Statistics Norway

Statistisk sentralbyrå

Ole Klungsøyr

Sammenligning av mikro-formler for prisindekser og modelltilpasning

Notater

Korrigert utgave

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Innholdsfortegnelse

1. Innledning

Et veiet aritmetisk / geometrisk gjennomsnitt inngår i beregningen av en pris-indeks på det mest detaljerte nivået, og kan kalles pris-indeksens mikroformel. Hensikten med dette notatet er å sammenligne det veiede aritmetiske gjennomsnittet og det veiede geometriske gjennomsnittet som parameterestimater i hhv en additiv og en multiplikativ modell mhp statistiske egenskaper. Innføringen av den additive og multiplikative modellen gir en enkel utledning av de generelle optimale mikroformlene som i forskjellige versjoner gir de mest brukte mikroformlene i dag.

Seksjon for Økonomiske Indikatorer skiftet i august 1999 mikroformel i konsumpris-indeksen, og gikk over til et geometrisk gjennomsnitt, i tråd med anbefalinger fra International Labour Office (ILO) og Boskin-rapporten (se [3]). Begrunnelsene er flere, men et viktig økonomisk argument er at det geometriske gjennomsnittet til en viss grad tar høyde for det at konsumentene endrer kjøpeadferd ved prisendringer (substitusjon), noe den gamle mikroformelen ikke gjorde (veiet aritmetisk gjennomsnitt). Vi ser her kun på rent statistiske egenskaper, og skal se at det fra denne synsvinkelen også finnes god begrunnelse for å velge det geometriske gjennomsnittet, nemlig at det er mer robust. Å undersøke sammenhengen mellom de økonomiske og statistiske momentene, og prøve å kvantifisere de økonomiske argumentene er interessant, og kanskje en naturlig fortsettelse av dette notatet.

Vi tar utgansgspunkt i faktiske prisobservasjoner fra konsumprisindeksen, men vil også simulere nye prisobservasjoner for å sjekke egenskapene til de forskjellige estimatene ved å kunne kontrollere hvilken modell som virkelig har generert dataene. Forskjellen på modellene er om "støyen" i modellen inngår additivt eller multiplikativt.

Prisobservasjonene er her slått sammen mht geografiske områder (hele landet er delt inn i 8 forskjellige områder), men beholdt på forskjellige representantvarenivå (her kalt varenr) som er ca. 850 forskjellige. Til sammen er det ca. 46 000 observasjoner totalt hver måned. Analysen her er gjort med utgangspunkt i observasjoner fra juli 98 (referansemåned) og desember 1998.

Notasjonen som vil bli brukt gjennomgående er:

$P_{t,i}$: prisobservasjon *i* for en bestemt representant-vare ved tidspunkt *t*

$P_{0,i}$: tilsvarende prisobservasjon ved referansetidspunktet (juli)

β_t : indeksparameteren som skal estimeres (felles for observasjonene innen en representantvare)

ε_i : tilfeldig feil (støy) i modellen, med gitte fordelingsantagelser

De mest brukte mikroformlene er $\frac{\sum_i P_{t,i}}{\sum_i P_{0,i}}$, $\frac{1}{n}\sum_i \frac{P_{t,i}}{P_{0,i}}$ og $\prod_i \left(\frac{P_{t,i}}{P_{0,i}}\right)^{\frac{1}{n}}$ og vi skal se i kapittel 2 og 3 (ligning (2.3) og (3.5)) at disse er spesialtilfeller av de generelle uttrykkene for de optimale estimatene under additiv og multiplikativ modell.

Vi skal se at det veiede geometriske gjennomsnittet (multiplikative estimatet) er mer robust overfor modell og fordeling på støy enn det aritmetiske (additive estimatet). Ved å bruke de faktiske prisobservasjonene til å beregne det additive estimatet $\hat{\beta}_a$ og det multiplikative estimatet $\hat{\beta}_m$ er vi istand til å generere nye prisobservasjoner under den modellen vi ønsker og med kjent fordeling på støyen. Vi bruker altså referanseprisene $P_{0,i}$, $\hat{\beta}_a$ eller $\hat{\beta}_m$, og simulerer den tilfeldige komponenten ε_i fra en kjent fordeling for så å generere nye $P_{t,i}$ under kjente betingelser. Dermed kan vi lage nye estimater og sammenligne egenskapene deres. Som resultatene vil vise er støyfordelingen av stor betydning. Når denne er skjev nok mot høyre vil det additive estimatet fungere bra når den additive modellen har generert dataene (estimering under sann modell) og dårlig når den multiplikative modellen har generert dataene (estimering under feil modell), mens det multiplikative estimatet fungerer bra i begge tilfeller. Målt i skjevhet, standard-feil og RMSE (root mean square error), beskrevet i kapittel 5, kan hoved-resultatet oppsummeres i flg. tabell (gjelder for alle 829 varenumrene og når støyen er gammafordelt):

	ADDITIVT ESTIMAT $\hat{\beta}_a$		MULTIPLIKATIVT ESTIMAT $\hat{\beta}_m$	
	Sann modell	Feil modell (*)	Sann modell	Feil modell (*)
$\lvert$Skjevhet$\rvert$	≤ 0.009	≤ 8.428 (37.39 %)	≤ 0.046	≤ 0.103 (1.09 %)
Standard-feil	≤ 0.168	≤ 177.5 (21.35 %)	≤ 0.473	≤ 0.175 (0 %)
RMSE	≤ 0.168	≤ 177.7 (21.59 %)	≤ 0.476	≤ 0.188 (0 %)

(*): Andel varenummer (av 829) som har nivå større enn max-nivået i sann modell

Tabell 1: Simuleringsresultater som viser forskjell på $\hat{\beta}_a$ og $\hat{\beta}_m$ (gammafordelt støy)

Tabellen viser tydelig hvordan det additive estimatet "bryter sammen" under feil modell. Det multiplikative estimatet er noe dårligere enn det additive under sann modell, men til gjengjeld omtrent like bra under feil og sann modell.

2. Additiv modell

Den additive modellen for pris-utvikling innen et varenr er gitt ved :

(2.1) $$P_{t,i} = \beta_t P_{0,i} + \varepsilon_i, \quad \text{eller} \quad \frac{P_{t,i}}{P_{0,i}} = \beta_t + \frac{1}{P_{0,i}}\varepsilon_i$$

der ε_i antas å ha flg. egenskaper:

(2.2) $$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = w_i \sigma_a^2$$

Vi har altså en enkel, veiet lineær regresjonsmodell, der w_i er en tenkt "varians-inflasjon" som gjerne er en funksjon av referanseprisen og som gjør modellen istand til å beskrive f.eks. at variansen øker med referanseprisen. $P_{0,i}$ (referanseprisen) kan betraktes som en kjent størrelse i modellen og estimatene, slik som regresjonsmodellen vanligvis betinger på forklaringsvariabelen. Minstekvadraters estimatet av β_t er gitt ved:

(2.3) $$\hat{\beta}_t = \frac{\sum_i \frac{1}{w_i} P_{0,i} P_{t,i}}{\sum_i \frac{1}{w_i} P_{0,i}^2} = \sum_i \left(\frac{\frac{P_{0,i}^2}{w_i}}{\sum_i \frac{P_{0,i}^2}{w_i}} \right) \frac{P_{t,i}}{P_{0,i}}$$

som kan ses på som et veiet aritmetisk gjennomsnitt av prisforholdene, som med $w_i = P_{0,i}^2$ gir det "vanlige" aritmetiske gjennomsnittet. Vanlige antagelser for w_i i økonomiske modeller er $w_i = P_{0,i}^{\alpha}$, for $1 \leq \alpha \leq 2$ (se f.eks. [2]). Her ser vi på $\alpha = 0,1,2$ slik at $w_i = 1, P_{0,i}, P_{0,i}^2$ (modellen uten varians-inflasjon er dermed også med i betraktning).

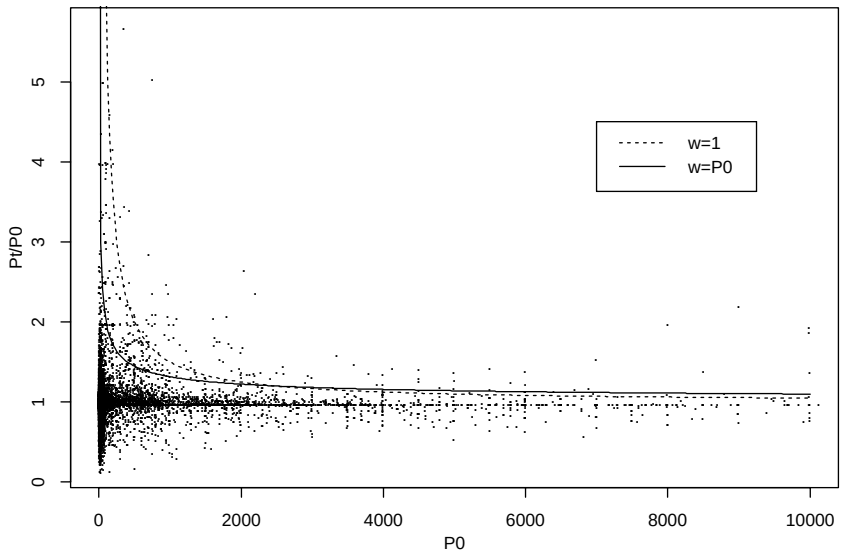
Estimatet for σ_a^2 er:

(2.4) $$\hat{\sigma}_a^2 = \frac{1}{n-1} \sum_i \left(\frac{1}{\sqrt{w_i}} (P_{t,i} - \hat{\beta}_t P_{0,i}) \right)^2$$

Fra (2.1) følger at

(2.5) $$Var\left(\frac{P_{t,i}}{P_{0,i}}\right) = \frac{w_i}{P_{0,i}^2} \sigma_a^2, \quad \text{eller} \quad SD\left(\frac{P_{t,i}}{P_{0,i}}\right) = \frac{\sqrt{w_i}}{P_{0,i}} \sigma_a$$

For å beskrive (2.5) grafisk er $\frac{P_{t,i}}{P_{0,i}}$ plottet mot $P_{0,i}$ i figuren nedenfor.



Figur 1: Valg av w_i i additiv modell

I figuren er alle pris-observasjoner under 10 000 tatt med, så det er ikke realistisk i forhold til at valg av w_i skal gjøres på varenr nivå, men såpass mange observasjoner i samme intervallet gir en god illustrasjon på varians-inflasjon. Den stiplede linjen er proporsjonal med $\frac{1}{P_{0,i}}$ og fra (2.5) ses at dette tilsvarer hvordan det teoretiske standard-avviket i $\frac{P_{t,i}}{P_{0,i}}$ varierer med $P_{0,i}$ når $w_i = 1$. Tilsvarende er den heltrukne linjen proporsjonal med $\frac{1}{\sqrt{P_{0,i}}}$ og beskriver standard-avviket i $\frac{P_{t,i}}{P_{0,i}}$ som funksjon av $P_{0,i}$ når $w_i = P_{0,i}$. Begge kurvene synes å følge standard-avviket i observasjonene brukbart og antyder at $w_i = P_{0,i}^2$ som tilsvarer konstant standard-avvik i $\frac{P_{t,i}}{P_{0,i}}$ ikke passer dette samlede utsnittet av dataene, selv om det også ved konstant standard-avvik vil være størst spredning der antall observasjoner er størst.

2.1. Symmetrisk støy (normalfordeling)

Med normalantagelse på støyen følger at prisforholdet har flg. betingede fordeling:

(2.6) $$\frac{P_{t,i}}{P_{0,i}} \mid P_{0,i} \sim N(\beta_t, \frac{w_i}{P_{0,i}^2}\sigma_a^2)$$

og (2.3) er lik "Maximum likelihood" estimatet. Heretter vil vi droppe den betingede notasjonen. All analyse og estimering gjøres innenfor hvert varenummer. Ved normalfordelt støy, w_i som funksjon av $P_{0,i}$ og sistnevnte som en kjent størrelse, vil $\hat{\beta}_t$ også være normal-fordelt (fra (2.3)) (en veiet sum av normal-fordelinger) med forventning og varians gitt ved:

(2.7) $$E(\hat{\beta}_t) = \beta_t, \quad Var(\hat{\beta}_t) = \frac{1}{\sum_i \frac{P_{0,i}^2}{w_i}}\sigma_a^2$$

Som (2.7) viser er altså parameter-estimatet forventningsrett under modell-antagelsene.

