Statistics Norway
Statistisk sentralbyrå

Ole Klungsøyr

Ekstremverdimodell for industrinæringenes investeringer i 90-årene

Notater

125 1876•2001 år

Seksjon for statistiske metoder og standarder
Emnegruppe: 10.90

Innholdsfortegnelse

1. Innledning

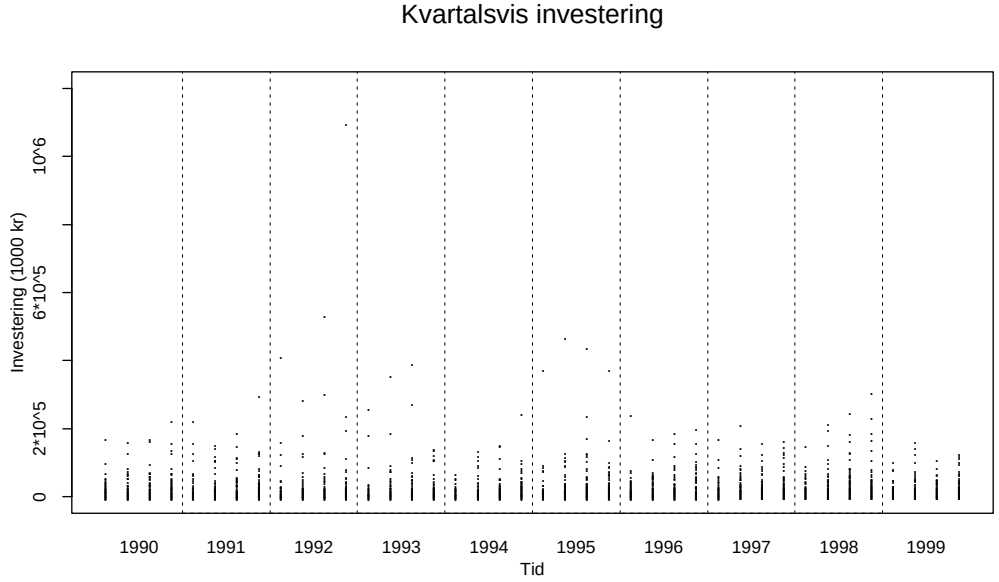
Den årlige statistikken over investeringer i industrinæringene i Norge bygger på data hentet inn på skjema til alle bedrifter i disse næringene - omtrent 12 000. Tallene publiseres i dag omtrent to år etter innsamling. Det er ønskelig å forbedre aktualiteten, samt publisere noen foreløpige totaltall basert på utvalget (ca. 2000 bedrifter) til den kvartalsvise investeringsstatistikken som sender inn skjema hvert kvartal. Det er også et mål å utnytte mer av informasjonen på skjemaene, og lage modeller slik at også foreløpige tall basert på utvalget kan publiseres.

2. Av ekstrem betydning

Ved estimering av totalen i en endelig populasjon der størrelsen på enhetene varierer mye er det opplagt at de største observasjonene har stor innflytelse på estimatet. I investeringsstatistikken kan de største bedriftene utgjøre mellom 5 og 20 % av utvalgs-totalen og har dermed stor innflytelse. Fordelingsestimater er ofte gode der man har mange observasjoner, og dårlige der man har få. Ekstremverditeori fokuserer nettopp på å estimere halen i fordelingen, der man har få og ekstreme observasjoner. Dette er motivasjonen for å prøve å anvende slike metoder på investeringsstatistikken. Hvis man klarer å lage gode modeller for de store bedriftenes adferd, kan dette kanskje gi bedre prediksjon av totaltallene.

3. Beskrivelse av data

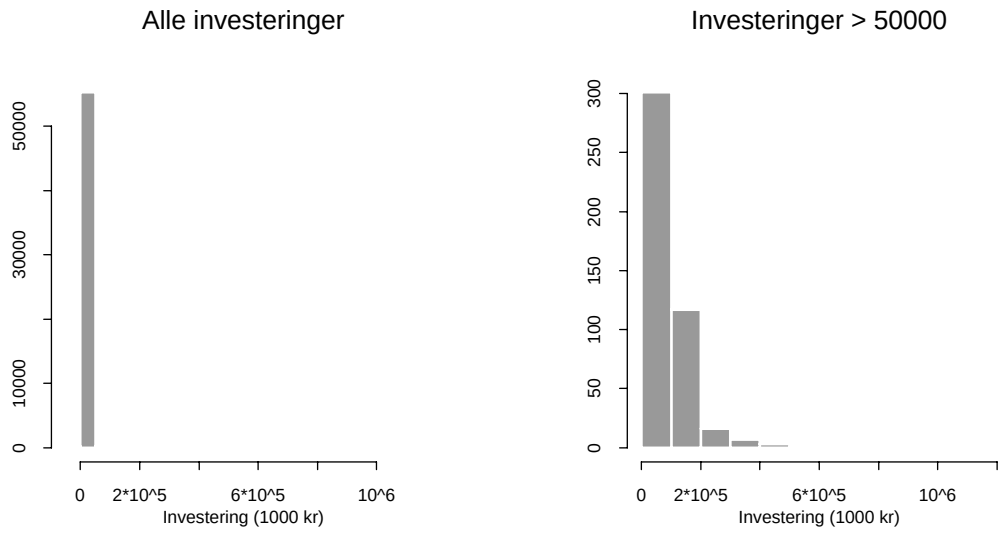
Figuren nedenfor viser dataene som analyseres her. Det er utvalgstall for industrinæringenes investeringer, hvert kvartal f.o.m 1990 t.o.m. 1999. Av hensyn til modellene som skal brukes er antall observasjoner pr. kvartal konstant, og lik 1 400 (den laveste utvalgs-størrelsen i perioden). Naturlig dynamikk i bedrifts-populasjonen (og uvalget) på grunn av konkurser, omdannelser, ny-skapninger gjør at uvalgs-bedriftene ikke er de samme hele tiden, men forandringene er små og spesielt blant de største. Ser man på de 1 400 største investeringene i utvalget hvert kvartal, er bedriftene stort sett de samme, og man er sikret å få med de ekstreme observasjonene.



Figur 1: Investering for hver bedrift, kvartalsvis fom 1990 tom 1999

Hver prikk i figuren er en bedrifts investering i et bestemt kvartal. Det er verdt å legge merke til de ekstreme investeringene, f.eks. i fjerde kvartal 1992.

Frekvens-histogrammene nedenfor viser kvartalene samlet. At dataene har "tung hale" kommer ikke frem ved å se på alle observasjonene siden de største er såpass få, men litt bedre når man ser på store observasjoner for seg (til høyre). Det er mange observasjoner med 0 (ingen investering).



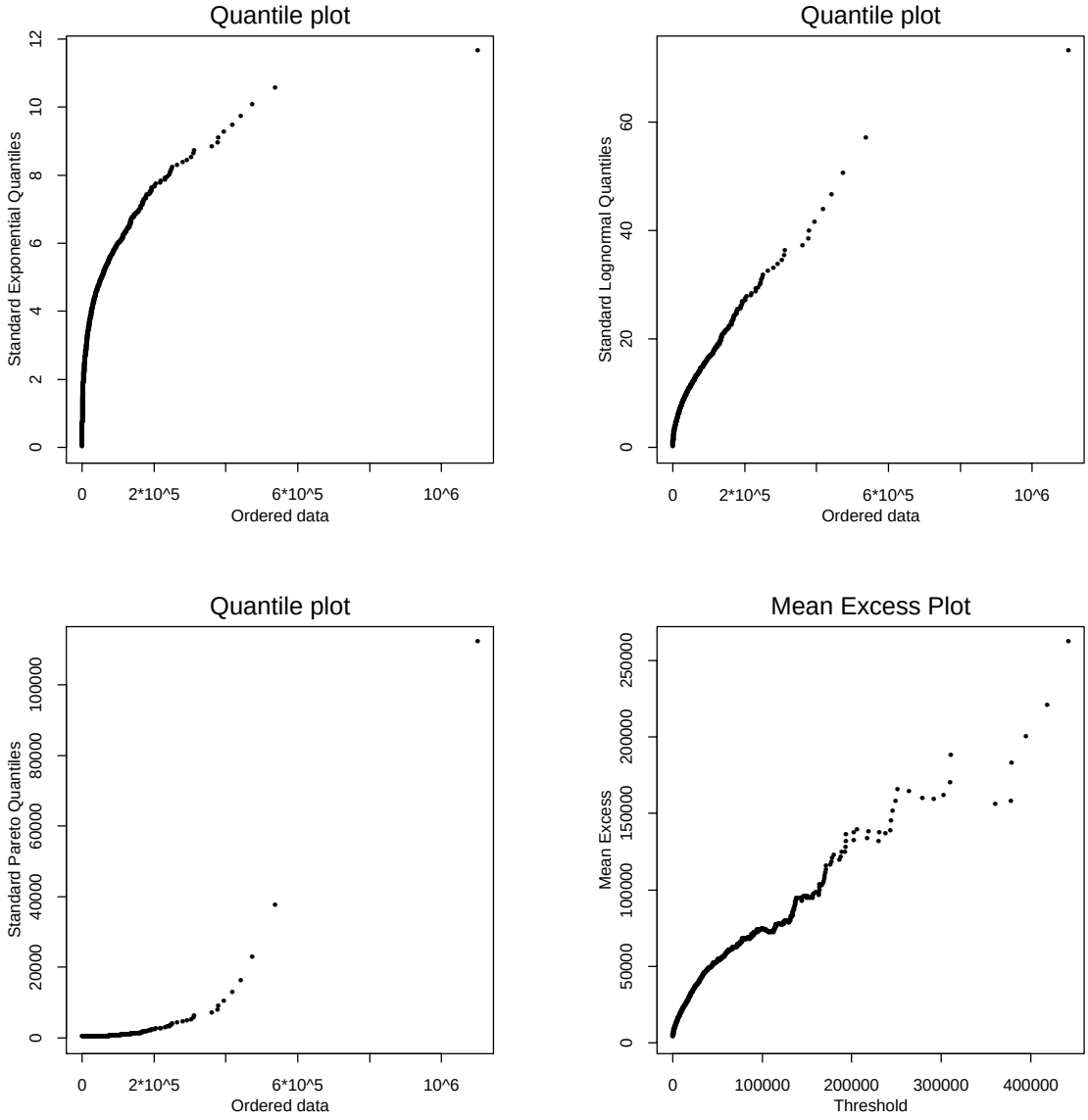
Figur 2: Histogram over kvartalene samlet

Tabellen nedenfor viser deskriptive fakta om noen utvalgte kvartaler som er forskjellige og alle kvartalene samlet. I de to siste kolonnene er det tatt med empiriske versjoner av skjevhets-koeffisienten og kurtosis altså $\frac{\hat{\mu}_3}{\hat{\sigma}^3}$ og $(\frac{\hat{\mu}_4}{\hat{\sigma}^4}) - 3$ som er mål på skjevhet og haletyngde for en fordeling, der μ_k er k-te ordens sentralmoment og σ er standardavviket. Disse størrelsene er begge 0 for standardnormal fordelingen og uberørte av lineærtransformasjoner med positive konstanter. For $k \geq 2$ har de empiriske momentene dårlige samplingsegenskaper ([1], s 321) så tallene her er nokså usikre. Kvartalene hver for seg har alle til felles at gjennomsnittet er mye større enn medianen, noe som beskriver skjevheten mot høyre. 4. kv. 92 er det kvartalet med den største enkeltobservasjonen og vi ser at dette kvartalet har størst forskjell mellom gjennomsnitt og median. Det har også størst verdi både på skjevhet og kurtosis. 1. kv. 93 er det kvartalet med minst median og har mindre skjevhet og kurtosis enn 4. kv. 92. 4. kv. 97 er det med størst median og samtidig ikke så stor max-verdi, og dette gir mindre skjevhet og kurtosis (blant de kvartalene med minst skjevhet og kurtosis), noe som også er intuitivt rimelig. Alle kvartalene samlet har karakteristikk som ligner litt på kvartalet med størst median, men med større kurtosis og mindre skjevhet.

	MIN	Q_1	MEDIAN	MEAN	Q_3	MAX	SD	SKJEVHET	KURTOSIS
4. kv. 92	63,17	224,3	650	3 847	1 987	1 101 000	31 425	30,96	1 060,4
1. kv. 93	36	147,2	390,5	2 088	1 209	263 900	9 886,4	18,97	443,6
4. kv. 97	147	455,5	1 226	4 576	3 572	170 000	11 949,1	7,66	77,8
Alle kv.	0	218	657	3 219	2 189	1 101 000	12 542,3	24,25	1 338,1

Tabell 1: Deskriptive fakta for noen kvartaler

For å illustrere haletyngden til dataene kan man ved kvantil-plott sammenligne med kjente fordelinger. Figuren nedenfor viser dataene sammenlignet med tre forskjellige fordelinger.

