$

Da får jeg resultatene i tabell 4.6.

	Alle skip		Tankskip	
	Estimat	Standardavvik	Estimat	Standardavvik
β_0	0.586	0.126	0.101	0.313
β_1	0.031	0.014	0.043	0.031
β_2	0.048	0.012		
β_3				
β_4	-0.042	0.018	0.060	0.032
β_5				
n	6500		1421	
$n - p_0$	6499		1420	
$n - p_1$	6496		1418	
AIC	14212		3082.6	

Tabell 4.6: GLM for Paretofordeling med egenandel

Gjennomsnittlig verdi av $\hat{\alpha}$ blir for henholdsvis alle skipene og tankskipene $(0.736, 0.771)$. Disse verdiene ligger i nærheten av verdiene fra kapittel 3 hvor estimatene er gjort med SME.

4.7.3 Sammenslått lognormal-Pareto fordeling

Fra ligning 3.33 i kapittel 3 er det kjent at tettheten for den sammenslåtte lognormal-Paretofordelingen med egenandel er som nedenfor.

$$f_d(y) = \begin{cases} \dfrac{\alpha x_0^{z\alpha} d^{\alpha(1-z)}}{y^{\alpha+1}} \dfrac{e^{-\frac{\alpha^2}{2k^2}\log^2\frac{y}{x_0}}}{(1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{x_0}+k))^z}, & \text{for } 0 < y \leq x_0, \\ \dfrac{\alpha x_0^{z\alpha} d^{\alpha(1-z)}}{y^{\alpha+1}(1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{x_0}+k))^z}, & \text{for } x_0 \leq y < \infty, \end{cases}$$

hvor

$$z = \begin{cases} 1, & d < y_0, \\ 0, & d \geq y_0. \end{cases}$$

Jeg velger å benytte samme transformasjon som for sammenslått lognormal-Paretofordelingen uten egenandel i avsnitt 3.4.1, med $T = \log\frac{Y}{y_0}$. Med det får jeg at tettheten for T blir som i ligning 4.21.

$$f_d(t) = \begin{cases} \dfrac{\alpha d^{\alpha(1-z)}}{[1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{x_0}+k)]^z} e^{-\frac{\alpha^2}{2k^2}t^2-\alpha t}, & \text{for} -\infty < t \leq 0, \\ \dfrac{\alpha d^{\alpha(1-z)}}{[1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{x_0}+k)]^z} e^{-\alpha t}, & \text{for } 0 \leq t < \infty. \end{cases} \tag{4.21}$$

Integrering på samme måte som tidligere gir den kumulative fordelingsfunksjonen.

$$F_d(t) = \begin{cases} \dfrac{d^{\alpha(1-z)}\Phi(\frac{\alpha t+k^2}{k})}{[1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k)]^z}, & \text{for } -\infty < t \leq 0, \\ 1 - \dfrac{d^{\alpha(1-z)}e^{-\alpha t}}{[1+\Phi(k)-\Phi(\frac{\alpha}{k}\log\frac{d}{y_0}+k)]^z}, & \text{for } 0 \leq t < \infty. \end{cases}$$

Jeg må benytte ordningsobservatoren når jeg skal finne scorefunksjonen S_j. Benytter da samme metode som i avsnitt 4.6 til å finne likelihoodfunksjonen.

$$L(\alpha, m, z; t, d) = n! \prod_{i=1}^{n} \frac{\alpha_{(i)} d_{(i)}^{\alpha_{(i)}(1-z_{(i)})} e^{-\alpha_{(i)} t_{(i)}}}{[1+\Phi(k)-\Phi(\frac{\alpha_{(i)}}{k}\log\frac{d_{(i)}}{y_0}+k)]^{z_{(i)}}} \prod_{l=1}^{m} e^{-\frac{\alpha_{(i)}^2}{2k^2} z_{(i)}^2}.$$

Loglikelihoodfunksjonen blir

$$\begin{aligned} l(\alpha, m, z; t, d) = \log n! &+ \sum_{i=1}^{n} \log \alpha_{(i)} + \sum_{i=1}^{n} \alpha_{(i)}(1-z_{(i)}) \log d_{(i)} \\ &- \sum_{i=1}^{n} \alpha_{(i)} t_{(i)} - \sum_{i=1}^{n} z_{(i)} \log[1+\Phi(k)-\Phi(\frac{\alpha_{(i)}}{k}\log\frac{d_{(i)}}{y_0}+k)] \\ &- \frac{1}{2k^2} \sum_{l=1}^{m} \alpha_{(l)}^2 t_{(l)}^2. \end{aligned}$$

Nå kan jeg finne loglikelihoodfunksjonen for hver enkelt observasjon, som jeg

igjen kan benytte til å finne scorefunksjonen.