2.2. Usymmetrisk støy (gammafordeling)

Vi skal senere se at formen på støyen og omfanget av den (variansen) er avgjørende for hvilken estimator som er best. Som usymmetrisk støy velger vi gamma-fordelingen. Modellen er som i (2.1), fremdeles med forventning og varians som i (2.2), men nå er ε_i gammafordelt, se appendix. Siden denne fordelingen kun er definert for positive verdier, må vi bruke en lineær transformasjon for at (2.2) skal holde. Vi velger altså, som før:

(2.8) $$\frac{P_{t,i}}{P_{0,i}} = \beta_t + \frac{1}{P_{0,i}}\varepsilon_i$$

og videre:

(2.9) $$\varepsilon_i = \sqrt{w_i}\,(\eta_i - \alpha_a)$$

der η_i er standard gammafordelt med parameter α_a ($E(\eta_i) = Var(\eta_i) = \alpha_a$). Dermed er (2.8):

(2.10) $$\frac{P_{t,i}}{P_{0,i}} = \beta_t + \frac{1}{P_{0,i}}\left(\sqrt{w_i}\,(\eta_i - \alpha_a)\right) = \frac{\sqrt{w_i}}{P_{0,i}}\eta_i + (\beta_t - \frac{\sqrt{w_i}}{P_{0,i}}\alpha_a) \quad \sim G(\alpha_a, \theta_{a,i}, \gamma_{a,i})$$

altså tre-parameter gamma-fordelt med:

(2.11) $$\theta_{a,i} = \frac{\sqrt{w_i}}{P_{0,i}}, \quad \gamma_{a,i} = \beta_t - \frac{\sqrt{w_i}}{P_{0,i}}\alpha_a$$

Forventning og varians til $\frac{P_{t,i}}{P_{0,i}}$ er da (se appendix):

(2.12) $$E(\frac{P_{t,i}}{P_{0,i}}) = \beta_t, \quad Var(\frac{P_{t,i}}{P_{0,i}}) = \frac{w_i}{P_{0,i}^2}\alpha_a$$

Fra (2.3) ses at $\hat{\beta}_t$ nå er en veiet sum av gamma-fordelte variable, og får forventning og varians som i (2.7):

(2.13) $$E(\hat{\beta}_t) = \beta_t, \quad Var(\hat{\beta}_t) = \frac{1}{\sum_i \frac{P_{0,i}^2}{w_i}}\alpha_a$$

Fordelingen til $\hat{\beta}_t$ er nå vanskelig å finne eksakt, men det er heller ikke nødvendig siden resultatene baseres på simuleringer.

3. Multiplikativ modell

Denne modellen er spesifisert ved:

(3.1) $$P_{t,i} = \beta_t P_{0,i}\varepsilon_i, \quad \text{eller} \quad \frac{P_{t,i}}{P_{0,i}} = \beta_t\varepsilon_i$$

og den blir additiv på log-skala:

(3.2) $$\log(\frac{P_{t,i}}{P_{0,i}}) = \log(\beta_t) + \log(\varepsilon_i)$$

Med utgangspunkt i log-skalaen blir resultatene tilsvarende som i forrige avsnitt, og tilbake-transformasjon gir resultatene på opprinnelig skala. De generelle antagelsene på støyen er nå

(3.3) $$E\left(\log(\varepsilon_i)\right) = 0, \quad Var\left(\log(\varepsilon_i)\right) = w_i\sigma_m^2$$

Minstekvadraters estimatet av $\log(\beta_t)$ er som i (2.3) et veiet aritmetisk gjennomsnitt:

(3.4) $$\log(\hat{\beta}_t) = \sum_i \left(\frac{\frac{1}{w_i}}{\sum_i \frac{1}{w_i}} \right) \log(\frac{P_{t,i}}{P_{0,i}})$$

eller på opprinnelig skala:

(3.5) $$\hat{\beta}_t = \prod_i \left(\frac{P_{t,i}}{P_{0,i}} \right)^{\frac{\frac{1}{w_i}}{\sum_i \frac{1}{w_i}}}$$

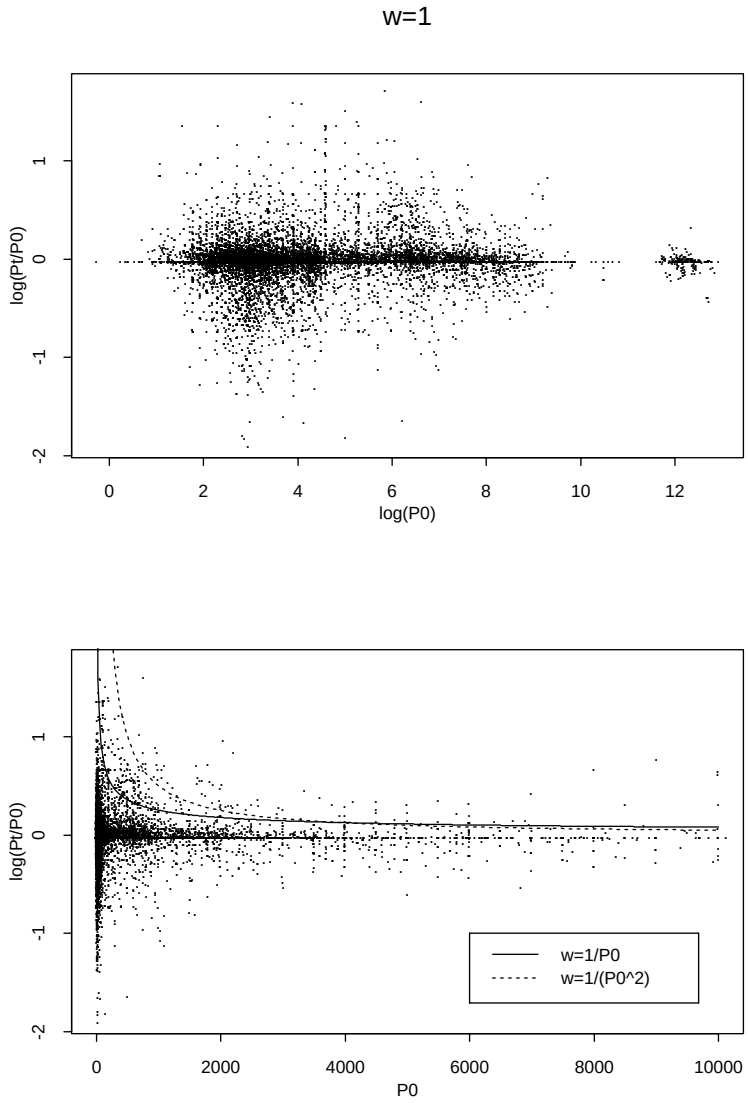
som kan betraktes som et generalisert geometrisk gjennomsnitt som ved $w_i = 1$ gir det "vanlige" geometriske gjennomsnittet. Som i den additive modellen er $w_i = P_{0,i}^{\alpha}$, men her har vi brukt $\alpha = -2, -1, 0$ ($w_i = \frac{1}{P_{0,i}^2}, \frac{1}{P_{0,i}}, 1$). Variansestimatet er også tilsvarende som i additiv modell :

(3.6) $$\hat{\sigma}_m^2 = \frac{1}{n-1} \sum_i \left(\frac{1}{\sqrt{w_i}} \left[\log(\frac{P_{t,i}}{P_{0,i}}) - \log(\hat{\beta}_t) \right] \right)^2$$

Fra (3.2) følger at:

(3.7) $$Var\left(\log(\frac{P_{t,i}}{P_{0,i}})\right) = w_i\sigma_m^2, \quad \text{eller} \quad SD\left(\log(\frac{P_{t,i}}{P_{0,i}})\right) = \sqrt{w_i}\,\sigma_m$$

og som i forrige avsnitt er forskjellige valg av w_i illustrert i figuren nedenfor.



Figur 2: Valg av w_i i multiplikativ modell

I den øverste figuren er det log-skala på x-aksen. Dermed er hele $P_{0,i}$ -området tatt med og området med høyest observasjons-tetthet (små $P_{0,i}$) er "strukket" ut. Den viser at et konstant standard-avvik ikke er urimelig, selv om nederste figuren synes å vise at $w_i = \frac{1}{P_{0,i}}, \frac{1}{P_{0,i}^2}$ passer best.

3.1. Symmetrisk støy (normal- / lognormal-fordeling)

Med normalantagelse på $\log(\varepsilon_i)$ i (3.2) vil vi ha

(3.8) $$\log(\frac{P_{t,i}}{P_{0,i}}) \sim N\left(\log(\beta_t), w_i\sigma_m^2\right)$$

og dette tilsvarer at $\frac{P_{t,i}}{P_{0,i}}$ er lognormal-fordelt på opprinnelig skala. Fra (3.4) ser vi at $\log(\hat{\beta}_t)$ dermed også er normalfordelt (veiet sum av normal-fordelinger) med flg. forventning og varians:

(3.9) $$E\left(\log(\hat{\beta}_t)\right) = \beta_t\,, \quad Var\left(\log(\hat{\beta}_t)\right) = \frac{1}{\sum_i \frac{1}{w_i}}\sigma_m^2$$

Siden $\log(\hat{\beta}_t)$ er normal-fordelt er $\hat{\beta}_t$ selv log-normalfordelt med forventning og varians gitt ved (se appendix):

(3.10) $$E\left(\hat{\beta}_t\right) = \exp\left(\log(\beta_t) + \tfrac{1}{2}\frac{1}{\sum_i \frac{1}{w_i}}\sigma_m^2 \right)$$

(3.11) $$Var\left(\hat{\beta}_t\right) = \exp\left(2\log(\beta_t) + 2\frac{1}{\sum_i \frac{1}{w_i}}\sigma_m^2 \right) - \exp\left(2\log(\beta_t) + \frac{1}{\sum_i \frac{1}{w_i}}\sigma_m^2 \right)$$

Som viser at parameter-estimatet ikke er forventningsrett under modellen, selv om skjevheten er liten og avhengig av w_i .

3.2. Usymmetrisk støy (gamma- / loggamma-fordeling)

Med modellen fra (3.2) og gammafordeling på $\log(\varepsilon_i)$ må vi som i avsnitt (2.2) bruke en lineær-transformasjon for å oppnå kravene i (3.3). Vi setter:

(3.12) $$\log(\varepsilon_i) = \sqrt{w_i}\,(\eta_i - \alpha_m)$$

der η_i er, som i additiv modell, standard gammafordelt med parameter α_m $(E(\eta_i) = Var(\eta_i) = \alpha_m)$. Da kan (3.2) skrives som:

(3.13) $$\log(\tfrac{P_{t,i}}{P_{0,i}}) = \log(\beta_t) + \sqrt{w_i}\,(\eta_i - \alpha_m) = \sqrt{w_i}\,\eta_i + \left(\log(\beta_t) - \sqrt{w_i}\,\alpha_m\right) \sim G(\alpha_m, \theta_{m,i}, \gamma_{m,i}),$$

altså tre-parameter gamma-fordelt med:

(3.14) $$\theta_{m,i} = \sqrt{w_i}\,, \quad \gamma_{m,i} = \log(\beta_t) - \sqrt{w_i}\,\alpha_m$$

Da blir forventning og varians til $\log(\tfrac{P_{t,i}}{P_{0,i}})$:

(3.15) $$E\left(\log(\tfrac{P_{t,i}}{P_{0,i}})\right) = \log(\beta_t), \quad Var\left(\log(\tfrac{P_{t,i}}{P_{0,i}})\right) = w_i\alpha_m$$

Når $\log(\tfrac{P_{t,i}}{P_{0,i}})$ er gamma-fordelt tilsvarer det at $\tfrac{P_{t,i}}{P_{0,i}}$ er log-gammafordelt (se appendix) med forventning og varians:

(3.16) $$E\left(\tfrac{P_{t,i}}{P_{0,i}}\right) = \left(\frac{1}{1-\sqrt{w_i}}\right)^{\alpha_m} \exp\left(\log(\beta_t) - \sqrt{w_i}\,\alpha_m\right), \quad \sqrt{w_i} < 1$$

(3.17) $$Var\left(\tfrac{P_{t,i}}{P_{0,i}}\right) = \left[\left(\frac{1}{1-2\sqrt{w_i}}\right)^{\alpha_m} - \left(\frac{1}{1-\sqrt{w_i}}\right)^{2\alpha_m} \right] \exp\left[2\left(\log(\beta_t) - \sqrt{w_i}\,\alpha_m\right)\right], \quad \sqrt{w_i} < \tfrac{1}{2}$$

Fra (3.4) ses at $\log(\hat{\beta}_t)$ er en veiet sum av gamma-fordelinger med varierende parametre og dermed vanskelig og regne ut analytisk, men det følger av (3.4) at:

(3.18) $$E\left(\log(\hat{\beta}_t)\right) = \log(\beta_t), \quad Var\left(\log(\hat{\beta}_t)\right) = \frac{1}{\sum_i \frac{1}{w_i}}\alpha_m$$

Siden vi ikke vet fordelingen til $\log(\hat{\beta}_t)$ kan vi heller ikke si noe om fordelingen til $\hat{\beta}_t$ på opprinnelig skala, eller om forventning og varians, men simulering gjør det mulig å undersøke disse egenskapene.