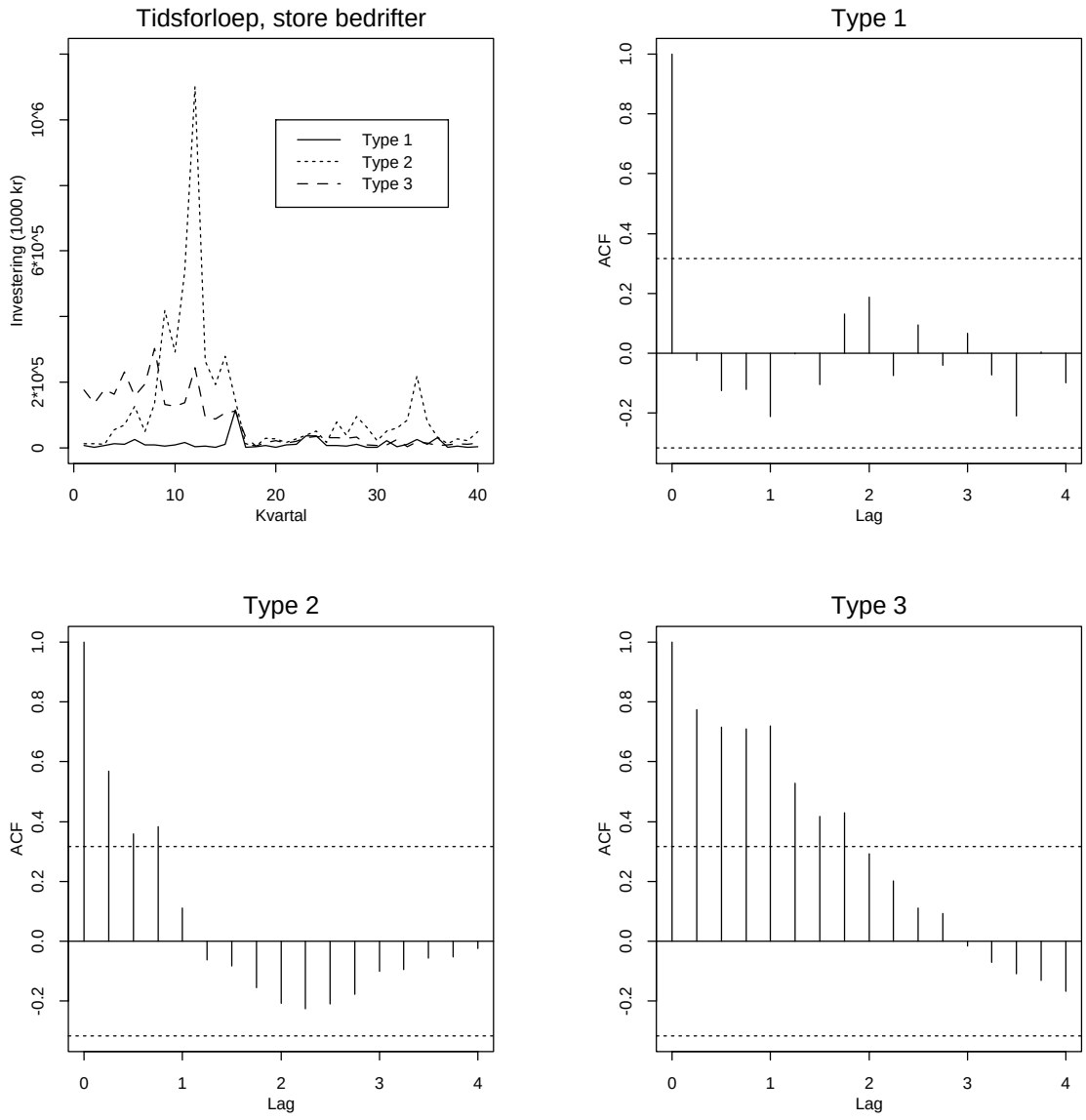


Figur 3: Investering sammenlignet med forskjellige fordelinger, og ME-plott

Hvis dataene er eksakt fordelt som den man sammenligner med, vil man få en rett linje. En "konkav" fasong på punktene indikerer tyngre hale enn referanse-fordelingen (gjelder for usymmetriske fordelinger med halen til høyre), og tilsvarende betyr en "konveks" fasong på punktene lettere hale. Øverst til venstre viser figuren at dataene har klart tyngre hale enn Eksponential fordelingen. Øverst til høyre ser vi at standard Lognormalfordelingen også har for lett hale til å gi god tilpasning til dataene. Begge disse fordelingene tilhører "Maximum Domain of Attraction" til Gumbel fordelingen (MDA(Λ)), se [1], selv om Eksponential fordelingen har lett hale mens Lognormal fordelingen har moderat tung hale. Pareto fordelingen har tung hale og tilhører "Maximum Domain of Attraction" til Frechet fordelingen (MDA(Φ_α)). Nederst til venstre sammenlignes dataene med standard Pareto fordeling og plottet antyder at dataene har lettere hale enn denne. Hvis man øker parameteren i Pareto fordelingen til ca. 2 vil kvantil plottet gi ganske rett linje (men dette kan også oppnås ved å øke parametrene tilstrekkelig i Lognormalfordelingen). Nederst til høyre er også "Mean Excess plottet" i

samsvar med tung hale, som gir stigende "Mean Excess" for stigende "Threshold". Hvis denne funksjonen er lineær over en viss grense (stor) tyder det på god tilpasning til Generalisert Pareto Fordeling (GPD) for "overskridelsene" av grensen. Dette kommer vi tilbake til.

En annen viktig karakteristikk av dataene er avhengighet. Siden det stort sett er samme bedrifter som er med hvert kvartal, vil investeringen til en bedrift et kvartal påvirke bedriftens investering neste kvartal, og kanskje mange kvartaler fremover. Dette vil rimeligvis ha uheldige konsekvenser for resultatene hvis man, som vi gjør her, bruker modeller som forutsetter uavhengighet mellom kvartalene. Det er imidlertid vanskelig å finne ut hvor mye det har å si, og det er også vanskelig å lage modeller som tar hensyn til akkurat den avhengigheten som finnes i våre data. Vi prøver å lage modeller som tar høyde for bestemte typer avhengighet som trend og sesong-effekt, men utover det vil vi analysere dataene som om de var uavhengige og identisk fordelte, noe som for øvrig også gjøres i [1] (s302 , avhengige data).



Figur 4: Tidsforløp for noen store bedrifter, og autokorrelasjon for disse

For å "fjerne" litt av avhengigheten kan man i GEV-modellen (neste avsnitt) som kun bruker maksimal-verdien, f.eks. en maksimal-verdi pr. kvartal, tenke seg å slå sammen flere kvartaler ("declustering") hvis det var slik at en investering stort sett hadde en bestemt varighet. Ved dermed kun å benytte en maksimal-verdi for hele investeringsperioden kunne man betrakte de etterfølgende periodene (maksimal-verdiene) som uavhengige.

I våre data varierer graden av avhengighet (autokorrelasjon) fra bedrift til bedrift, noe figur 4 (forrige side) viser. Vi har tatt med tre store bedrifter som har forskjellig form for og grad av avhengighet i tid. Figuren viser tidsforløp og autokorrelasjon for de tre bedriftene (prikkete linjer er omtrentlig 95 % konfidens-intervall). Her vises bare tre typer som er ganske forskjellige, det finnes også andre typer og mellomting i dataene. Den ene typen (Type 1) viser ingen korrelasjon i tid og her er det antakelig små konsekvenser å anta uavhengighet. Type 2 viser den bedriften med størst enkelt-observasjon, som viser signifikant positiv korrelasjon over en periode på ca. tre kvartal. Dvs. at man kunne kanskje slå sammen tre og tre kvartaler for å få mer uavhengighet. Type 3 viser sterk autokorrelasjon over mange kvartaler og dermed sterk avhengighet. Det kan hende at næringstilhørighet vil avsløre fellestrekk for avhengighet og kan dermed behandles separat, men dette er ikke gjort her.

Avhengigheten i dataene må vi ta i betraktning når vi tolker resultatene. Den vil påvirke nøyaktigheten i p-verdier og kan gjøre at vi undervurderer standard-feil til estimater.

4. Blokk Maksima Metode (GEV modell)

Denne metoden deler dataene først i blokker og benytter kun maksimal-verdiene i disse blokkene. Når man skal bruke denne metoden på et bestemt data-sett, er blokk-størrelsen kritisk. Fra teorien ([1,2]) har vi at maksimal-verdien for en blokk konvergerer i fordeling mot Generalisert Ekstrem Verdi Fordeling (GEV), altså når antall observasjoner i blokken går mot uendelig. På den annen side er det viktig med flest mulig blokker for god parameter-estimering siden det bare benyttes en observasjon fra hver blokk. Så for et gitt data-sett fås god tilnærming til GEV fordeling når blokken er stor, men på bekostning av kvaliteten til parameter-estimeringen i modellen.

I vårt tilfellet gir blokk-størrelsen seg selv som et kvartal. Vi har altså totalt 40 blokker med 1400 observasjoner i hver blokk og tenker oss følgende fremstilling:

$$\begin{array}{cccc|c} I_{1,1} & I_{1,2} & \cdots & I_{1,n} & M_{1,n} \\ I_{2,1} & \cdots & & I_{2,n} & M_{2,n} \\ \vdots & & & \vdots & \vdots \\ I_{m,1} & \cdots & & I_{m,n} & M_{m,n} \end{array} \tag{4.1}$$

Der $I_{i,j}$ er investering til bedrift j i kvartal i, $n = 1400$ og $m = 36$ (mrk. at $I_{1,1}$ og $I_{2,1}$ ikke nødvendigvis er samme bedriften). Investeringene innen et kvartal betraktes som uavhengige med identisk, men ukjent fordeling. $M_{1,n}, \cdots, M_{m,n}$ representerer maksimal-verdiene ($M_{1,n}$ er maksimum for $I_{1,1}, \cdots, I_{1,n}$ osv.) og betraktes som uavhengige og identisk fordelte. Siden n er stor approksimeres fordelingen til maksimal-verdiene med grensefordelingen, altså GEV fordelingen. Med vanlig notasjon (bytter ut $M_{i,n}$ med $X_{i,n}$) har vi flg. fordelingsantagelse:

$$H_{\xi;\mu,\sigma}(x) = \Pr\{X \le x\} = \exp\left\{-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}, \quad \text{der } 1+\xi\frac{x-\mu}{\sigma} > 0 \tag{4.2}$$

som gir tettheten:

(4.3) $$h_{\xi;\mu,\sigma}(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

Når $\xi = 0$ blir fordelingen litt annerledes (lar ξ gå mot 0 i (4.2)), men vi skal se at for våre data estimeres ξ til å være > 0. Dette tilsvarer den av tre mulige grensefordelinger for maksimal-verdiene (normaliserte) som har tyngst hale, nemlig Frechet fordelingen ($\Phi_\alpha(x)$). (4.2) er en reparametrisering av denne. For å estimere parametere i modellen vil vi se på to metoder. Den ene metoden er Maksimum Likelihood estimering og består i å maksimere log-likelihooden (se nedenfor). Den andre er den såkalte "Probability Weighted Moments" metode og ligner på andre moment-metoder (se nedenfor).

Når det gjelder kvantil-estimering, tas utgangspunkt i den teoretiske kvantilen og man setter inn estimater for parametrene. $(1-p)$ kvantilen q_p til tettheten i (4.3) er gitt ved:

(4.4) $$q_p = \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right]$$

Dermed er estimatet $\hat{q}_p$ gitt ved å sette inn estimerte parametere i (4.4).

Blokk Maksima Metode er benyttet på investerings-dataene med forskjellige modeller. Først er alle dataene analysert samlet med antagelse om identisk fordelte maksimalverdier (stasjonaritet), deretter er modeller med trend og sesong-effekter prøvd.

4.1. Enkel modell (stasjonaritet)

4.1.1. "Maximum Likelihood" estimering (ML)

Log-likelihooden for m maksimal-verdier kan lett uttrykkes fra (4.3) ved:

(4.5) $$l(\mu,\sigma,\xi) = \sum_{i=1}^{m}\left\{-\log\sigma - (1+\tfrac{1}{\xi})\log\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

Fra [2] har vi at når $\xi > -0.5$ (som vi skal se gjelder her) er alle regularitetsbetingelser for ml-estimatene oppfylt og numerisk maksimering av (4.5) er uproblematisk. Maksimerings-rutinen kan imidlertid få problemer når tallene er for store (gjelder for våre data). En enkel transformasjon er nok til at konvergens av rutinen oppnås. Vi har at hvis $X \sim H_{\xi;\mu,\sigma}(x)$ gjelder at $Y = \frac{X}{k} \sim H_{\xi;\frac{\mu}{k},\frac{\sigma}{k}}(y)$ ([3]), for vilkårlig k. Dermed kan man gjøre nødvendig estimering og analyse på transformerte data, og så transformere tilbake etterpå for å tolke resultatene på opprinnelig skala. Dette er gjort her med k=1000 for enkelhets skyld.