$$l_{(i)}(m, \alpha_{(i)}, z_{(i)}, t_{(i)}, d_{(i)}) =$$

$$\begin{cases} \log \alpha_{(i)} + \alpha_{(i)}(1 - z(i)) \log d_{(i)} - \alpha_{(i)} z_{(i)} \\ - z_{(i)} \log(1 + \Phi(k) - \Phi(\frac{\alpha}{k} \log \frac{d_{(i)}}{x_0} + k)) - \frac{\alpha_{(i)}^2}{2k^2} t_{(i)}, & \text{for } i \leq m, \\ \log \alpha_{(i)} + \alpha_{(i)}(1 - z(i)) \log d_{(i)} - \alpha_{(i)} z_{(i)} \\ - z_{(i)} \log(1 + \Phi(k) - \Phi(\frac{\alpha}{k} \log \frac{d_{(i)}}{x_0} + k)), & \text{for } i > m. \end{cases} \tag{4.22}$$

Når jeg deriverer loglikelihoodfunksjonen i ligning 4.22 med hensyn på $\alpha_{(i)}$, får jeg

$$\frac{\delta l_{(i)}}{\delta \alpha_{(i)}} = \begin{cases} \frac{1}{\alpha_{(i)}} - z_{(i)} - \frac{\Phi(\frac{\alpha_{(i)}}{k} \log \frac{d_{(i)}}{x_0} + k)}{k} \log \frac{d_{(i)}}{x_0} - \frac{\alpha_{(i)}}{k^2} t_{(i)}, & \text{for } i \leq m, \\ \frac{1}{\alpha_{(i)}} - z_{(i)} - \frac{\Phi(\frac{\alpha_{(i)}}{k} \log \frac{d_{(i)}}{x_0} + k)}{k} \log \frac{d_{(i)}}{x_0}, & \text{for } i > m. \end{cases}$$

Som i kapittel 4.6, må jeg finne tetthetsfunksjonen for ordningsobservatoren for å finne forventningsverdien til $T_{(i)}$. Tetthetsfunksjonen blir et meget stort uttrykk, som igjen må integreres for å finne forventningsverdien. Utifra forventningsverdien må man så velge seg en linkfunksjon. Etter å ha utført alle nødvendige derivasjoner, kan disse settes inn i uttrykket for scorefunksjonen. Også her må man benytte en iterasjonsprosedyre for å finne verdier for β.

5 Oppsummering

I det foregående har jeg tilpasset et datasett med skadegrader på skip til henholdsvis lognormalfordelingen, Paretofordelingen og den sammenslåtte lognormal-Paretofordelingen. Ved tilpasning av fordelingene med SME, syntes ingen av fordelingsfunksjonene å gi en god tilpasning til skadegradene. Det ser heller ikke ut til at den sammenslåtte lognormal-Paretofordelingen gir en bedre tilpasning til dataene enn det lognormalfordelingen og Paretofordelingen kan gjøre hver for seg. Som nevnt i kapittel 3, er det mulig å tilpasse en sammenslått lognormal-Paretofordeling med tre parametere. Dette ville gitt en større fleksibilitet enn det jeg har fått i den modellen jeg har tilpasset. Utifra QQ-plottene for lognormalfordelingen og Paretofordelingen var det imidlertid ikke helt tydelig at lognormalfordelingen passet godt til de minste skadene, eller at Paretofordelingen ga en god tilpasning til de største skadene. I utgangspunktet var det dette som var begrunnelsen for å slå sammen de to fordelingene. Derfor ville heller ikke en tilpasning av den sammenslåtte lognormal-Paretofordelingen med tre parametere gi et veldig godt resultat. En måte å få bedre tilpasning til dette datasettet vil være å prøve en annen fordeling. Man kan også prøve å slå sammen to andre fordelinger. Hadde lognormalfordelingen og Paretofordelingen passet bedre til skadegradene i utgangspunktet hadde resultatet blitt et annet, og en sammenslått fordeling av disse to vil absolutt kunne gi et godt resultat i andre situasjoner. Ved tilpasning av lognormalfordelingen og Paretofordelingen med GLM fikk jeg resultater som samsvarte med de jeg fikk med SME. Det er også mulig å tilpasse den sammenslåtte lognormal-Paretofordelingen med to parametere med GLM. For denne fordelingen får man veldig store uttrykk, noe som gjør den vanskelig å jobbe med. Det kan imidlertid være vel verdt det dersom modellen blir betraktelig bedre enn alternativet.

Bibliografi

Casella, G, & Berger, R.L. 2002. *Statistical inference*. 2 edn. Pacific Grove, California: Duxbury. ISBN 0-534-24312-6.

Cooray, K, & Ananda, MMA. 2005. Modeling actuarial data with a composite lognormal-pareto model. *Scandinavian actuarial journal*, 321–334.

Dobson, A. 2002. *An introduction to generalized linear models*. 2 edn. London: Chapman & Hall. ISBN 1-58488-165-8.

Heuch, I. 2007. *Stat201,generaliserte lineære modeller*. Forelesningsrekke ved Matematisk institutt ved UiB.

Klugman, S.A, Panjer, H.H, & Willmot, G.E. 2004. *Loss models:from data to decisions*. 2 edn. Hoboken, New Jersey: Wiley Interscience. ISBN 978-0-471-68787-0.

Lindsey, J.K. 1997. *Applying generalized linear models*. 1 edn. New York: Springer-Verlag.

Pawitan, Y. 2001. *In all likelihood: statistical modelling and inference using likelihood*. Oxford: Oxford Science Publications. ISBN 978-0-19-850765-9.

R, Development core team. 2007. *An introduction to R*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-12-7.