4. Hvordan bestemme w_i

Fra figur 1 og 2 har vi sett grafisk at både additiv og multiplikativ modell har bra tilpassning til dataene. Det er imidlertid vanskelig ut fra figurene å si om additiv eller multiplikativ modell passer best, og videre hvilken av de additive / multiplikative modellene. For best mulig estimering på varenr nivå må w_i kunne variere fra varenr til varenr. For å finne den best mulige w_i tas utgangspunkt i å minimere residualene. I den additive modellen f.eks., er residualene gitt ved:

(4.1) $$\varepsilon_i = P_{t,i} - \beta_t P_{0,i}$$

og estimert residual:

(4.2) $$\hat{\varepsilon}_i = P_{t,i} - \hat{\beta}_t P_{0,i}$$

der $\hat{\beta}_t$ er gitt ved (2.3) og funksjon av w_i. For å bestemme "optimal" w_i beregner vi den w_i, blant tre forskjellige, som minimerer den totale residual-kvadrat-summen, altså:

(4.3) $$\hat{w}_{add} = \underset{w}{\arg\min}\left[\sum_i \left(P_{t,i} - \hat{\beta}_t(w)P_{0,i}\right)^2\right]$$

og i den multiplikative modellen

(4.4) $$\hat{w}_{mult} = \underset{w}{\arg\min}\left[\sum_i \left(\log(\tfrac{P_{t,i}}{P_{0,i}}) - \log(\hat{\beta}_t(w))\right)^2\right]$$

Både i den additive og den multiplikative modellen viser det seg at på de dataene vi har prøvd minimeres kvadratsumresidualene med konstant varians, altså $\hat{w}_{add} = 1$ og $\hat{w}_{mult} = 1$, for alle varenummer.

5. Estimering under sann / feil modell

Vi tar utgangspunkt i modellene fra de forrige avsnittene, men vil nå skille mellom sann modell og antatt modell. For å sammenligne den additive og den multiplikative modellen vil vi simulere skjevhet, standard-feil og RMSE (root mean square error) til $\hat{\beta}_t$ i additiv og multiplikativ modell, når antatt og sann modell er like og når de er ulike. Vi bruker $P_{0,i}$ pris-observasjonene og parameter-estimatene under begge modellene til å simulere nye pris-data $P_{t,i}$, og kan dermed beregne nye parameter-estimater under sann og feil modell. Dette er så grunnlaget for å estimere skjevhet, standard-feil og RMSE som gjør oss i stand til å sammenligne estimatene.

For å unngå unødig komplisert notasjon dropper vi indeks "i" (for pris observasjon innen et varenr) og indeks "t" der det er hensiktsmessig. Det skal fremkomme hvilken modell som er sann og hvilken som er antatt. Indeks "a" og "m" markerer hhv additiv og multiplikativ modell og "s" betyr sann modell. Antatt modell vil ha indeks "j" ($j = 0,1,2$) for å skille mellom de tre forskjellige "w" som antas. Videre vil sann "w" i f.eks. den additive modellen betegnes med $w_{a,s}$, og den estimerte (kfr. avsnitt 4) betegnes med $\hat{w}_{a,s}$. Tilsvarende er $\hat{\beta}_{a,j}$ parameter-estimatet i de tre antatt additive modellene og $\hat{\beta}_{a,s}$ er estimatet under sann modell og med optimal "w".

Analysen gjøres på varenr. nivå, dvs. at all estimering og simulering gjøres for hvert av 829 varenr (her er varenr som har kun 1 observasjon eller få og "avvikende" observasjoner fjernet, samt alle pris-observasjoner som er 0). Simuleringen gjøres ved å "trekke" fra angitt fordeling. Dette gjøres K ganger som hvis stor nok gir et fornuftig estimat på teoretisk skjevhet, varians, MSE og RMSE i de forskjellige estimatene. De teoretiske størrelsene (skjevheten betegnes med "bias"):

$$\text{(5.1)} \qquad \text{bias}(\hat{\beta}_j) = \text{E}(\hat{\beta}_j - \beta), \quad \text{MSE}(\hat{\beta}_j) = \text{E}\left((\hat{\beta}_j - \beta)^2\right), \quad \text{Var}(\hat{\beta}_j) = \text{E}\left((\hat{\beta}_j - \text{E}\hat{\beta}_j)^2\right)$$

estimeres ved (vi bruker samme navnene på de empiriske størrelsene, selv om det egentlig er estimater):

$$\text{(5.2)} \qquad \text{bias}(\hat{\beta}_{.,j}) = \tfrac{1}{K}\sum_{k=1}^{K}\left(\hat{\beta}^k_{.,j} - \hat{\beta}_{a,s}\right)$$

$$\text{(5.3)} \qquad \text{MSE}(\hat{\beta}_{.,j}) = \tfrac{1}{K}\sum_{k=1}^{K}\left(\hat{\beta}^k_{.,j} - \hat{\beta}_{a,s}\right)^2$$

$$\text{(5.4)} \qquad \text{Var}(\hat{\beta}_{.,j}) = \text{MSE}(\hat{\beta}_{.,j}) - \left(\text{bias}(\hat{\beta}_{.,j})\right)^2$$

Her er den sanne modellen additiv, og det blir helt tilsvarende for den multiplikative. Prikk-indeksene står for enten "a" eller "m", antatt modell. "K" er her satt lik 500. Dette gir et standard-avvik til bias-estimatet på under 5 % av standard-avviket til parameter-estimatet som er rimelig bra presisjon. 500 simuleringer tar ca. 14 timer og dette er årsaken til at ikke "K" er valgt høyere.

Når den sanne og den antatte modellen er like, er beregningene enklest og følger direkte av avsnitt 2 og 3.

5.1. Sann modell

Her er antatt og sann modell like. Dette gir enklest regning og først ser vi på de teoretiske egenskapene under disse forutsetningene. Resultat-avsnittet følger samme inndeling slik at når antatt og sann modell er like betyr det at vi simulerer (sann modell) og estimerer (antatt modell) under samme betingelser.

5.1.1. Additiv

Først betrakter vi den additive modellen. Fra avsnitt 2 har vi at ved normal-fordeling / gamma-fordeling på støyen er modellen spesifisert ved:

(5.5) $$\frac{P_t}{P_0} \sim N(\beta_{a,s}, \frac{w_{a,s}}{P_0^2}\sigma_a^2) \text{, eller } \frac{P_t}{P_0} \sim G\left(\alpha_a, \frac{\sqrt{w_{a,s}}}{P_0}, (\beta_{a,s} - \frac{\sqrt{w_{a,s}}}{P_0}\alpha_a)\right)$$

og fra (2.3) at parameter-estimatet er:

(5.6) $$\hat{\beta}_{a,j} = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right) \frac{P_t}{P_0}, \quad j = 0,1,2$$

Det følger dermed at:

(5.7) $$E(\hat{\beta}_{a,j}) = \beta_{a,s}, \quad j = 0,1,2$$

(5.8) $$Var(\hat{\beta}_{a,j}) = \sum_i \frac{P_0^2 \frac{w_{a,s}}{w_{a,j}^2}}{\left(\sum_i \frac{P_0^2}{w_{a,j}}\right)^2}\sigma_a^2, \quad \text{eller} \quad Var(\hat{\beta}_{a,j}) = \sum_i \frac{P_0^2 \frac{w_{a,s}}{w_{a,j}^2}}{\left(\sum_i \frac{P_0^2}{w_{a,j}}\right)^2}\alpha_a$$

der variansuttrykket refererer til normal-fordeling / gamma-fordeling og $w_{a,0} = 1, w_{a,1} = P_0, w_{a,2} = P_0^2$.

5.1.2. Multiplikativ

Den multiplikative modellen er som i avsnitt 3 lettest spesifisert på log-skala ved:

(5.9) $$\log(\frac{P_t}{P_0}) \sim N\left(\log(\beta_{m,s}), w_{m,s}\sigma_m^2\right), \text{ eller } \log(\frac{P_t}{P_0}) \sim G\left(\alpha_m, \sqrt{w_{m,s}}, \left[\log(\beta_{m,s}) - \sqrt{w_{m,s}}\alpha_m\right]\right)$$

og parameter-estimatet fra (3.4):

(5.10) $$\log(\hat{\beta}_{m,j}) = \sum_i \left(\frac{\frac{1}{w_{m,j}}}{\sum_i \frac{1}{w_{m,j}}} \right) \log(\frac{P_t}{P_0}), \quad j = 0,1,2$$

Herav:

(5.11) $$E\left(\log(\hat{\beta}_{m,j})\right) = \log(\beta_{m,s}), \quad j = 0,1,2$$

(5.12) $$Var\left(\log(\hat{\beta}_{m,j})\right) = \sum_i \frac{\frac{w_{m,s}}{w_{m,j}^2}}{\left(\sum_i \frac{1}{w_{m,j}}\right)^2}\sigma_m^2, \quad \text{eller} \quad Var\left(\log(\hat{\beta}_{m,j})\right) = \sum_i \frac{\frac{w_{m,s}}{w_{m,j}^2}}{\left(\sum_i \frac{1}{w_{m,j}}\right)^2}\alpha_m$$

der $w_{m,0} = \frac{1}{P_0^2}$, $w_{m,1} = \frac{1}{P_0}$, $w_{m,2} = 1$. I normal-fordelingstilfellet vil $\hat{\beta}_{m,j}$ være lognormal-fordelt og vi har eksakt forventning og varians gitt ved:

$$(5.13) \qquad E(\hat{\beta}_{m,j}) = \exp\left(\log(\beta_{m,s}) + \frac{1}{2} \sum_i \frac{\frac{w_{m,s}}{w_{m,j}^2}}{(\sum_i \frac{1}{w_{m,j}})^2} \sigma_m^2 \right), \quad j = 0,1,2$$

$$(5.14) \qquad Var(\hat{\beta}_{m,j}) = \exp\left(2\log(\beta_{m,s}) + 2\sum_i \frac{\frac{w_{m,s}}{w_{m,j}^2}}{(\sum_i \frac{1}{w_{m,j}})^2} \sigma_m^2 \right) - \exp\left(2\log(\beta_{m,s}) + \sum_i \frac{\frac{w_{m,s}}{w_{m,j}^2}}{(\sum_i \frac{1}{w_{m,j}})^2} \sigma_m^2 \right)$$

Når støyen er gamma-fordelt er det vanskelig å finne analytisk uttrykk for forventning og varians, men lett ved simulering (neste avsnitt).

5.2. Feil modell

Denne situasjonen beskriver effekten på estimeringen av at den sanne modellen er en annen enn den man tror den er. Resultatene i neste avsnitt fremkommer ved å simulere fra en modell og bruke estimatorene som er optimale under den andre modellen.