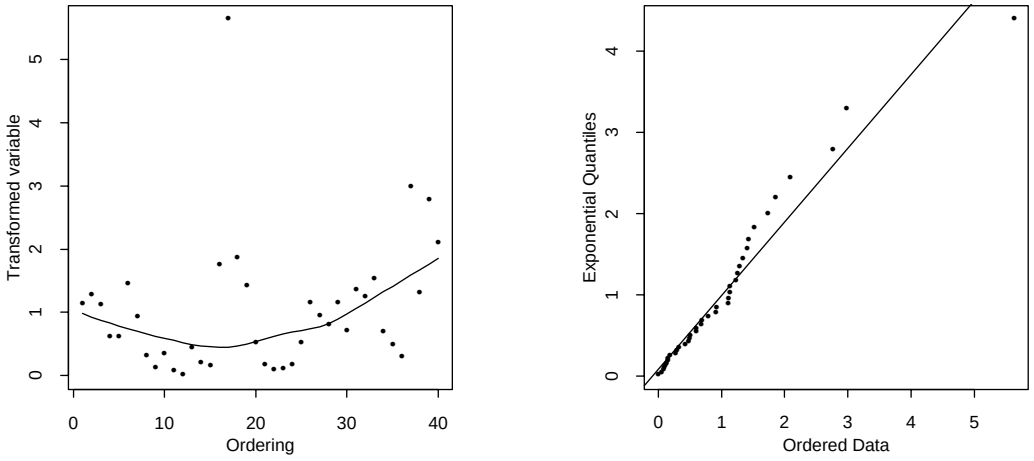
I denne første enkle, stasjonære modellen er alle dataene samlet og (4.5) er maksimert numerisk mhp parametrene. Resultatet av denne maksimeringen og de tilsvarende estimatene på opprinnelig skala er gitt i tabellen nedenfor:

	Transformert skala			Opprinnelig skala		
	$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$
Estimat	0,254	81,095	186,115	0,254	81 094,8	186 115,3
Standard-feil	0,118	11,644	14,274	0,118	11 644	14 274,1

Tabell 2: ML-estimater i enkel GEV modell

Legg merke til at $\hat{\xi}$ er den samme på opprinnelig og på transformert skala. Vi ser at den er signifikant positiv (i den forstand at den er lenger fra 0 enn 2 ganger standard-feilen sin, noe som tilsvarer mindre enn 5 % nivå, uten å ta hensyn til ev. effekt av avhengighet og ved å anta omtrentlig normalitet av ML estimatene). Dette tyder på at halen til dataene er tunge ("Pareto-aktig" hale) og at Frechet fordelingen ($\Phi_\alpha(x)$) er den korrekte grensefordelingen til de normaliserte maksima.

For å undersøke hvor god tilpasningen er, er to diagnostiske plott vist nedenfor. Legg først merke til flg. transformasjon: Når $X \sim h_{\xi;\mu,\sigma}(x)$ (som (4.3)) er $\left[1+\xi\frac{X-\mu}{\sigma}\right]^{-\frac{1}{\xi}} \sim \text{Exp}(1)$, altså standard eksponential fordelt.



Figur 5: Diagnostiske plott for enkel GEV modell

Plottet til venstre i figuren viser et enkelt "scatterplott" av transformerte maksima som funksjon av kvartal. Den heltrukne kurven er en glattet kurve (glatting av punktene i et scatter-plott) og hvis denne hadde vært konstant lik 1 så hadde denne vært i samsvar med at alle kvartalene hadde vært standard eksponentialfordelt (med parameter = 1). Vi kan dermed se hvilke kvartaler som skiller seg ut. Til høyre ser vi et kvantil-plott som også illustrerer overenstemmelsen med standard eksponential fordelingen. Den heltrukne kurven i plottet til høyre er en lineær OLS (minste kvadraters) tilpasning av punktene, og illustrerer hvor godt punktene ligger på en rett linje. (Hvis de transformerte dataene hadde vært eksponential-fordelt med en parameter forskjellig fra 1 ville vi fremdeles fått en rett linje, men med stigning forskjellig fra 1.) Tilpasningen er middels bra.

Tabellen nedenfor viser ml-estmater for noen utvalgte kvantiler, ved bruk av (4.4). De tilsvarende empiriske kvantil-estimatene er 420,87, 476,25, 881,17, 1 078,61, så det er brukbart samsvar inntil data-området er slutt (den største maksimalverdien er 1 101 000).

	Transformert skala				Opprinnelig skala			
p	0,1	0,05	0,01	0,001	0,1	0,05	0,01	0,001
$\hat{q}_p$	432,33	545,8	894,12	1 712,98	432 331	545 801,5	894 121	1 712 976

Tabell 3: ML kvantil estimater

4.1.2. "Probability Weighted Moments" estimering (PWM)

Et alternativ til Maximum Likelihood estimater er en metode som omtales som Probability Weighted Moments ([1] s 321). Den er tilsvarende andre moment-metoder (setter teoretiske momenter lik de tilsvarende empiriske momenter og løser mhp parametrene), men er veiet med kumulative sannsynligheter slik at den er følsom for hvordan halene varierer når observasjonene er store. Definer først:

(4.6) $$w_r = E\left(X \;\; H^r_{\xi;\mu,\sigma}(X)\right),$$

der r er et vilkårlig ikke-negativt heltall. Vi antar $\xi < 1$. Da følger det at $w_0 < \infty$ (og dermed de høyere ordens også). Det empiriske motstykke til (4.5) er:

(4.7) $$\hat{w}_r = \frac{1}{m}\sum_{i=1}^{m} x_i \left\{\hat{H}_{\xi;\mu,\sigma}(x_i)\right\}^r,$$

der $\hat{H}$ er den empirisk kumulative fordelingsfunksjonen utregnet i data-punktene. Det kan vises at når $\xi < 1$ og $\neq 0$ kan w_r uttrykkes som funksjon av parametrene:

(4.8) $$w_r(\xi,\mu,\sigma) = \frac{1}{r+1}\left\{\mu - \frac{\sigma}{\xi}\left(1 - \Gamma(1-\xi)(1+r)^{\xi}\right)\right\}$$

Ved å sette (4.7) lik (4.5) for $r = 0,1,2$ oppnår man et ulineært lignings-sett som kan approksimeres og gi tilnærmede løsninger for hver parameter for seg (se [3]). En slik tilnærmet løsning er på formen:

(4.9) $$\hat{\xi} = -7.859c - 2.9554c^2, \qquad \text{der } c = \frac{2\hat{w}_1 - \hat{w}_0}{3\hat{w}_2 - \hat{w}_0} - \frac{\log 2}{\log 3}$$

(4.10) $$\hat{\sigma} = \frac{(\hat{w}_0 - 2\hat{w}_1)\hat{\xi}}{\Gamma(1-\hat{\xi})(1-2^{\hat{\xi}})}$$

(4.11) $$\hat{\mu} = \hat{w}_0 - \frac{\hat{\sigma}}{\hat{\xi}}\left\{\Gamma(1-\hat{\xi}) - 1\right\}$$

På dataene våre gir dette flg. estimater:

	Transformert skala			Opprinnelig skala		
	$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$
Estimat	0,354	76,345	176,544	0,354	76 345,1	176 544,1

Tabell 4: PWM-estimater i enkel GEV modell

Uttrykk for variansen til disse estimatene er kompliserte og ikke tatt med her. Tilsvarende kvantiler som i tabell 3 er estimert for PWM og vist i tabell 5 nedenfor.

	Transformert skala				Opprinnelig skala			
p	0,1	0,05	0,01	0,001	0,1	0,05	0,01	0,001
$\hat{q}_p$	439,15	577,9	1 059,29	2 445,44	439 154,1	577 898,5	1 059 286	2 445 438

Tabell 5: PWM kvantil estimater

PWM-metoden har sin styrke for små utvalg. Den har vist seg bedre enn ML-estimatene mhp skjevhet og "Mean Square Error" for små utvalg, og ikke så god som ML for større utvalg, avhengig av størrelsen på ξ. Noe av grunnen til at den er bedre enn ML kan være at man forutsetter $\xi < 1$. Dette er en antagelse som kan inngå i ML-estimeringen også for å gjøre denne bedre (ikke gjort her). For våre data som har en utvalgsstørrelse på $m = 40$ (ikke veldig liten) burde vi få brukbart sammenlignbare resultater, noe som vi kan se vi gjør ved å sammenligne tabell 2 og tabell 4. Vi ser f.eks. at PWM estimatene er godt innenfor en avstand lik to ganger standard-feilen til ML-estimatene noe man kan tolke som en kontroll av ML-estimatene. En faktor som påvirker PWM estimatene er valget av empirisk kumulativ fordelingsfunksjon. Her brukes $\hat{F}_n(x) = \frac{1}{n}\sum I(X_i \le x)$.

4.2. Ikke stasjonaritet

Måten å modellere ikke-stasjonaritet på her er å la parametrene i (4.2) variere med tiden etter enkle kjente funksjoner, som f.eks. trend og sesong funksjoner. Forandringer i lokasjons-parameteren μ tilsvarer at det underliggende nivået på ekstrem verdi atferden varierer og er det mest iøyenfallende når man betrakter dataene. Forandringer i σ tolkes som forandringer i variabiliteten i de ekstreme verdiene. Forandringer i ξ er vanskeligere å modellere. Vi har prøvd de to førstnevnte formene for ikke-stasjonaritet i GEV modellen.

4.2.1. Trend i GEV modellen

For å illustrere en lineær trend i lokasjonsparameteren har vi altså flg. modell μ:

(4.11) $$\mu_k = \mu_0 + \alpha k, \quad k = 1, \cdots, 40$$

der k er kvartal. Estimering av parametrene i modellen gjøres her ved å maksimere likelihooden i (4.5) som før med μ_k i stedet for μ (og k i stedet for i). Da får man en parameter til å maksimere med hensyn på, α. Resultatet av denne tilpasningen er vist i tabellen nedenfor.

	$\hat{\mu}_0$	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\xi}$
Estimat	208,63	-1	80,63	0,238
Standard-feil	24,98	0,94	11,37	0,111

Tabell 6: ML estimater i GEV trend modell

Parameterverdiene i tabell 6 er på transformert skala, men ved å multiplisere med k (alle parametrene untatt ξ) fås opprinnelig skala. Som vi ser er ikke $\hat{\alpha}$ signifikant forskjellig fra 0, som viser at dataene ikke tyder på lineær trend i lokasjonsparameteren. Tallene er på transformert skala. Andre trend modeller er også forsøkt (kvadratisk ledd, og med tredje grads ledd), men med ennå mindre signifikante koeffisienter. Vi har gjort tilsvarende modellering med σ, også log-lineær modell er prøvd her, $\log(\sigma_k) = \alpha + \beta k$ (for å sikre positiv skala-parameter), men ingen av modellene for σ gir gode tilpasninger.

4.2.2. Sesong i GEV modellen

Som figur 1 viser er det antydning til sesong-effekt i de største investeringene, dvs. vi ser at for flere av årene har fjerde kvartal større investeringer enn de tre første kvartalene.

For å undersøke hvor tydelig denne sesong-variasjonen er, prøver vi modeller med sesong variasjon i parametrene. Dette blir tilsvarende som en trend, men med parametre som har periodisk variasjon i tid, og ett år (fire kvartaler) som periodetid.

Sesongvariasjon i lokasjonsparameteren lar seg formulere ved modellen:

$$\mu_{k,j} = \mu_0 + \beta_1 s_{1,k,j} + \beta_2 s_{2,k,j} + \beta_3 s_{3,k,j} \tag{4.12}$$

der $k = 1,\cdots,40$ er kvartal og $j = 1,\cdots,10$ er år. Videre er $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ kvartalsdummyer, som er vektorer med dimensjon [40×1] og med egenskapene $\sum_k s_{1,k,j} = \sum_k s_{2,k,j} = \sum_k s_{3,k,j} = 0$ og $s_{1,k,j_1} = s_{1,k,j_2}$, $s_{2,k,j_1} = s_{2,k,j_2}$, $s_{3,k,j_1} = s_{3,k,j_2}$ med $j_1 \neq j_2$. F.eks. er $\mathbf{s}_1' = (\frac{3}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad \frac{3}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \quad -\frac{1}{4} \cdots)'$ med $\frac{3}{4}$ på alle førstekvartals-posisjonene slik at denne veies opp. Helt tilsvarende har $\mathbf{s}_2$ $\frac{3}{4}$ på alle andrekvartaler og $\mathbf{s}_3$ har $\frac{3}{4}$ på alle tredjekvartaler.

Den mer intuitive modellen med fire kvartalsdummyer,

$$\mu^*_{k,j} = \gamma_1 s_{1,k,j} + \gamma_2 s_{2,k,j} + \gamma_3 s_{3,k,j} + \gamma_4 s_{4,k,j} \tag{4.13}$$

altså med $\mathbf{s}_4$ i stedet for μ_0, som er normalisert rundt 0 (i stedet for μ_0) og derfor beskriver kun den rent periodiske delen, kan oppnås som en lineær-transformasjon av (4.12). Den har singulær designmatrise siden $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_4$ er lineært avhengige ($-\mathbf{s}_1 - \mathbf{s}_2 - \mathbf{s}_3 = \mathbf{s}_4$) og er derfor ikke entydig. En av mange mulige $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)'$ er flg.:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix} = \begin{pmatrix} 1 & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 1 & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ 1 & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\ 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \tag{4.14}$$

Dette er samme transformasjonen som i (4.12), altså:

$$\begin{pmatrix} \mu_{1,j} \\ \mu_{2,j} \\ \mu_{3,j} \\ \mu_{4,j} \end{pmatrix} = \begin{pmatrix} 1 & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ 1 & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ 1 & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \\ 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} \mu_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \tag{4.15}$$

Dermed, hvis (4.14) er oppfylt må $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)' = (\mu_{1,j}, \mu_{2,j}, \mu_{3,j}, \mu_{4,j})'$ og ved innsetting i (4.13) fås $\mu^*_{1,j} = \cdots = \mu_{1,j} - \mu_0, \cdots, \mu^*_{4,j} = \mu_{4,j} - \mu_0$, altså at modellene (4.13) og (4.12) er like. I (4.13) kan $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)'$ tolkes direkte som de fire sesong-nivåene.

Resultatet av en sesong modell tilpasning for lokasjons-parameteren er vist nedenfor:

$$\begin{pmatrix}\hat{\mu}_{1,j}\\ \hat{\mu}_{2,j}\\ \hat{\mu}_{3,j}\\ \hat{\mu}_{4,j}\end{pmatrix}=\begin{pmatrix}144,92\\ 197,12\\ 182,38\\ 194,32\end{pmatrix},\quad \mathrm{COV}_{\hat{\mu}_{k,j}}=\begin{pmatrix}477,35 & . & . & .\\ 90,31 & 316,51 & . & .\\ 170,78 & 75,16 & 389,87 & .\\ 159,19 & 72,51 & 122,71 & 385,53\end{pmatrix},$$

$$\begin{pmatrix}\hat{\sigma}\\ \hat{\xi}\end{pmatrix}=\begin{pmatrix}70,06\\ 0,4276\end{pmatrix},\quad \begin{pmatrix}\mathrm{St.feil}(\hat{\sigma})\\ \mathrm{St.feil}(\hat{\xi})\end{pmatrix}=\begin{pmatrix}12,24\\ 0,190\end{pmatrix}$$

Tabell 7: ML-estimater i GEV sesong modell (lokasjon) fra (4.12)

Med normalfordelings egenskapene for ML-estimater kan man her teste hypoteser for å sammenligne kvartals-nivåene for lokasjon. De seks mulige sammenligningene (første mot andre, første mot tredje osv.) gir ingen signifikante forskjeller med signifikansnivå på 0,417 % som er 5 % nivå justert for to-sidig testing og for at det er seks forskjellige delhypoteser. Det er imidlertid en indikasjon på at første kvartal har lavere nivå enn fjerde kvartal (p=0,017). Det er verdt å merke seg at parameter-estimatene er nokså like de under stasjonaritet. $\hat{\xi}$ er større nå, men har også større usikkerhet.

Det er også mulig å konstruere kvartals-dummyer som spesifiserer hypoteser. F.eks. ser det fra figur 1 ut som om fjerde kvartal skiller seg ut fra de andre med høyere lokasjons-parameter. En sammenligning mellom fjerde kvartal og de tre første kan gjøres ved modellen:

(4.16) $$\mu_{k,j}=\mu_0+\beta s_{123,k,j}$$

der $\mathbf{s}_{123}=(\frac{1}{4},\frac{1}{4},\frac{1}{4},-\frac{3}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4},-\frac{3}{4},\cdots)'$. Resultatet av en slik tilpasning er gitt nedenfor:

$$\begin{pmatrix}\hat{\mu}_{1,j}\\ \hat{\mu}_{2,j}\\ \hat{\mu}_{3,j}\\ \hat{\mu}_{4,j}\end{pmatrix}=\begin{pmatrix}179,71\\ 179,71\\ 179,71\\ 204,17\end{pmatrix},\quad \mathrm{COV}_{\hat{\mu}_{k,j}}=\begin{pmatrix}243,84 & . & . & .\\ 243,84 & 243,84 & . & .\\ 243,84 & 243,84 & 243,84 & .\\ 77,07 & 77,07 & 77,07 & 522,52\end{pmatrix},$$

$$\begin{pmatrix}\hat{\sigma}\\ \hat{\xi}\end{pmatrix}=\begin{pmatrix}79,49\\ 0,2714\end{pmatrix},\quad \begin{pmatrix}\mathrm{St.feil}(\hat{\sigma})\\ \mathrm{St.feil}(\hat{\xi})\end{pmatrix}=\begin{pmatrix}11,64\\ 0,125\end{pmatrix}$$

Tabell 8: ML-estimater i GEV sesong modell (lokasjon) fra (4.16)

Det er ingen signfikant forskjell på fjerde kvartal og de tre første kvartalene (p=0,16). ξ estimeres til å være lavere i denne modellen enn forrige.

En sesongmodell for skala-parameteren tilsvarende (4.12) kan skrives:

(4.17) $$\sigma_{k,j} = \sigma_0 + \beta_1 s_{1,k,j} + \beta_2 s_{2,k,j} + \beta_3 s_{3,k,j}$$

og en tilpasning til dataene gir flg. resultat:

$$\begin{pmatrix} \hat{\sigma}_{1,j} \\ \hat{\sigma}_{2,j} \\ \hat{\sigma}_{3,j} \\ \hat{\sigma}_{4,j} \end{pmatrix} = \begin{pmatrix} 104,72 \\ 45,25 \\ 71,13 \\ 65,11 \end{pmatrix}, \quad \mathrm{COV}_{\hat{\sigma}_{k,j}} = \begin{pmatrix} 472,9 & . & . & . \\ 70,63 & 197,24 & . & . \\ 83,87 & 76,37 & 312,56 & . \\ 16,82 & 69,87 & 48,18 & 415,62 \end{pmatrix},$$

$$\begin{pmatrix} \hat{\mu} \\ \hat{\xi} \end{pmatrix} = \begin{pmatrix} 181,84 \\ 0,459 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{St.feil}(\hat{\mu}) \\ \mathrm{St.feil}(\hat{\xi}) \end{pmatrix} = \begin{pmatrix} 11,79 \\ 0,19 \end{pmatrix}$$

Tabell 9: ML-estimater i GEV sesong (skala) modell fra (4.17)

Som den eneste markante forskjellen og med 0,417 % signifikansnivå har første kvartal nesten signifikant høyere skala-nivå enn andre kvartal (p=0,00485). Dette kan virke rart når man betrakter dataene, at første kvartal har større spredning enn andre kvartal. Det kan forklares ved at under beregning av $\hat{\sigma}$ holdes lokasjonsparameteren konstant , $\hat{\mu} = 181,84$. Denne verdien er lengre fra det estimerte lokasjonsnivået i første kvartal fra modell (4.12), enn det estimerte lokasjonsnivået i andre kvartal. Så i forhold til det nye estimerte lokasjonsnivået er observasjonene i første kvartal mer spredt. I tillegg må man huske på at med positiv ξ er fordelingen blant de med veldig tung hale og variansen er dermed veldig stor, hvis den i det hele tatt eksisterer. Dermed vil usikkerheten i estimeringen være stor.

Med en modell som sammenligner fjerde kvartals skala-nivå mot resten, tilsvarende som i (4.16) og spesifisert ved:

(4.18) $$\sigma_{k,j} = \sigma_0 + \beta s_{123,k,j}$$

fås flg. resultat:

$$\begin{pmatrix} \hat{\sigma}_{1,j} \\ \hat{\sigma}_{2,j} \\ \hat{\sigma}_{3,j} \\ \hat{\sigma}_{4,j} \end{pmatrix} = \begin{pmatrix} 82,86 \\ 82,86 \\ 82,86 \\ 73,63 \end{pmatrix}, \quad \mathrm{COV}_{\hat{\sigma}_{k,j}} = \begin{pmatrix} 177,14 & . & . & . \\ 177,14 & 177,14 & . & . \\ 177,14 & 177,14 & 177,14 & . \\ -1,5 & -1,5 & -1,5 & 619,74 \end{pmatrix},$$

$$\begin{pmatrix} \hat{\mu} \\ \hat{\xi} \end{pmatrix} = \begin{pmatrix} 186,48 \\ 0,272 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{St.feil}(\hat{\mu}) \\ \mathrm{St.feil}(\hat{\xi}) \end{pmatrix} = \begin{pmatrix} 14,18 \\ 0,134 \end{pmatrix}$$

Tabell 10: ML-estimater i GEV sesong (skala) modell fra (4.18)

Her er det ikke signifikant forskjell i skala-parameteren mellom fjerde kvartal og de tre første kvartalene (p=0,37).

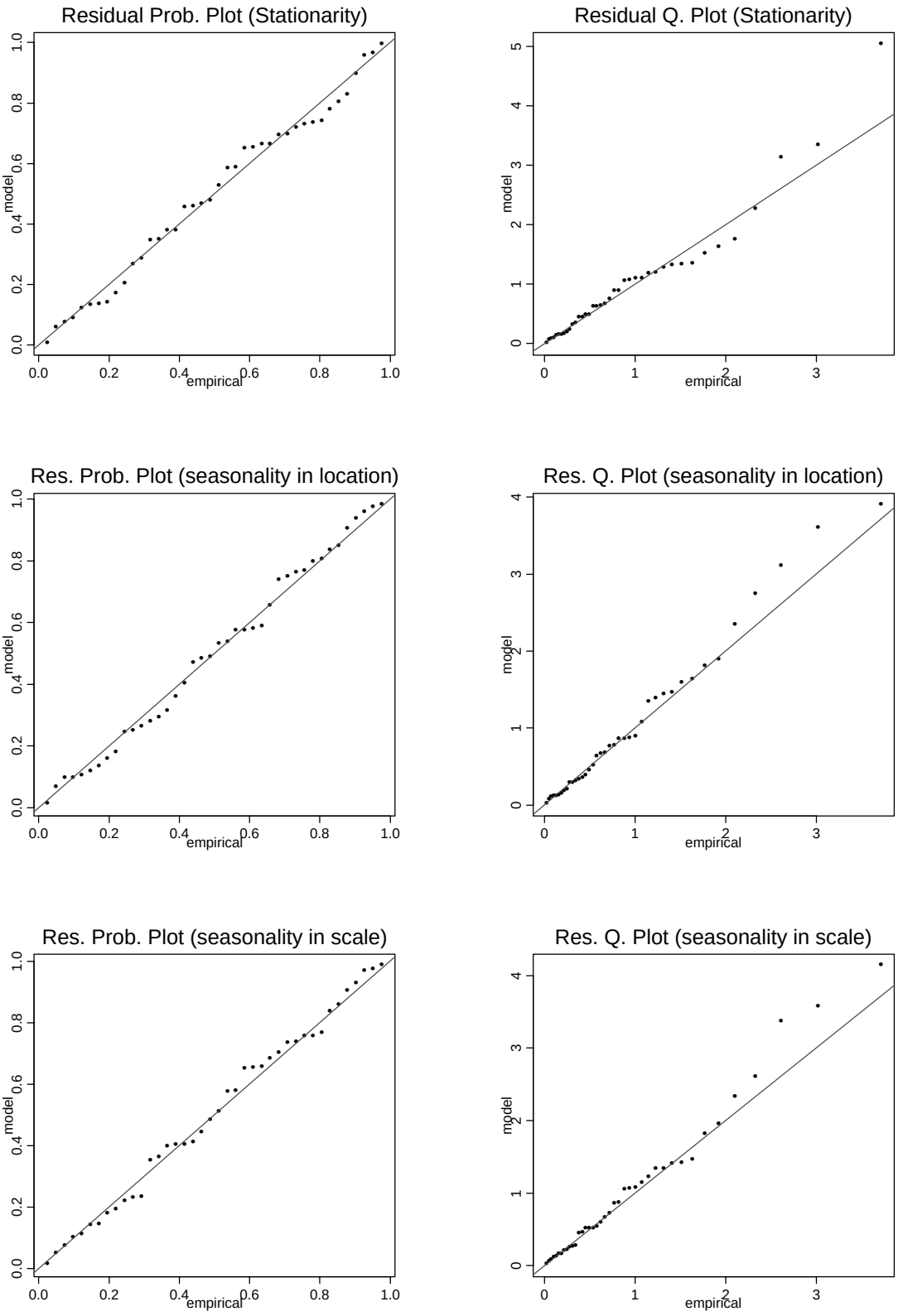
Figuren på neste side viser noen diagnostiske plott for residualene (observasjonene minus den estimerte sesong-effekten) i sesong modellene og for den stasjonære modellen for sammenligning. De heltrukne linjene er rette linjer med stigning=1 for å markere perfekt samsvar mellom observasjoner og modell. Vi ser at sesong modellen for lokasjons parameteren gir noe bedre tilpasning til dataene enn modellen uten sesong. Vi ser videre at det er liten forskjell i tilpasning mellom sesong modellen for lokasjon og for skala parameter, men velger sesong modellen for skala som den beste siden den har lavest p-verdi i sammenligning av kvartalene.