5.2.1. Additiv

Her antas at modellen som genererer dataene er additiv, mens man bruker estimatene fra den multiplikative modellen. Modellen er altså som i (5.5):

$$(5.15) \qquad \frac{P_t}{P_0} \sim N(\beta_{a,s}, \frac{w_{a,s}}{P_0^2}\sigma_a^2), \text{ eller } \frac{P_t}{P_0} \sim G\left(\alpha_a, \frac{\sqrt{w_{a,s}}}{P_0}, (\beta_{a,s} - \frac{\sqrt{w_{a,s}}}{P_0}\alpha_a) \right)$$

mens estimatet blir som i (5.10):

$$(5.16) \qquad \log(\hat{\beta}_{m,j}) = \sum_i \left(\frac{\frac{1}{w_{m,j}}}{\sum_i \frac{1}{w_{m,j}}} \right) \log(\frac{P_t}{P_0}) \quad \Leftrightarrow \quad \hat{\beta}_{m,j} = \prod_i \left(\frac{P_t}{P_0} \right)^{\frac{\frac{1}{w_{m,j}}}{\sum_i \frac{1}{w_{m,j}}}}, \quad j = 0,1,2$$

Modellen (5.15) er definert for negative $\frac{P_t}{P_0}$, men det er ikke estimatet (5.16). Dette gjør at det er umulig å finne et analytisk uttrykk for forventning og varians til estimatet. Det tydelig-gjør også modellens uegnethet siden det fra et tolkningssynspunkt bør være umulig med negative pris-forhold.

I praksis løser simulering problemet med å finne forventning og varians til estimatet, selv om man ikke har analytiske uttrykk. Varenumrene som genererer negative prisforhold kuttes ut. Vi skal se at dette problemet oppstår kun ved normal-fordelt støy, noe som tyder på at støy som er fordelt med skjevhet mot høyre fungerer bedre for disse modellene, og stemmer bedre med fysisk tolkning. Modellen med normal-fordelt støy kan også være egnet, men først og fremst med såpass liten varians på støyen at ingen negative priser genereres.

5.2.2. Multiplikativ

Modellen er nå som i (5.9), og på opprinnelig skala gjelder:

$$(5.17) \qquad \frac{P_t}{P_0} \sim \log N\left(\log(\beta_{m,s}), w_{m,s}\sigma_m^2 \right), \text{ eller } \frac{P_t}{P_0} \sim \log G\left(\alpha_m, \sqrt{w_{m,s}}, \left[\log(\beta_{m,s}) - \sqrt{w_{m,s}}\alpha_m \right] \right)$$

altså at prisforholdet er lognormal-fordelt eller loggamma-fordelt. Med estimatet fra (5.6):

(5.18) $$\hat{\beta}_{a,j} = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right) \frac{P_t}{P_0}, \quad j = 0,1,2$$

ser man at med lognormal-fordelt prisforhold fås at estimatet er en veiet sum av lognormal-fordelinger og med forventning og varians:

(5.19) $$E\left(\hat{\beta}_{a,j}\right) = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right) \exp\left(\log(\beta_{m,s}) + \tfrac{1}{2} w_{m,s} \sigma_m^2 \right), \quad j = 0,1,2$$

(5.20) $$Var\left(\hat{\beta}_{a,j}\right) = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right) \left\{ \exp\left(2\log(\beta_{m,s}) + 2w_{m,s}\sigma_m^2 \right) - \exp\left(2\log(\beta_{m,s}) + w_{m,s}\sigma_m^2 \right) \right\}$$

Med loggamma-fordelt prisforhold fås at estimatet er en tilsvarende sum av loggamma-fordelinger og har forventning og varians:

(5.21) $$E\left(\hat{\beta}_{a,j}\right) = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right) \left\{ \left(\frac{1}{1-\sqrt{w_{m,s}}} \right)^{\alpha_m} \exp\left(\log(\beta_{m,s}) - \sqrt{w_{m,s}}\alpha_m \right) \right\}, \quad \sqrt{w_{m,s}} < 1, \; j = 0,1,2$$

(5.22)

$$Var\left(\hat{\beta}_{a,j}\right) = \sum_i \left(\frac{\frac{P_0^2}{w_{a,j}}}{\sum_i \frac{P_0^2}{w_{a,j}}} \right)^2 \left\{ \left[\left(\frac{1}{1-2\sqrt{w_{m,s}}} \right)^{\alpha_m} - \left(\frac{1}{1-\sqrt{w_{m,s}}} \right)^{2\alpha_m} \right] \exp\left(2\log(\beta_{m,s}) - 2\sqrt{w_{m,s}}\alpha_m \right) \right\}, \quad \sqrt{w_{m,s}} < \tfrac{1}{2}$$

Disse estimeres i neste avsnitt ved simulering.

6. Resultater

6.1. Sann modell

Her estimerer vi under den riktige modellen, dvs. at vi bruker de optimale estimatene. Dette er situasjonen beskrevet teoretisk i avsnitt 5.1. Vi regner ut skjevhet, standard-feil og RMSE fra uttrykk (5.2), (5.3), og (5.4) for å vurdere hvor god estimeringen er. Den additive og den multiplikative modellen er begge tatt med i samme figur for sammenligning. Først vises figurene for normal-fordelt / lognormal-fordelt støy, deretter for gamma-fordelt / loggamma-fordelt støy.

Simuleringen tar utgangspunkt i modell (5.5) og (5.9) og erstatter ukjente størrelser med estimater (parameterne med indeks "s"). Dette gjør det mulig å trekke fra kjente fordelinger.

Først regner vi ut skjevhet etter (5.2), f.eks. for additiv modell. Da blir uttrykket for skjevhet, for hvert varenr:

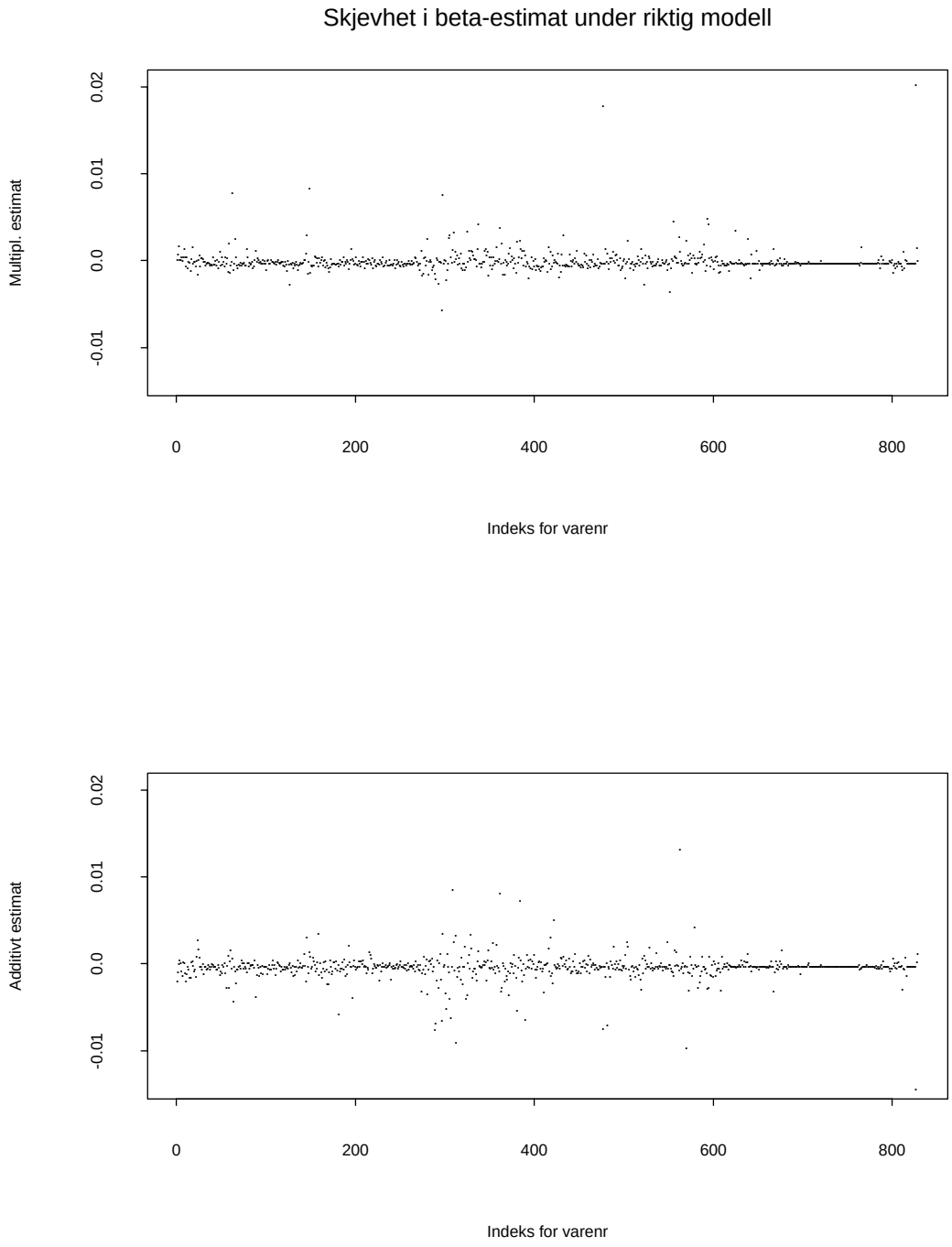
$$\text{bias}(\hat{\beta}_a) = \tfrac{1}{K}\sum_{k=1}^{K} \text{sign}\left[\min_j \left|\hat{\beta}_{a,j}^k - \hat{\beta}_{a,s}\right|\right] \tag{6.1}$$

Dette gir oss et estimat på $E(\hat{\beta}_{a,j} - \beta_{a,s})$ med gjennomsnitt over mange simuleringer i stedet for forventning, og $\hat{\beta}_{a,s}$ i stedet for $\beta_{a,s}$.

Bakgrunnen for å velge den minste skjevheten (over $j = 0,1,2$) i (6.1), er at i praksis har vi en metode for å finne optimal "w", da må dette gjelde for de simulerte dataene også. For hver "k" i simuleringen velges altså den $\hat{\beta}_{a,j}^k$ (over $j = 0,1,2$) som gir minst absolutt avvik $\left|\hat{\beta}_{a,j}^k - \hat{\beta}_{a,s}\right|$ (differansen kan være negativ og fortegnet beholdes). Det er ikke nødvendigvis den samme "j" for hver "k", men ved å minimere for hver "k" blir også total-summen (og gj.snittet) minimert. Dette måler altså skjevheten i estimatet under optimal "w".