Vi har prøvd modeller med både sesong og trend, men dette gir ikke bedre resultat. Man kan også tenke seg sesong for flere parametere, men dette har vi ikke gjort. Med bare 40 observasjoner må man begrense antall parametere i modellen.

Noen kvantil estimater for den beste sesong modellen (skala) er vist i tabell 11 nedenfor. Vi ser at kvantilene nå varierer fra kvartal til kvartal og de fleste er større enn kvantil estimatene uten sesong fra tabell 3, spesielt utenfor data-området, noe som viser at usikkerheten i kvantilene er stor her (største maksimalverdi er 1 101).

	Transformert skala				Opprinnelig skala			
	Kv.1	Kv.2	Kv.3	Kv.4	Kv.1	Kv.2	Kv.3	Kv.4
$\hat{q}_{0.1}$	594,7	360,2	462,2	438,5	594 667	360 206	462 236	438 508,6
$\hat{q}_{0.05}$	845,7	468,6	632,7	594,6	845 656,8	468 647,9	632 710,5	594 557,1
$\hat{q}_{0.01}$	1 838,7	897,7	1 307,2	1 211,9	1 838 667	897 684,3	1 307 171	1 211 943
$\hat{q}_{0.001}$	5 389,6	2 431,9	3 719	3 419,7	5 389 598	2 431 886	3 718 991	3 419 670

Tabell 11: ML kvantil estimater i GEV sesong modell (sesong i skala)



Figur 6: Diagnostiske residual plott for GEV modell uten (øverst) og med (fire nederste) sesong

5. "Peak Over Threshold" Metode (GPD modell)

Denne metoden er nært knyttet til den forrige, men utnytter flere observasjoner (GEV modellen brukte bare 1 observasjon pr. blokk, maksimalverdien). Forbindelsen mellom disse metodene er lett å se med utgangspunkt i flg. situasjon. Vi tenker oss at vi har:

(5.1) $X_1, \cdots, X_n$ u. i. f. med ukjent fordeling $F \in MDA\ (H_\xi)$,

en forkortelse for at F er med i "Maximum Domain of Attraction" til H_ξ som er fordelingen i (4.2) (normalisert til bare å ha en parameter). Dette passer på investeringsdataene våre og er nettopp den situasjonen vi hadde i kapittel 4. Dette betyr at, maksimal verdiene til blokker av investeringer (passe normalisert) har H_ξ som grensefordeling. Her velger vi i stedet for å se på maksimal verdiene, å se på overskridelser av en høy grenseverdi u slik at når X er større enn u kaller vi denne overskridelsen for Y, $Y = X - u$. Vi kaller antall overskridelser for N_u, altså $N_u = \left|\{i : X_i > u\}\right|$. Fordelingen til disse overskridelsene ("eksessene") defineres som en betinget fordeling:

(5.2) $$F_u(y) = \Pr\{X - u \le y \mid X > u\} = \Pr\{Y \le y \mid X > u\}, \quad y \ge 0$$

som kan omformes til:

(5.3) $$\bar{F}(u + y) = \bar{F}(u)\ \ \bar{F}_u(y),$$

der $\bar{F}(.) = 1 - F(.)$ angir halesannsynligheten. (5.3) sier at halesannsynligheten i fordelingen til investeringene lengre ute enn en stor verdi u er gitt ved produktet av halesannsynligheten i u og halesannsynligheten i eksess fordelingen i y. For stor nok u gjelder flg. tilnærming:

(5.4) $$\bar{F}_u(y) \approx \bar{G}_{\xi,\beta(u)}(y), \quad \text{der}$$

(5.5) $$G_{\xi,\beta(u)}(y) = 1 - (1 + \xi \tfrac{y}{\beta(u)})^{-\frac{1}{\xi}}, \text{ for } \xi > 0 \text{ (som er vårt tilfellet)}$$

er den Generaliserte Pareto Fordelingen (GPD). GPD har en parameter mindre enn GEV fordelingen (har "mistet" lokasjons-parameteren) pga. at den betinger på u. Det er imidlertid samme ξ verdien, selv om skala parameteren er endret som følge av reparametriseringen. Det er mulig å lage en empirisk versjon av (5.4) (bruker vanlig empirisk estimat for halesannsynligheten $\hat{\bar{F}}(u) = \frac{N_u}{n}$) slik at

(5.6) $$\hat{\bar{F}}(u + y) = \tfrac{N_u}{n}(1 + \hat{\xi}\tfrac{y}{\hat{\beta}})^{-\frac{1}{\hat{\xi}}}$$

gir et estimat for halesannsynligheten til X_i (langt ute). Dette gir direkte et estimat for kvantilen i F som:

(5.7) $$\hat{q}_p = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\tfrac{n}{N_u}p\right)^{-\hat{\xi}} - 1\right)$$

For å finne ut hvor stor u må være er det nyttig med "mean excess" funksjonen som for GPD kan vises å være lik:

(5.8) $$e(u) = E(X - u \mid X > u) = \frac{\beta + \xi u}{1 - \xi} \ , \ \xi < 1$$

så $e(u)$ er lineær i u. Denne kan estimeres som:

(5.9) $$\hat{e}(u) = \frac{1}{N_u} \sum_{i:X_i > u} (X_i - u)$$

Teoretisk er det slik at hvis tilpasningen til GPD er god for en bestemt stor nok u skal den også være god for alle verdier av u større enn det. Derfor kan u velges stor og slik at (5.9) som funksjon av u er mest mulig lineær ovenfor. Hvis u velges for stor har man god GPD tilpasning, men få eksesser og dermed dårlig parameter estimering og hale estimering. Dette er en avveining som ligner på blokk størrelses problemet i kapittel 4.

Vi bruker her ML estimering for parametrene, men pga. at Y_i og N_u er avhengige trenger man en betinget likelihood, eller (som vi gjør her) bruke en punktprosess-likelihood, se neste avsnitt.

For ML-estimatene gjelder følgende approksimative fordelingsresultat (for n endelig) :

(5.10) $$\begin{pmatrix} \hat{\xi} \\ \hat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \xi \\ \beta \end{pmatrix}, \frac{1+\xi}{n} \begin{bmatrix} 1+\xi & \beta \\ \beta & 2\beta^2 \end{bmatrix} \right)$$

som gir konfidensintervaller for parametrene osv. F.eks. er et approksimativt 95 % konfidens-intervall for ξ gitt ved

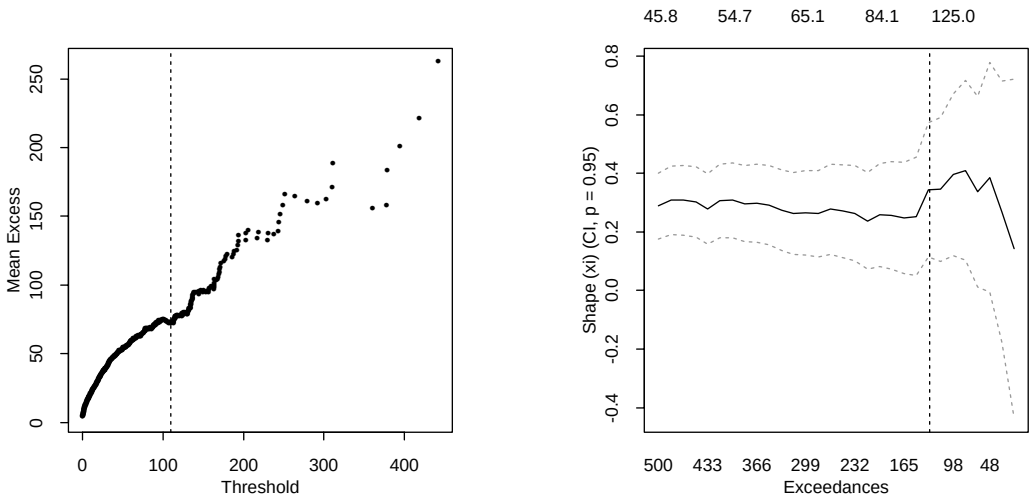
(5.11) $$\xi \in \left(\frac{\hat{\xi} - \frac{1.96}{\sqrt{n}}}{1 + \frac{1.96}{\sqrt{n}}} \ , \ \frac{\hat{\xi} + \frac{1.96}{\sqrt{n}}}{1 - \frac{1.96}{\sqrt{n}}} \right),$$

og for β:

(5.12) $$\beta \in \left(\frac{\hat{\beta}}{1 + 1{,}96\sqrt{2\frac{1+\xi}{n}}} \ , \ \frac{\hat{\beta}}{1 - 1{,}96\sqrt{2\frac{1+\xi}{n}}} \right)$$

5.1. Enkel modell (stasjonaritet)

Som i kapittel 4 prøver vi først metoden på alle investeringene samlet. For å velge riktig u ser vi først på mean excess plottet. Nedenfor er dette vist til venstre i figuren. Det er samme plottet som i figur 3, bare på transformert skala. Den heltrukne linjen er vårt initielle valg av u, nemlig $u = 110$. Ved siden av er det vist et såkalt "horror plot" som illustrerer ML-estimatet for ξ for forskjellige valg av u med u langs øvre aksen og antall overskridelser langs nedre aksen (horror fordi det ofte viser at ML-estimatet forandrer seg mye som funksjon av en mer eller mindre tilfeldig valgt u , en opplagt dårlig egenskap). Horror plottet viser også 95 % konfidensintervall for ξ, og vi ser at 0 ligger utenfor dette, untatt for de ekstreme verdiene for u, noe som er nok en bekreftelse på den tunge halen.



Figur 7: Mean Excess (transformert skala) og Horror plot, med inntegnet "threshold" (u=110)

Med $u = 110$ får vi 130 eksesser (investeringer som er over 110) og altså flere observasjoner å gjøre inferens på i forhold til GEV modellen. Tabellen nedenfor viser ML-estimater.

	Transformert skala		Opprinnelig skala	
	$\hat{\xi}$	$\hat{\beta}$	$\hat{\xi}$	$\hat{\beta}$
Estimat	0,359	46,256	0,359	46 255,7
Standard-feil	0,118	6,657	0,118	6 657,1
Konf.intervall (95%)	0,1598_0,6414	36,04_64,55	0,1598_0,6414	36 040,7_64 551,5

Tabell 12: ML-estimering i enkel GPD modell

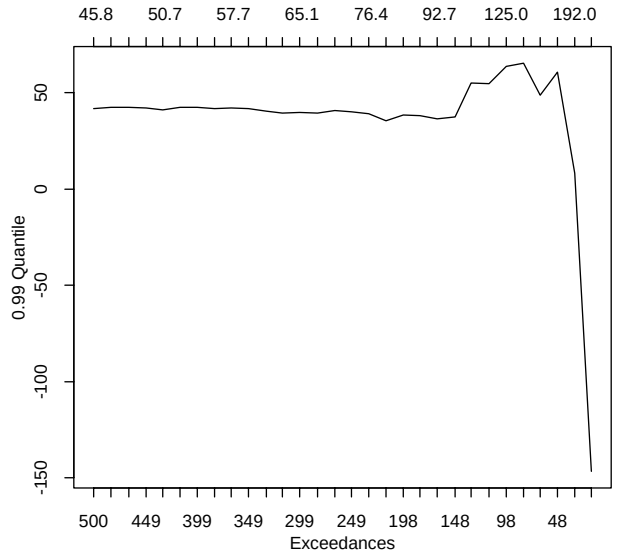
Vi legger merke til bedre estimering (lavere standard feil) av ξ enn i GEV modellen. Konfidens-intervallene er utregnet ved (5.11) og (5.12).

Tabellen nedenfor viser utvalgte estimerte kvantiler i halen til fordelingen til investeringene selv, på transformert og opprinnelig skala (utregnet ved (5.7)). Hvis vi sammenligner med kvantilene for GEV modellen ser vi at GEV modellens kvantiler er atskillig større (GEV modellen er fordelingen til maksimalverdiene).