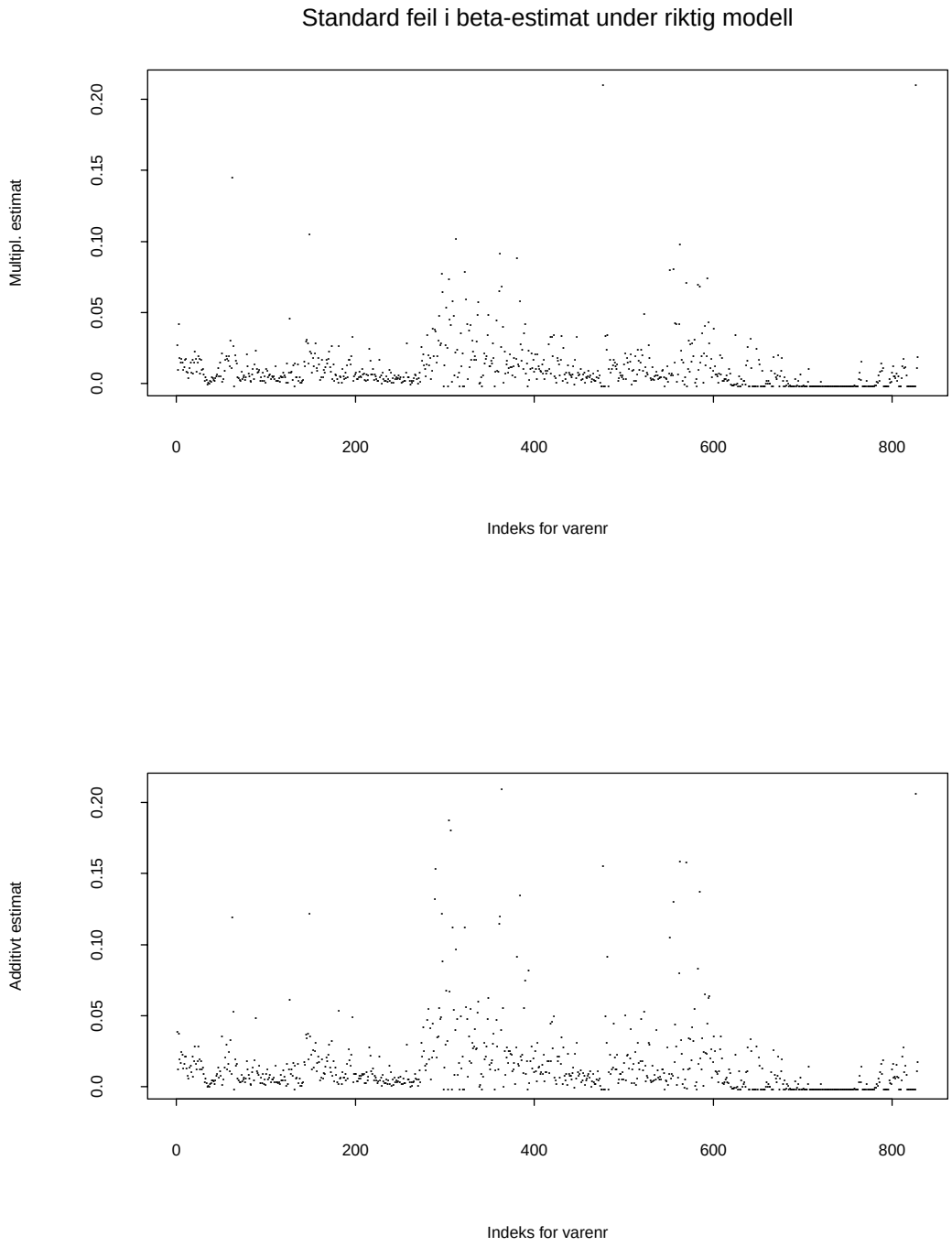
Tilsvarende beregnes standard feil (roten av (5.4)) og RMSE (roten av (5.3).

Siden figurene viser begge modellene og på samme målestokk er det enkelt å sammenligne. Hver prikk i figurene representerer et varenr og verdien har fremkommet ved 500 simuleringer.



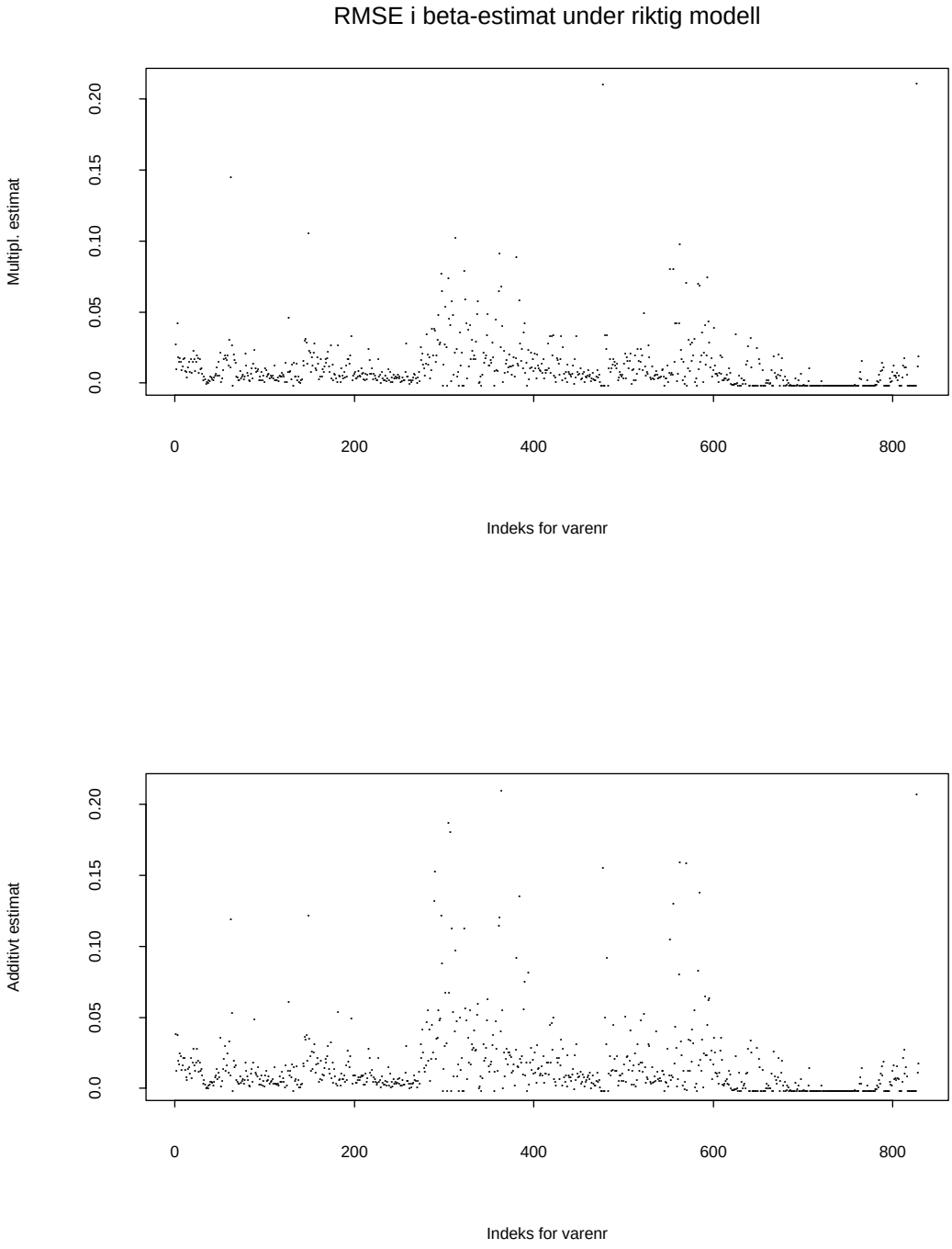
Figur 3: Sammenligning av skjevhet, riktig modell og normal- /lognormal-fordeling

Figuren viser ingen forskjell mellom modellene, og skjevheten er liten, noe vi kunne forvente fra uttrykkene (5.7) og (5.13).



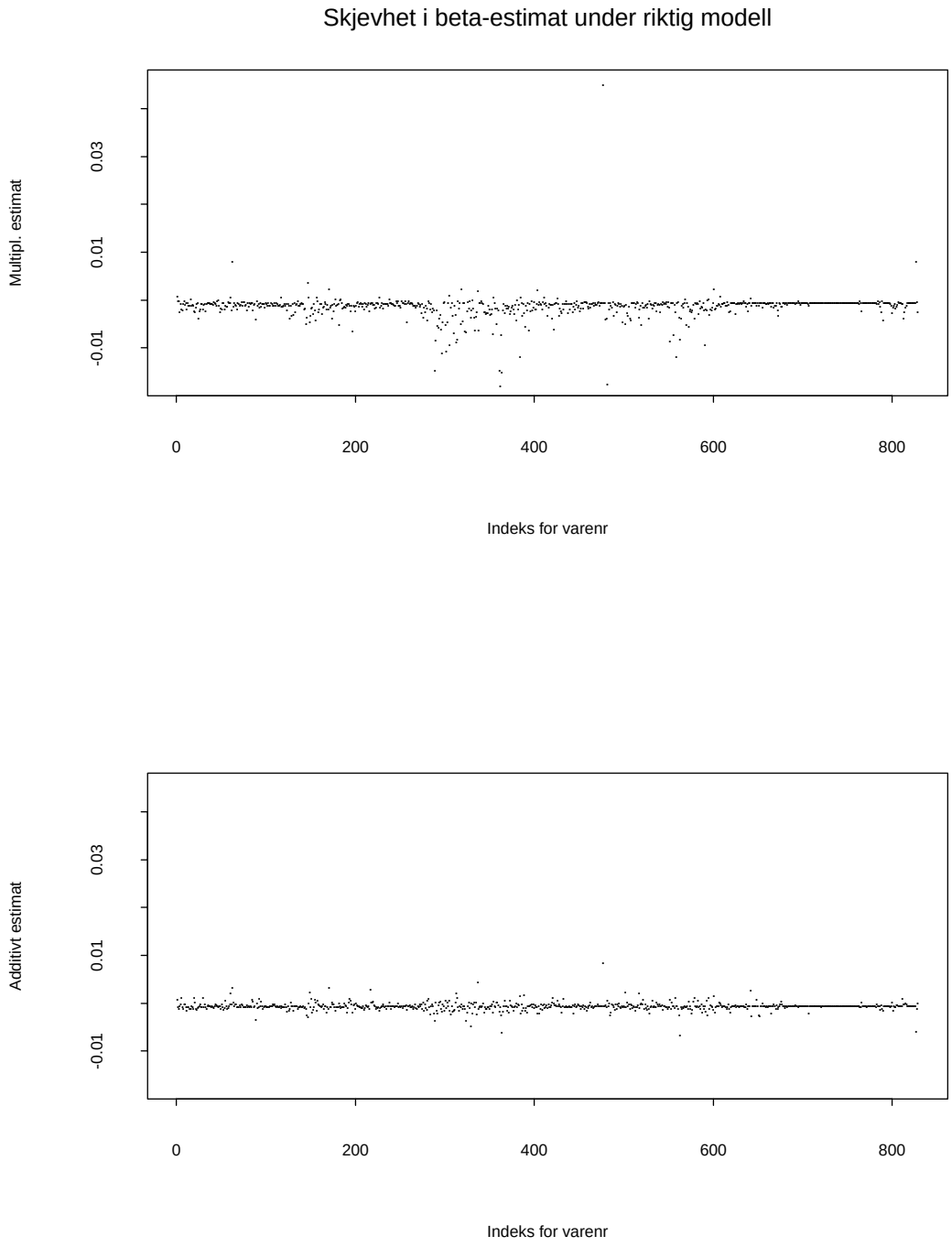
Figur 4: Sammenligning av standard-feil, riktig modell og normal- / lognormal-fordeling

Figuren viser ubetydelig forskjell i standard-feil mellom de to modellene. Videre ses at nivået på standard-feilen er lavt i forhold til verdien på parameter-estimatet (rundt 1), som betyr god estimering. Standard-avviket til bias-estimatet (forrige figur) kan beregnes ved å ta ca. 5 % av nivået på standard-feilen i denne figuren.



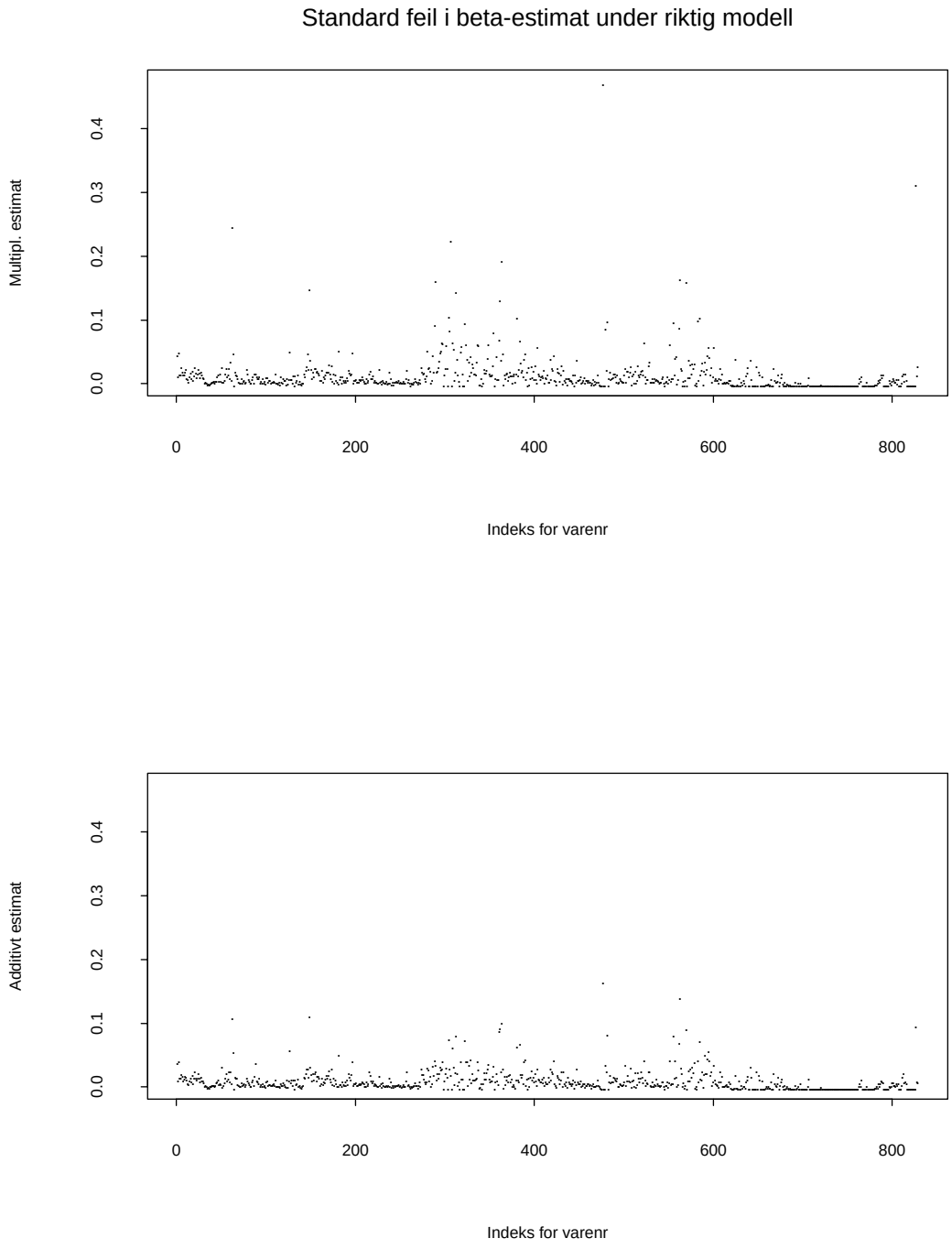
Figur 5: Sammenligning av RMSE, riktig modell og normal- / lognormal-fordeling

Siden nivået på skjevheten er lite i forhold til standard-feilen vil bildet av RMSE være dominert av bidraget fra standard-feilen og dette ser vi i figur 5. Figuren viser også at forskjellen mellom modellene er liten.



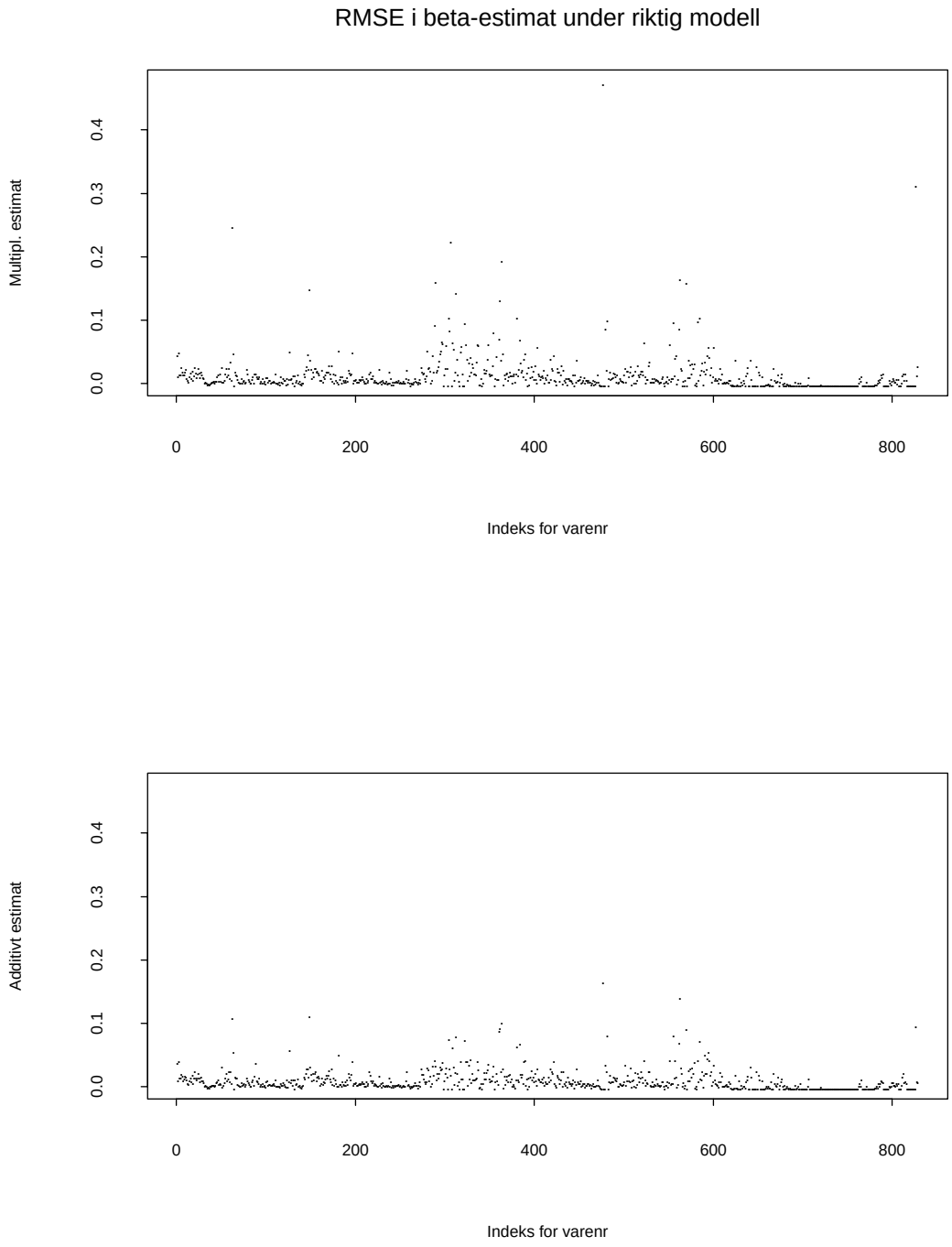
Figur 6: Sammenligning av skjevhet, riktig modell og gamma- / loggamma-fordeling

Med usymmetrisk støy, er fremdeles forskjellen mellom modellene liten (kanskje litt bedre estimering i den additive modellen) og skjevheten er liten. Skalaen går litt høyere pga. et varenummer i det multiplikative estimatet som har skjevhet på over 0.04. I det multiplikative estimatet er skjevheten nå i større grad negativ enn i figur 3.



Figur 7: Sammenligning av standard-feil, riktig modell og gamma- / loggamma-fordeling

Figuren viser lav standard-feil i estimatene med usymmetrisk støy også, med litt bedre estimering i det additive estimatet, og litt større standard-feil generelt enn for symmetrisk støy (figur 4).



Figur 8: Sammenligning av RMSE, riktig modell og gamma- / loggamma-fordeling

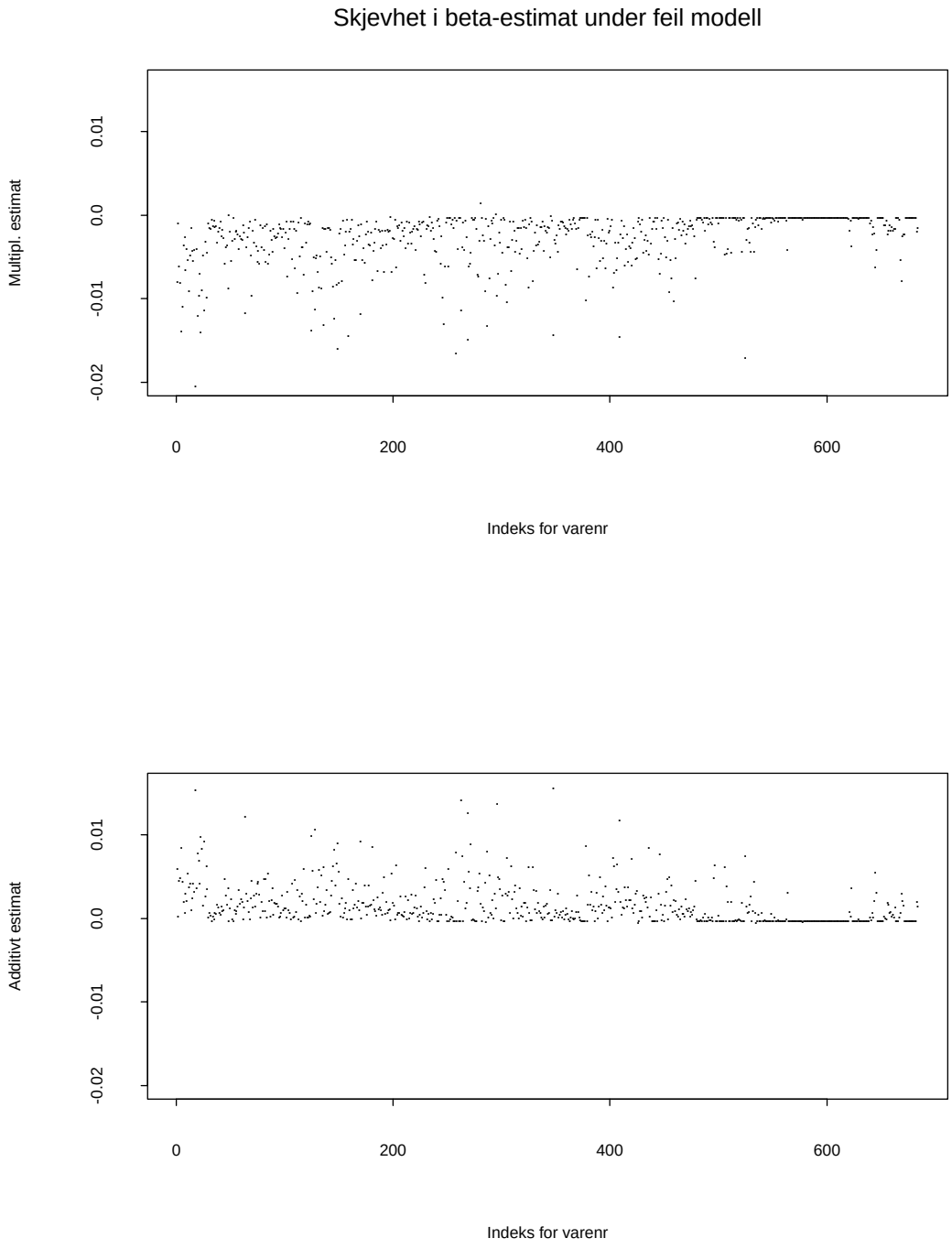
Bildet fra standard-feilen i forrige figur gjentas her. RMSE er dominert av standard-feilen (samme skala). Figuren viser litt bedre estimering for additivt estimat enn for multiplikativt, ved noen varenummer, og litt dårligere generelt her enn ved symmetrisk støy (figur 5).