	Transformert skala				Opprinnelig skala			
p	0,1	0,05	0,01	0,001	0,1	0,05	0,01	0,001
$\hat{q}_p$	14,53	23,95	57,43	155,49	14 525	23 953,8	57 427,8	155 486,3

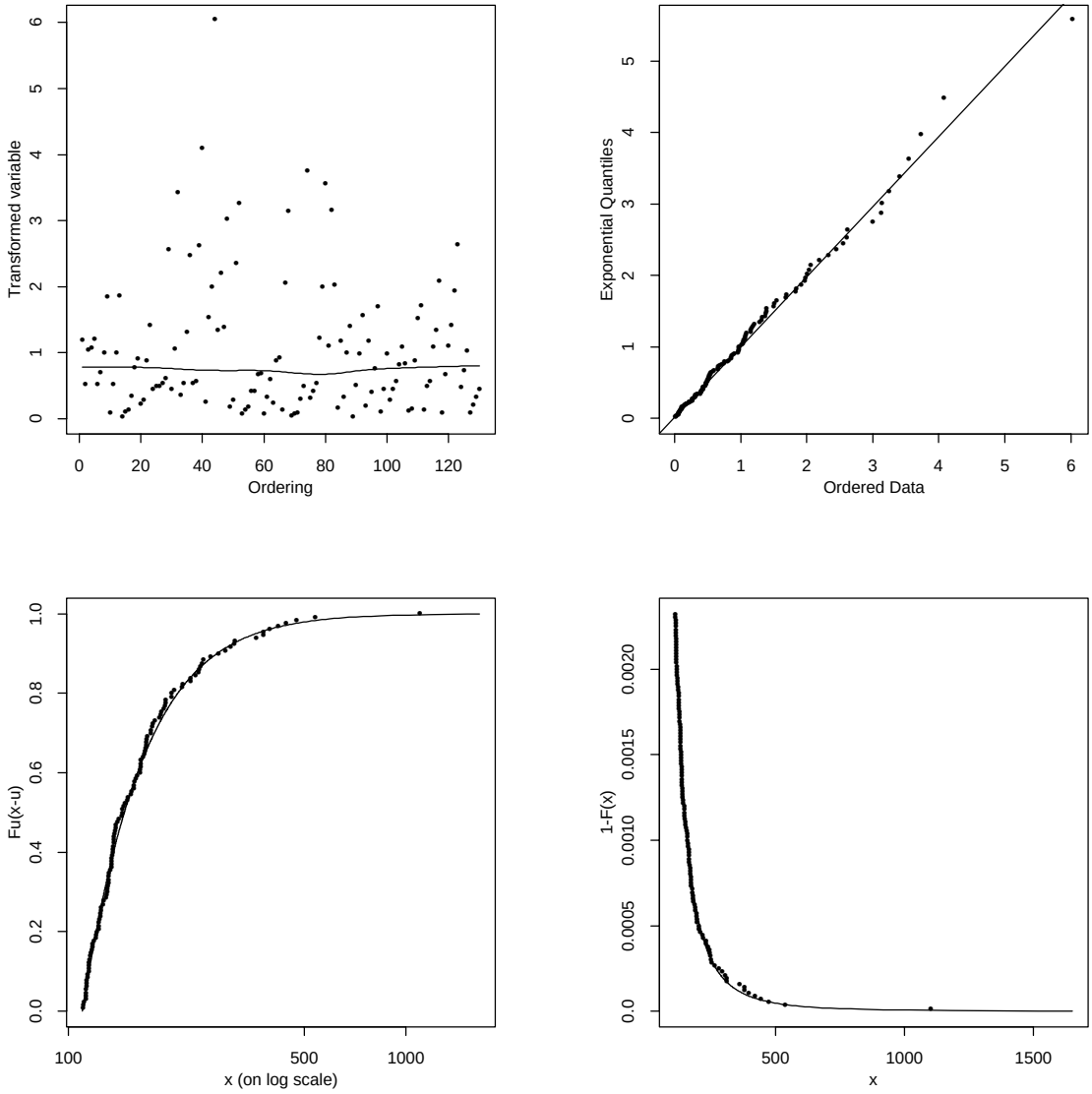
Tabell 13: Kvantil estimater i investerings-fordelingen

Figuren nedenfor viser 99 % kvantilen til investerings-fordelingen for forskjellige valg av u (øvre horisontale akse) og antall eksesser (nedre horisontale akse). Som vi ser fluktuerer kvantilestimatet voldsomt for stor u. Det skyldes at hale estimatet i (5.6) er beregnet for positiv *y*, og det betyr at den fungerer først for verdier som er større enn u. Når man da prøver å estimere noe som er lavere enn u fungerer ikke estimatet. For vårt valg (110) ser vi at plottet stemmer med tabellen ovenfor ($p = 0,01$).



Figur 8: 99 % kvantilen i investeringsfordelingen

Neste side følger noen diagnostiske plott for denne tilpasningen, tilsvarende figur 5 for GEV modellen. Legg først merke til at når $Y \sim G_{\xi,\beta}(y)$ som i (5.5) er $\frac{1}{\beta}\log\left(1+\xi\frac{Y}{\beta}\right) \sim \text{Exp}(1)$ altså standard eksponential fordelt (tilsvarende transformasjon som i figur 5). De to øverste plottene er tilsvarende figur 5 og beskriver hvor godt eksessene er GPD fordelt. Den viser bedre tilpasning enn i figur 5. Nederst til venstre viser figuren også hvor godt tilpasset eksessene er til den estimerte GPD fordelingen, utransformert (og log skala på x-aksen). Nederst til høyre ser vi haletilpasningen i investeringsfordelingen selv, estimert som i (5.6).



Figur 9: Diagnostiske plott for GPD tilpasning

5.2. Ikke stasjonaritet

For å modellere ikke-stasjonaritet i denne modellen har vi benyttet en tredje parametrisering. Den tar utgangspunkt i en punkt prosess karakterisering som er den mest generelle karakteriseringen og som omfatter de andre som spesialtilfeller ([2]).

Vi tenker som i (5.1) at vi har

(5.13) $X_1, \cdots, X_n$ u. i. f. med ukjent fordeling $F \in MDA(H_\xi)$,

Videre konstruerer vi en punktprosess i R^2

(5.13) $$P_n = \left\{\left(\frac{i}{n+1}, X_i\right) : i = 1, \cdots, n\right\}$$

Denne har intensitets-funksjon

$$\Lambda\{(t_1,t_2)\times(x,\infty)\} = n_y(t_2-t_1)\left[1+\xi\tfrac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}, \tag{5.14}$$

der n_y er tilsvarende antall blokker som i kapittel 4 og resten av parametrene er som i GEV modellen (og samme ξ som i GPD). Dette betyr at med n observasjoner totalt over en periode på n_y blokker (i tid) og et område av formen $A_v = [0,1]\times(v,\infty)$, for $v > u$ hvor u er vår høye grenseverdi, får vi flg. punktprosess likelihood:

$$L(A_v;\mu,\sigma,\xi) = \exp\left\{-n_y\left(1+\xi\tfrac{v-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\}\prod_{i=1}^{N_{A_v}}\tfrac{1}{\sigma}\left(1+\xi\tfrac{x_i-\mu}{\sigma}\right)^{-\frac{1}{\xi}-1}, \tag{5.15}$$

der $x_1, \cdots, x_{N_{A_v}}$ er de N_{A_v} observasjonene som er over grensen v.

For å modellere ikke-stasjonaritet i GPD modellen bruker vi tidsvarierende funksjoner for parametrene i (5.15). For å anvende dette på investeringsdataene våre lar vi observasjonene betegnes med:

$$X_{i,k,j}\left(I_{i,k,j}\right), \quad \text{der} \quad \begin{array}{l} i = 1, \cdots, 1\,400 \text{ er kvartalsinvestering} \\ k = 1,2,3,4 \text{ er kvartal} \\ j = 1, \cdots, 10 \text{ er år} \end{array} \tag{5.16}$$

5.2.1. Sesongvariasjon

Vi oppnår en noe bedre tilpasning ved å la modellen ta høyde for ikke-stasjonæritet i form av sesongvariasjon over året. Ved å analysere alle første kvartalene for seg, deretter alle andre kvartalene osv finner vi først de fire forskjellige kvartalsvise optimale grenseverdiene (u) og deretter formulerer en sesong-modell og maksimerer likelihooden som i forrige avsnitt, for estimering. Figuren nedenfor viser Mean Excess plot for hvert kvartal og inntegnede valgte grenseverdier. Disse grenseverdiene, $u_k, k = 1,\cdots,4$ med antall eksesser i parantes er gitt i tabellen under.

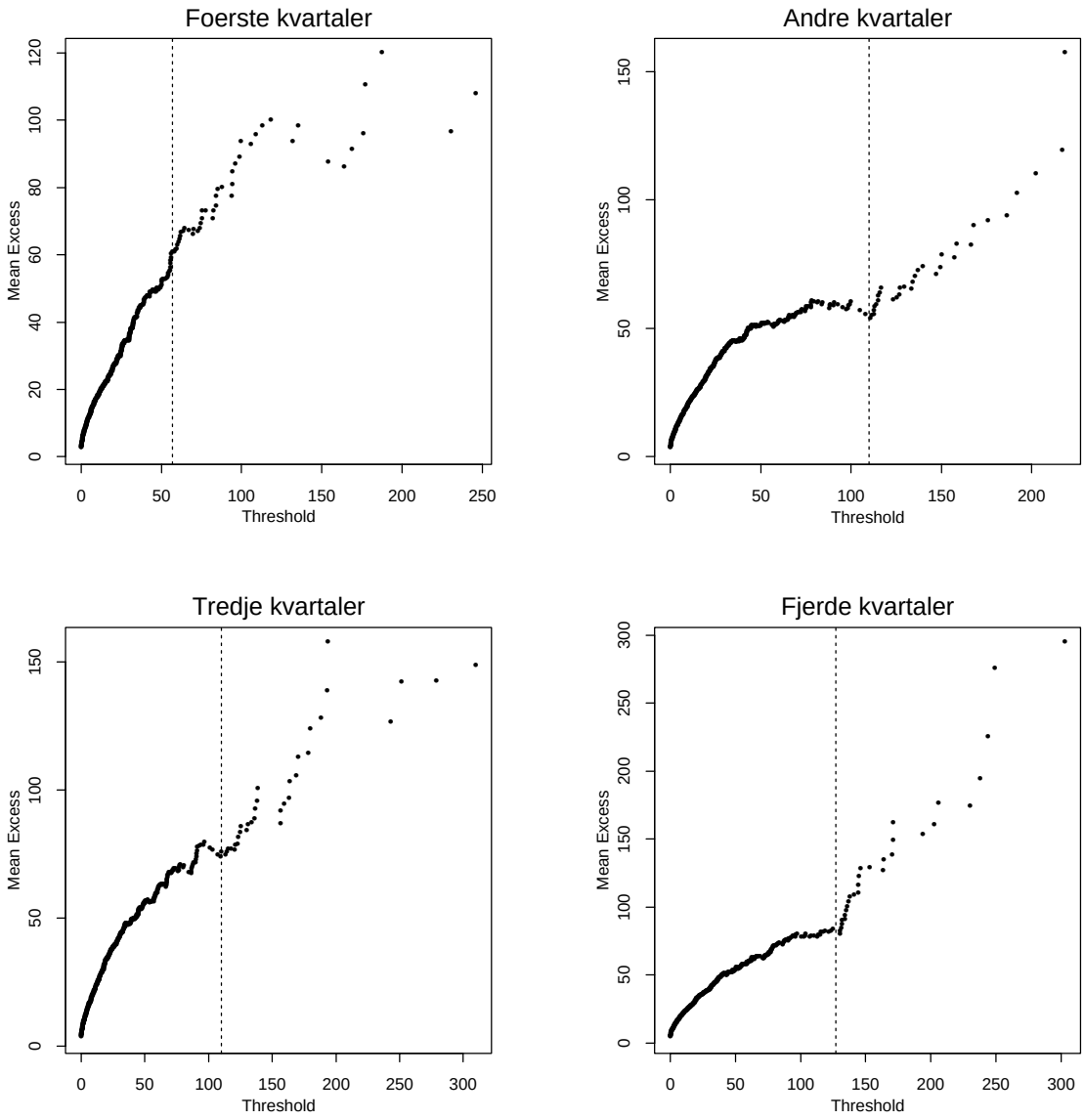
u_1 (n_1)	u_2 (n_2)	u_3 (n_3)	u_4 (n_4)
57 (48)	110 (36)	110 (36)	127 (33)

Tabell 14: Varierende u_k

Tilsvarende som i GEV sesong modellen lar sesongvariasjon i lokasjonsparameteren seg formulere ved modellen:

$$\mu_{k,j} = \mu_0 + \beta_1 s_{1,k,j} + \beta_2 s_{2,k,j} + \beta_3 s_{3,k,j} \tag{5.17}$$

hvor indeks "i" er droppet for å markere at for hver kombinasjon av "k" og "j" er lokasjonsparameteren den samme for alle 1 400 observasjonene. Tilsvarende som i kap. 4 er $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ kvartalsdummyer, som nå er vektorer med dimensjon [56 000×1] siden for hvert kvartal er 1 400 verdier like og med de samme egenskapene som i avsnitt 4.2.2.



Figur 10: Mean Excess plot for kvartalene hver for seg

Den totale likelihooden kan skrives som:

$$(5.18) \qquad L(A_{v_1}, A_{v_2}, A_{v_3}, A_{v_4}; \mu_{k,j}, \sigma, \xi) = \prod_{i,k,j}^{N} \exp\left\{-n_y \left(1 + \xi \frac{v_k - \mu_{k,j}}{\sigma}\right)^{-\frac{1}{\xi}}\right\} \frac{1}{\sigma}\left(1 + \xi \frac{x_{i,k,j} - \mu_{k,j}}{\sigma}\right)^{-\frac{1}{\xi}-1}$$

der $N = 153$ som er totalt antall observasjoner som er over grensene i tabell 14 og $n_y = 40$ som er antall kvartaler.

Resultatet av en sesong modell tilpasning for lokasjons-parameteren er vist i tabellen nedenfor.

$$\begin{pmatrix}\hat{\mu}_{1,j}\\ \hat{\mu}_{2,j}\\ \hat{\mu}_{3,j}\\ \hat{\mu}_{4,j}\end{pmatrix}=\begin{pmatrix}261,22\\ 315,56\\ 312,05\\ 331,24\end{pmatrix},\quad \mathrm{COV}_{\hat{\mu}_{k,j}}=\begin{pmatrix}1140,37 & . & . & .\\ 1115,64 & 1155,57 & . & .\\ 1105,14 & 1109 & 1140,31 & .\\ 1114,49 & 1118,39 & 1107,85 & 1159,05\end{pmatrix},$$

$$\begin{pmatrix}\hat{\sigma}\\ \hat{\xi}\end{pmatrix}=\begin{pmatrix}127,7\\ 0,4345\end{pmatrix},\quad \begin{pmatrix}\mathrm{St.feil}(\hat{\sigma})\\ \mathrm{St.feil}(\hat{\xi})\end{pmatrix}=\begin{pmatrix}34,4\\ 0,121\end{pmatrix}$$

Tabell 15: ML-estimater i sesong (lokasjon) punktprosess modell fra (5.17)

Som i avsnitt (4.2.2) kan man her teste hypoteser for å sammenligne kvartals-nivåene for lokasjon. De seks mulige sammenligningene (første mot andre, første mot tredje osv.) gir at første kvartal har signifikant lavere nivå på lokasjonsparameteren enn de tre andre kvartalene, mens det er ikke grunnlag for forkastning av at de tre siste kvartalene er like. Signifikansnivået er 0,417 % som er 5 % nivå justert for to-sidig testing og for at det er seks forskjellige delhypoteser. Det er verdt å merke seg bedre estimering av ξ (mindre st.feil) enn tidligere og høyere lokasjons nivå i alle kvartaler.

Som i avsnitt (4.2.2) tester vi også om fjerde kvartal er forskjellig fra de andre, ved spesielle kvartals-dummyer. Det ser i figur 1 ut som om fjerde kvartal har høyere nivå i lokasjons-parameteren. Modellen er gitt ved:

(5.19) $$\mu_{k,j}=\mu_0+\beta s_{123,k,j}$$

der $\mathbf{s}_{123}$ er som i avsnitt (4.2.2) bare av dimensjon [56 000×1]. Resultatet av en slik tilpasning er gitt nedenfor:

$$\begin{pmatrix}\hat{\mu}_{1,j}\\ \hat{\mu}_{2,j}\\ \hat{\mu}_{3,j}\\ \hat{\mu}_{4,j}\end{pmatrix}=\begin{pmatrix}300,9\\ 300,9\\ 300,9\\ 347,97\end{pmatrix},\quad \mathrm{COV}_{\hat{\mu}_{k,j}}=\begin{pmatrix}693,7 & . & . & .\\ 693,7 & 693,7 & . & .\\ 693,7 & 693,7 & 693,7 & .\\ 675,5 & 675,5 & 675,5 & 791,1\end{pmatrix},$$

$$\begin{pmatrix}\hat{\sigma}\\ \hat{\xi}\end{pmatrix}=\begin{pmatrix}105,27\\ 0,207\end{pmatrix},\quad \begin{pmatrix}\mathrm{St.feil}(\hat{\sigma})\\ \mathrm{St.feil}(\hat{\xi})\end{pmatrix}=\begin{pmatrix}20,16\\ 0,079\end{pmatrix}$$

Tabell 16: ML-estimater i sesong (lokasjon) punktprosess modell fra (5.19)

Her har fjerde kvartal signifikant høyere lokasjons-nivå enn de tre første kvartalene med p-verdi=0,000024. ξ estimeres til å være lavere i denne modellen enn forrige og usikkerheten er også høyere relativt.

En sesongmodell for skala-parameteren tilsvarende (5.17) kan skrives:

(5.20) $$\sigma_{k,j} = \sigma_0 + \beta_1 s_{1,k,j} + \beta_2 s_{2,k,j} + \beta_3 s_{3,k,j}$$

og en tilpasning til dataene gir flg. resultat:

$$\begin{pmatrix} \hat{\sigma}_{1,j} \\ \hat{\sigma}_{2,j} \\ \hat{\sigma}_{3,j} \\ \hat{\sigma}_{4,j} \end{pmatrix} = \begin{pmatrix} 162,24 \\ 128,77 \\ 131,69 \\ 119,34 \end{pmatrix}, \quad \mathrm{COV}_{\hat{\sigma}_{k,j}} = \begin{pmatrix} 1\,639,02 & . & . & . \\ 1\,450,15 & 1\,308,76 & . & . \\ 1\,460,17 & 1\,306,03 & 1\,330,56 & . \\ 1\,394,64 & 1\,248,88 & 1\,257,45 & 1\,215,7 \end{pmatrix},$$

$$\begin{pmatrix} \hat{\mu} \\ \hat{\xi} \end{pmatrix} = \begin{pmatrix} 311,1 \\ 0,462 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{St.feil}(\hat{\mu}) \\ \mathrm{St.feil}(\hat{\xi}) \end{pmatrix} = \begin{pmatrix} 34,98 \\ 0,122 \end{pmatrix}$$

Tabell 17: ML-estimater i sesong (skala) punktprosess modell fra (5.20)

Med 0,417 % signifikansnivå har første kvartal signifikant høyere skala-nivå enn de tre andre kvartalene, som eneste signifikante forskjell. Dette kan virke rart når motsatt er tilfelle for lokasjonsparameteren, men kan som i avsnitt (4.2.2) forklares ved at når likelihooden maksimeres med hensyn på skala holdes lokasjonen konstant. Estimert lokasjonsnivå er mye nærmere de tre siste kvartalenes nivå fra tabell 15 enn det første. Derfor vil spredningen rundt denne verdien være større i første kvartal, altså et større skala-nivå.

Med en modell som sammenligner fjerde kvartal mot resten, tilsvarende som i (5.19) og spesifisert ved:

(5.21) $$\sigma_{k,j} = \sigma_0 + \beta s_{123,k,j}$$

fås flg. resultat:

$$\begin{pmatrix} \hat{\sigma}_{1,j} \\ \hat{\sigma}_{2,j} \\ \hat{\sigma}_{3,j} \\ \hat{\sigma}_{4,j} \end{pmatrix} = \begin{pmatrix} 112,81 \\ 112,81 \\ 112,81 \\ 90,44 \end{pmatrix}, \quad \mathrm{COV}_{\hat{\sigma}_{k,j}} = \begin{pmatrix} 441,22 & . & . & . \\ 441,22 & 441,22 & . & . \\ 441,22 & 441,22 & 441,22 & . \\ 382,28 & 382,28 & 382,28 & 353,82 \end{pmatrix},$$

$$\begin{pmatrix} \hat{\mu} \\ \hat{\xi} \end{pmatrix} = \begin{pmatrix} 312,15 \\ 0,223 \end{pmatrix}, \quad \begin{pmatrix} \mathrm{St.feil}(\hat{\mu}) \\ \mathrm{St.feil}(\hat{\xi}) \end{pmatrix} = \begin{pmatrix} 26,4 \\ 0,078 \end{pmatrix}$$

Tabell 18: ML-estimater i sesong (skala) punktprosess modell fra (5.21)

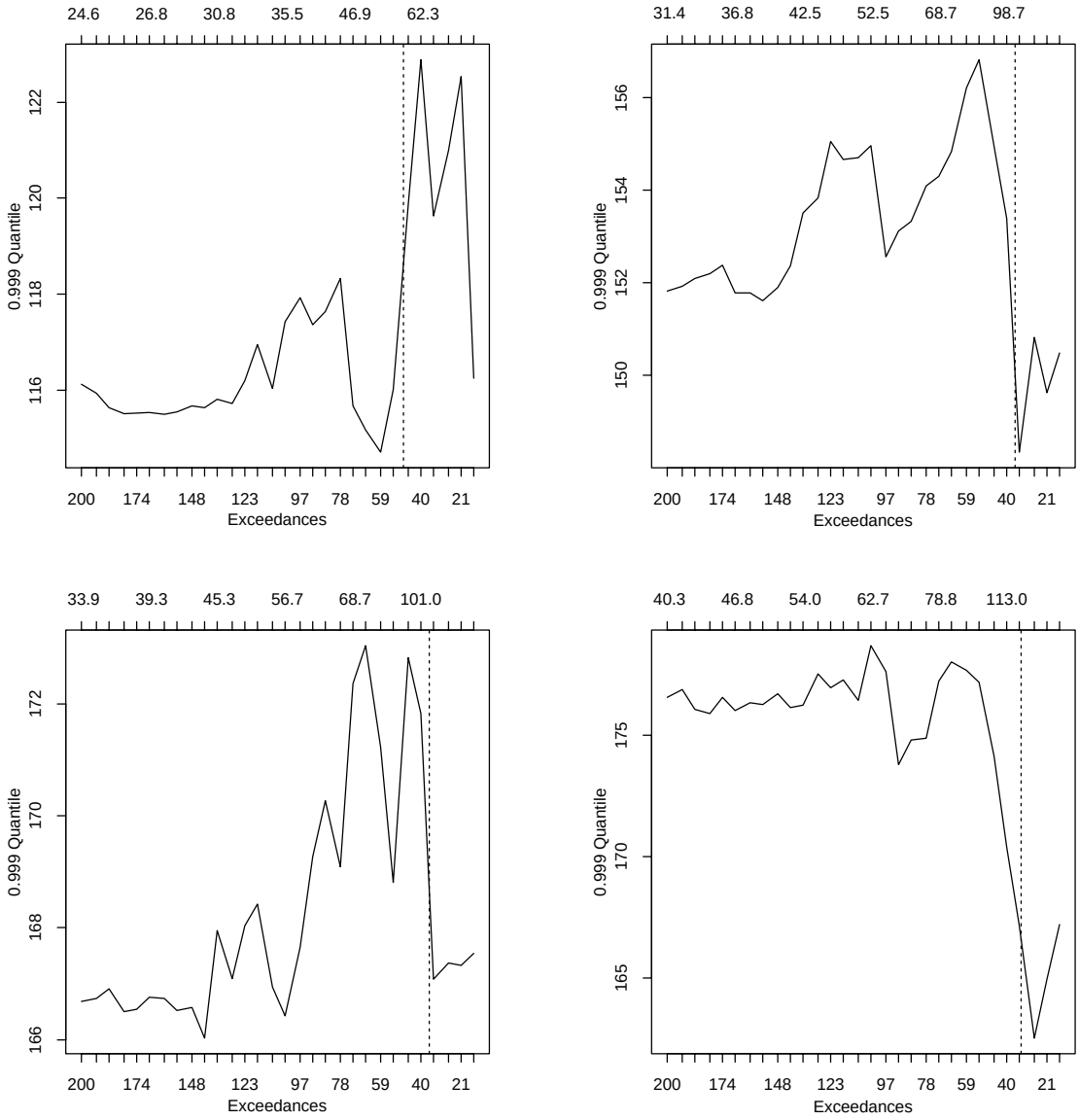
Her har fjerde kvartal signifikant lavere nivå på skala-parameteren enn de tre første kvartalene med p-verdi=0,000026. Dette er motsatt av resultatene for lokasjonsparameteren, helt tilsvarende som i den forrige skala modellen, og forklaringen er den samme. Estimert lokasjonsnivå er omtrent likt og man får den samme effekten på skalanivået. Igjen er estimeringen av ξ relativt litt mer usikker i denne modellen.