6.2. Feil modell

Her er situasjonen som i avsnitt 5.2. Vi trekker for øvrig fra samme fordelingene som i avsnitt 6.1 og følger presentasjonen derfra. Uttrykket for skjevhet blir nå:

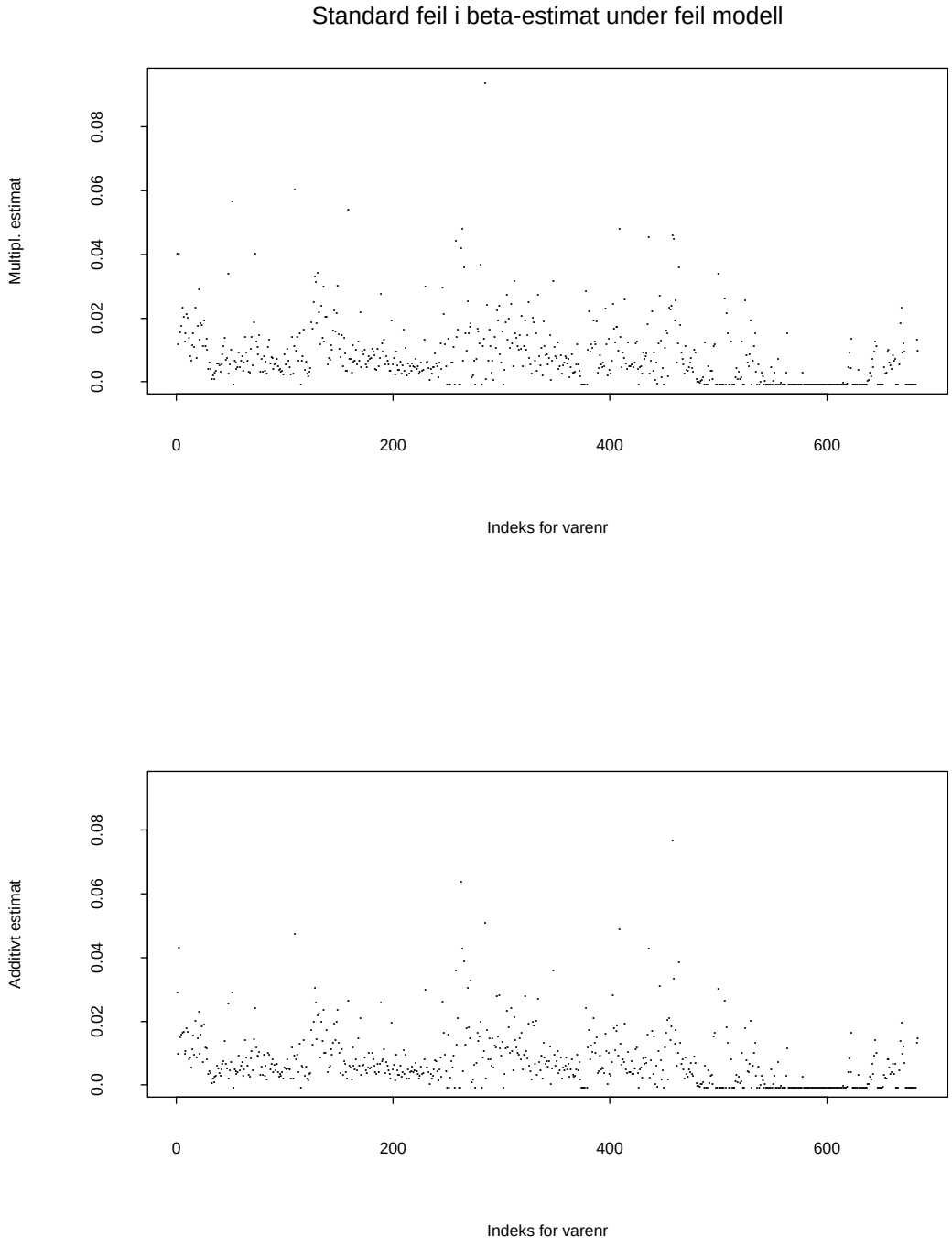
(6.2) $$\text{bias}(\hat{\beta}_a) = \frac{1}{K}\sum_{k=1}^{K} \text{sign}\left[\min_j \left|\hat{\beta}^k_{a,j} - \hat{\beta}_{m,s}\right|\right]$$

(byttet ut "a" med "m" i modell-estimatet), slik at dette estimerer skjevheten i det additive estimatet under den multiplikative modellen $E(\hat{\beta}_{a,j} - \beta_{m,s})$.



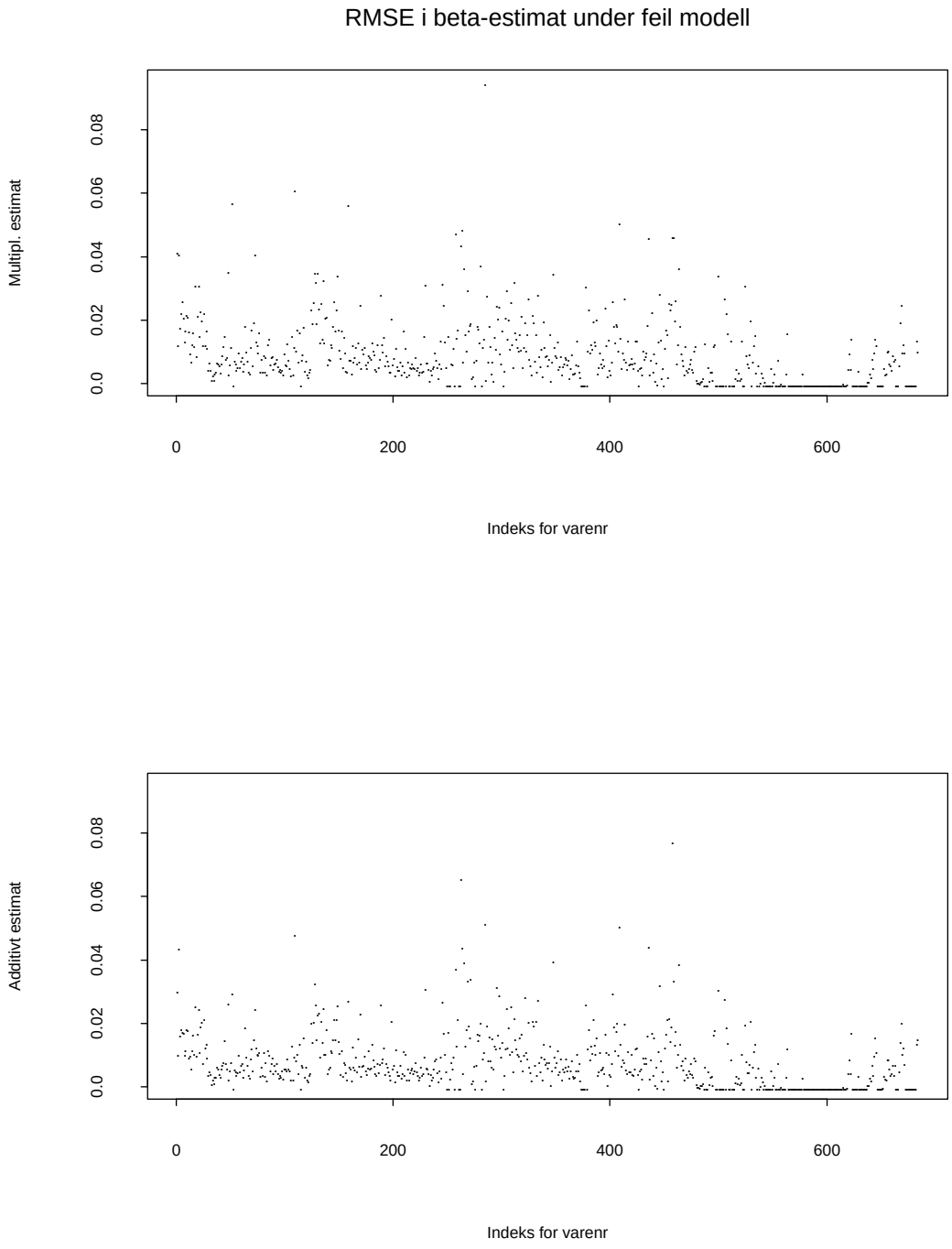
Figur 9: Sammenligning av skjevhet, feil modell og normal- / lognormal-fordeling

Figuren viser at det ikke er forskjell i absoluttnivå av skjevhet mellom modellene, men fortegn. Forskjellen i fortegn kommer av at det geometriske gjennomsnittet oftest er lavere enn det aritmetiske (se appendix). Nivået i seg selv er lavt her også, under feil modell (kanskje litt større enn under riktig modell, figur 3). Det er verdt å merke seg at antall varenummer her er kun 684 (i motsetning til 829). Dette er pga. at variansen i 145 varenummer er så stor at prisforholdene i modellen (5.15) blir negative og dermed gir ulovlige verdier til det multiplikative estimatet i 5.16.



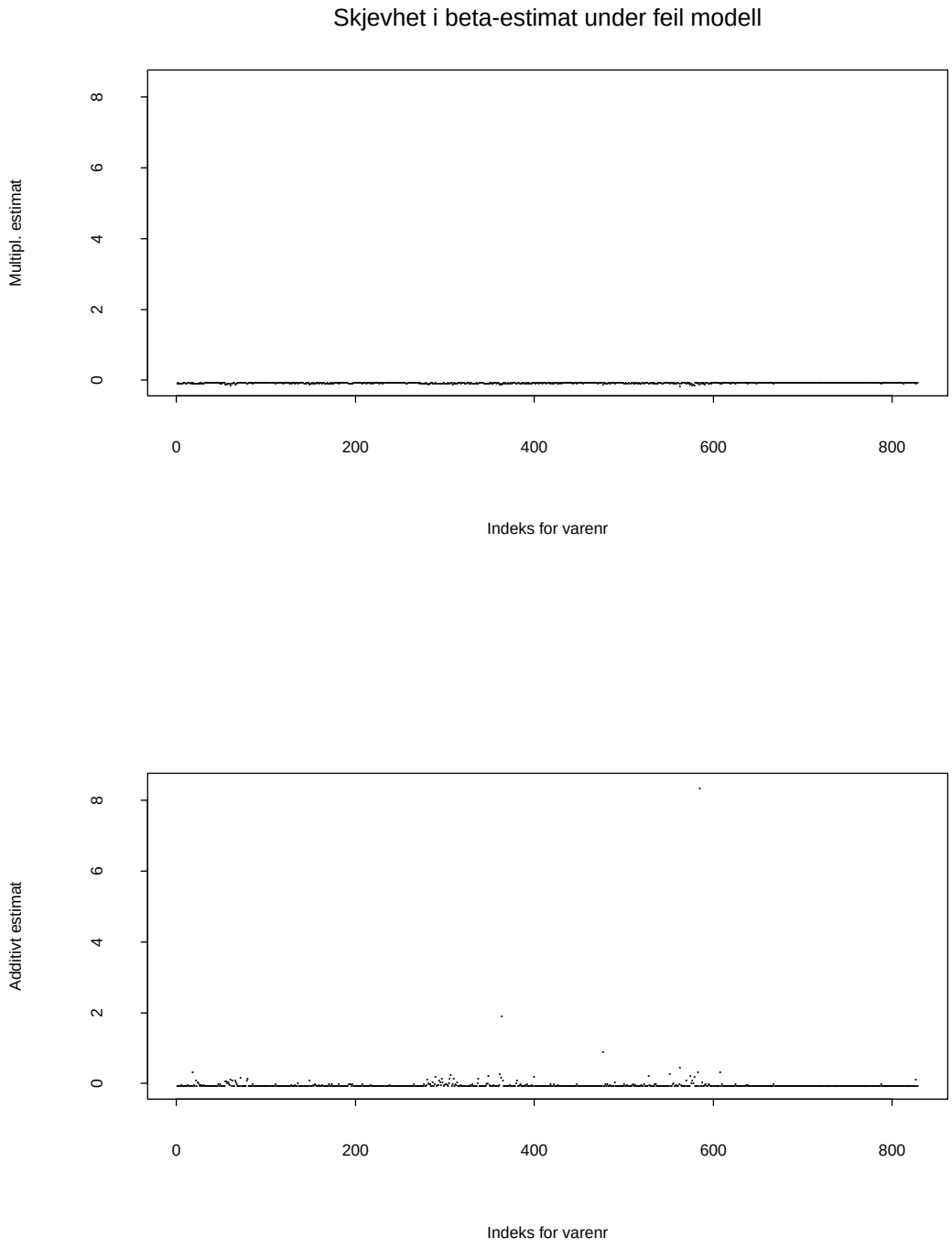
Figur 10: Sammenligning av standard feil, feil modell og normal- / lognormal-fordeling

Figuren viser ingen forskjell i standard-feil til estimatene under de to gale modellene. Antall varenummer er 684, og dette er årsaken til at nivået på standard-feilen er såpass lavt, i forhold til estimering under riktig modell (figur 4).