Tabellen nedenfor viser ML-kvantil estimater i investeringsfordelingen for hvert kvartal for seg (skala sesong modell), beregnet etter (5.7) og med forskjellige u for hvert kvartal som i tabell 14. De laveste kvantilene er negative og her er det stor variasjon fra kvartal til kvartal som illustrerer usikkerheten når man prøver å estimere kvantiler som er lavere enn u (5.7 forutsetter kvantiler som er høyere enn u). Den største kvantilen ($q_{0,001}$) er større enn u, her er det minst variasjon mellom kvartalene og de stemmer bra med kvantilen for modellen uten sesong (tabell 13) og de empiriske kvantilene som er 112,89, 149,57, 163,85 og 163,87 for de fire kvartalene.

	Transformert skala				Opprinnelig skala			
	Kv.1	Kv.2	Kv.3	Kv.4	Kv.1	Kv.2	Kv.3	Kv.4
$\hat{q}_{0.1}$	-24,82	47,07	-8,08	82,26	-24 822	47 075	-8 082	82 265
$\hat{q}_{0.05}$	-14,74	53,07	5,47	84,65	-14 744	53 072	5 471	84 650
$\hat{q}_{0.01}$	20,33	75,57	50,5	96,96	20 333	75 575	50 500	96 958
$\hat{q}_{0.001}$	119,74	147,89	168,51	162,22	119 736	147 888	168 514	164 221

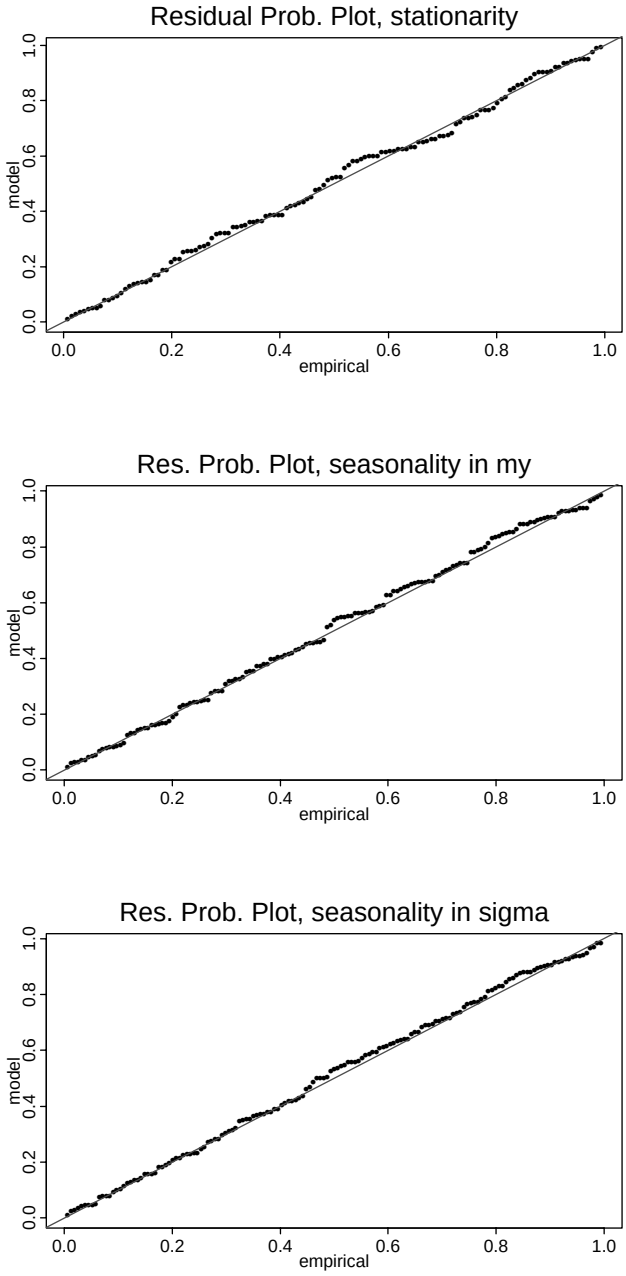
Tabell 19: ML kvantil estimater i GPD sesong modell (kvartalene hver for seg)

Figuren neste side viser den største kvantilen for hvert kvartal med varierende u. De valgte verdiene for u er også vist med stiplet linje. Kurvene ser urolige ut, men skalaen på y-aksen viser at de beveger seg ganske lite, og er nokså stabile rundt de valgte u-verdiene.



Figur 11: 99.9 % kvantilestimater i investeringsfordelingen for 1. (øverst til venstre), 2. (øverst til høyre), 3. (nederst til venstre) og 4. kvartal (nederst til høyre)

Neste side vises noen diagnostiske plott for residualene (observasjonene minus sin forventning) i sesong modellene og sammenlignet med den enkle modellen (tilsvarende plott som i figur 6). Øverst ser vi den enkle modellen uten tidsvarierende parametere. Kvantil plott for denne modellen (transformert til eksponential fordeling) er vist i figur 9. Her er $u = 110$ antall eksesser er 130. I plottene nedenfor er tilsvarende plott for de to sesong modellene vist. Her er u varierende etter tabell 14 og antall eksesser er 153. Vi kan se en noe bedre (jevnere) tilpasning i sesong modellene enn i den enkle modellen, med den beste modellen nederst (sesongvariasjon i σ). Dette er som i GEV modellen, men her er sesong forskjellene signifikante.



Figur 12: Diagnostiske residual plott for punkt prosess tilpasning

6. Konklusjon

Vi har sett at både GEV modellen og GPD (punkt-prosess) modellen gir bra tilpasning til investeringsdataene. Begge modellene viste forbedring ved sesong variasjon i parametre, mest i GPD modellen. I GEV modellen er det ikke signifikante sesong forskjeller, men klare indikasjoner for signifikant sesong modell i skala-nivået. I GPD modellen er det klart signifikante sesongforskjeller. Lokasjons nivået er klart lavere i første kvartal enn i de tre andre, og hvis man er interessert i å sammenligne fjerde kvartal mot de tre første kvartalene som enhet er det signifikant høyere nivå i fjerde kvartal. Dette stemmer intuitivt med det vi kan se fra dataene i figur 1. For skala-nivået derimot er det motsatt. Det er signifikant høyere nivå i første kvartal enn i de tre andre og signifikant lavere i fjerde kvartal enn i de tre første kvartalene når disse betraktes som en enhet. Forklaringen på dette er at maksimeringen av likelihooden for begge parametrene må ses i sammenheng. Med skala sesong modell er nivået i lokasjons estimatet såpass høyt at spredning rundt denne høye verdien blir stor, spesielt i første kvartal.

Akkurat som blokkstørrelse i GEV modellen er kritisk og en avveiing mellom god grensefordelings-approksimasjon og god parameter-estimering, er valg av u kritisk i GPD modellen. Vi har sett at en styrke med GPD modellen er at den har flere observasjoner å gjøre inferens på som gjør at estimatet av nøkkelparameteren ξ blir bedre. Denne parameteren beskriver hvor tung halen i fordelingen er. Alle punktestimatene av ξ har vært mellom 0,2 og 0,47 som plasserer investerings-dataene i området til typiske eksempler eksempler innen finans (ligger mellom 0,25 og 0,33) og litt over, men likevel under de typiske anvendelsene fra forsikring (mellom 0,5 og 1) [1, side 291]. Estimering av lokasjon og skala i GPD modellen gir generelt høyere verdier enn i GEV modellen men usikkerheten (variasjonskoeffisienten) er omtrent den samme. Siden estimatene for ξ er sikrere i GPD modellen, velger vi å tro mest på estimatene derfra, også når det gjelder lokasjon og skala (skulle vært like i de to modellene). Videre har vi sett at kvantil estimering for dataene er mulig i begge modeller, for kvantiler som er større enn u.

Vi har ikke tatt høyde for avhengigheten i modellene våre og må derfor regne med litt feilmargin i beregning av standardfeil og p-verdier. Likevel viser analysen at modeller fra ekstremverditeori passer godt til investeringsdataene, og kan forhåpentlig være grunnlag for videre arbeid. Disse modellene har ikke vært benyttet i SSB før (så vidt jeg vet). Noen interessante oppgaver videre er å prøve å bruke noen av disse resultatene på prediksjon av totalinvesteringen for utvalget, og hvor godt modellene passer på hele bedriftspopulasjonen. En annen er analyse på næringsnivå, for å finne ut hva som skiller de forskjellige industrinæringene. Tilsvarende datasett som det som er brukt her finnes mange av, og andre økonomiske variable kan også undersøkes tilsvarende.

7. Referanser

[1]: "Modelling Extremal Events": Embrechts, P., Klüppelberg, C., Mikosch Th.
Springer 1999

[2]: "Extreme value theory and applications": Coles, Stuart
Notater fra Internett

[3]: Frigessi, Arnoldo
Forelesningsnotater i kurset ST396, høst 99

De sist utgitte publikasjonene i serien Notater

2000/64 R. N. Johnsen: Undersøking om foreldrebetaling i barnehagar, august 2000. 36s.

2000/65 A. Thomassen: Byggekostnadsindeks for rørleggerarbeid i kontor- og forretningsbygg. 14s.

2000/67 A.G. Hustoft og G. Olsen: Metadata for statistikk om personer og husholdninger : Forprosjektrapport. 34s.

2000/68 A. Bruvoll, K. Flugsrud og H. Medin: Dekomponering av endringer i utslipp til luft i Norge - dokumentasjon av data. 19s.

2000/69 M. Vik Dysterud og E. Engelien: Tettstedsavgrensing: Teknisk dokumentasjon 2000. 53s.

2000/70 A. Akselsen, G. Dahl, J. Lajord og Ø. Sivertstøl: FD - Trygd: Variabelliste. 48s.

2000/71 B.O. Lagerstrøm: Kompetanse i grunnskolen , del 2: Dokumentasjonsrapport. 19s.

2000/72 B.O. Lagerstrøm: Kompetanse i grunnskolen: Hovdresultater 1999/2000 170s.

2000/73 J.H. Wang: Kvartalsvis investeringsstatistikk. 57s.

2000/74 P.O. Lande og T. Hoel: Dødsårsaksregisteret: Systemdokumentasjon. 90s.

2000/75 A.G. Pedersen, P.O. Lande og T. Hoel: Dødsårsaksregisteret: Brukerdokumentasjon. 99s.

2000/76 A.G. Hustoft, B. Vannebo: En undersøkelse av frafallet i utvalgsundersøkelser i perioden 1997-2000. 56s.

2000/77 P.O. Lande og J. Kittelsen: Forbruksundersøkinga 2000. Innlasting/Innsjekking: Brukardokumentasjon. 17s.

2000/78 J. Fosen, A.K. Johnsen og G. Røyne: Frafall blant innvandrere. En undersøkelse av frafall i Utdanningsundersøkelsen 1999 og i valgundersøkelser blant innvandrere. 53s.

2000/79 J. Kittelsen og P.O. Lande: OPPSLAG - Forbruksundersøkelsen. Brukerdokumentasjon. 39s.

2000/80 J. Kittelsen og P. O. Lande: Forbruksundersøkinga 2000. Systemdokumentasjon. 156s.

2000/81 J.T. Lind: Testing av stokastiske individuelle effekter i paneldatamodeller. 17s.

2001/2 D.Q. Pham: Innføring i tidsserier - sesongjustering og X-12-AMIRA. 110s.

2001/3 O. Rognstad: Eiendomsomsetning. Dokumentasjon av datagrunnlag og bearbeidingsrutine. 72s.

2001/4 T. Nøtnæs: Innføring i kognitiv kartlegging. 20s.

2001/5 T. Bye, M. Hansen og B. Strøm: Hvordan framskrive utslipp av klimagasser? 16s.

2001/6 A. Langørgen og R. Aaberge: KOMMODE II estimert på data for 1998. 16s.

2001/7 B.R. Joneid og J. Lajord: FD - Trygd: Dokumentasjonsrapport. Stønader til enslig forsørger. 1992-1999. 39s.

2001/8 T. Karlsen, E. Karstensen og E. Evensen: Beregningsrutiner og teknisk programstruktur for fylkesfordelt nasjonalregnskap. 27s.

2001/9 L. Rognstad, N.M. Stølen, T. Jakobsen og P. Schøning: Regional statistikk og analyse - strategi og prioriteringer. 45s.

2001/10 A. Akselsen og B.R. Joneid: FD - Trygd: Dokumentasjonsrapport. Pensjoner. Grunn- og hjelpestønader. 1992-1998. 94s.

2001/11 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 1999. 34s.

2001/12 A. Rognan og N. Barrabés: NUS2000. Dokumentasjonsrapport. 36s.

2001/13 K.I. Bøe, J. Johansen og Ø. Sivertstøl: FD - Trygd: Dokumentasjonsrapport. Attføringspenger, 1992-1998. 88s.