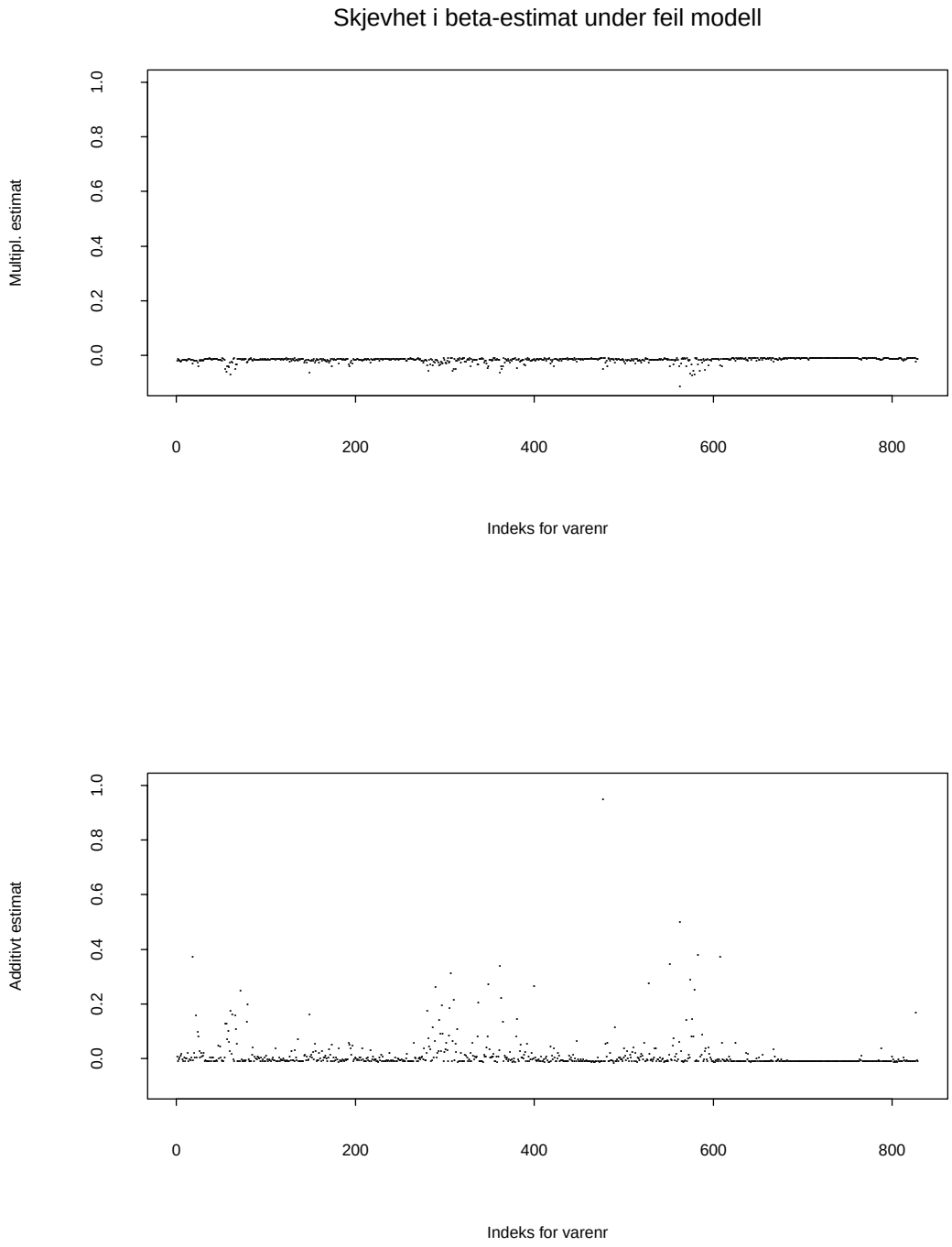
Figur 11: Sammenligning av RMSE, feil modell og normal- / lognormal-fordeling

Som i figur 10 er det ingen forskjell mellom modellene med hensyn til RMSE under feil modell, og nivået er dominert av standard-feilen. Det er også mindre enn under riktig modell (figur 5) pga. lavt varenummer-antall (684).



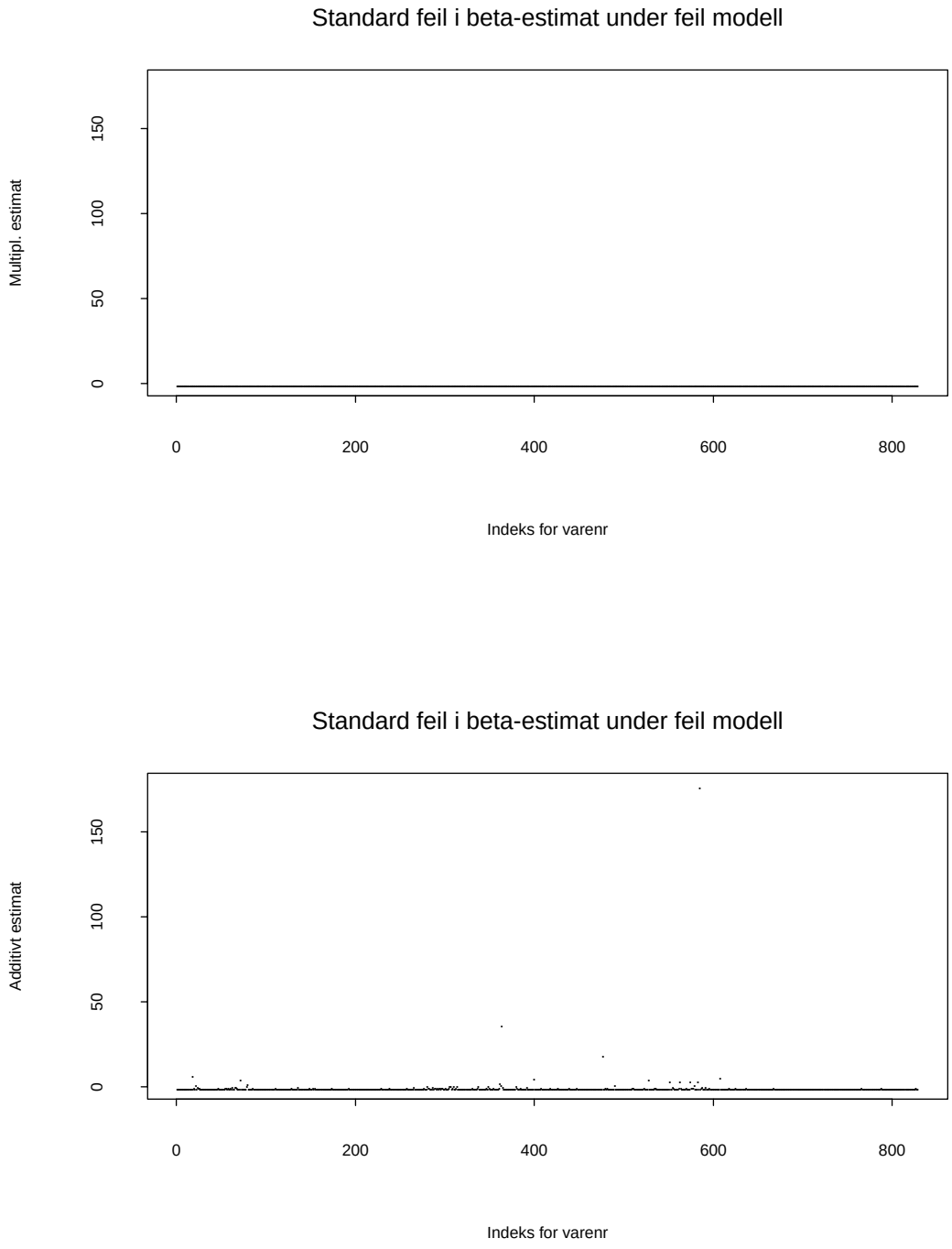
Figur 12: Sammenligning av skjevhet, feil modell og gamma- / loggamma-fordeling

Figuren viser dramatisk forskjell i skjevhet ved estimering under feil modell når støyen er usymmetrisk. Det multiplikative estimatet har tydelig minst skjevhet. Skjevheten i det additive estimatet er på over 8 for et varenummer, noe som selvfølgelig gjør estimatet ubrukelig med tanke på nivået i β_t selv som er rundt 1. Neste side er samme figuren presentert med en annen målestokk.



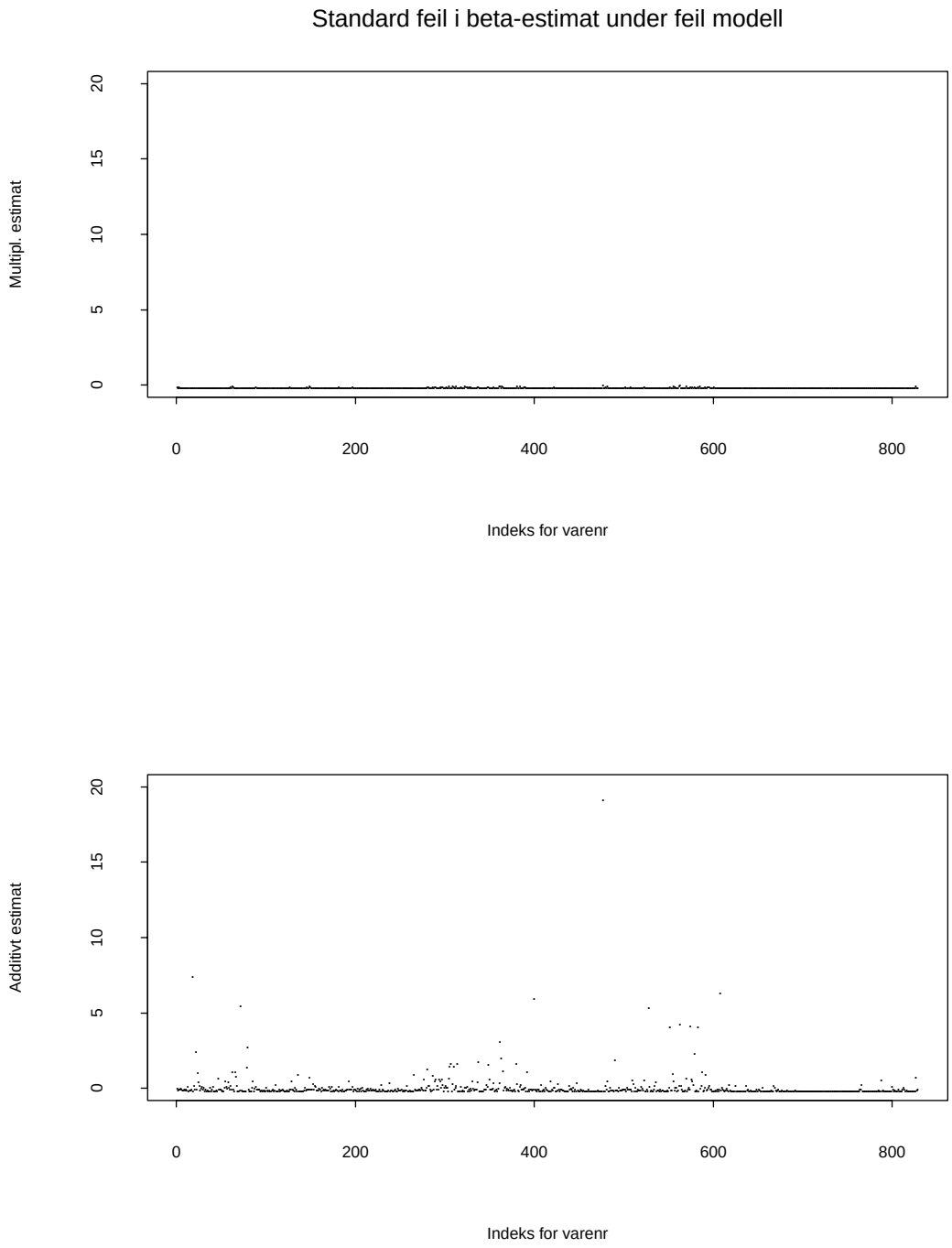
Figur 13: Som figur 12, annen målestokk

Denne figuren er som figuren på forrige side, men med de største verdiene fjernet for å gi et mer detaljert bilde. Den viser at de fleste varenumrene har i det additive estimatet skjevhet under 1, men likevel betydelig større enn i det multiplikative estimatet, og dessuten betydelig større enn i de andre situasjonene (sann modell med symmetrisk og usymmetrisk støy, samt feil modell med symmetrisk støy).



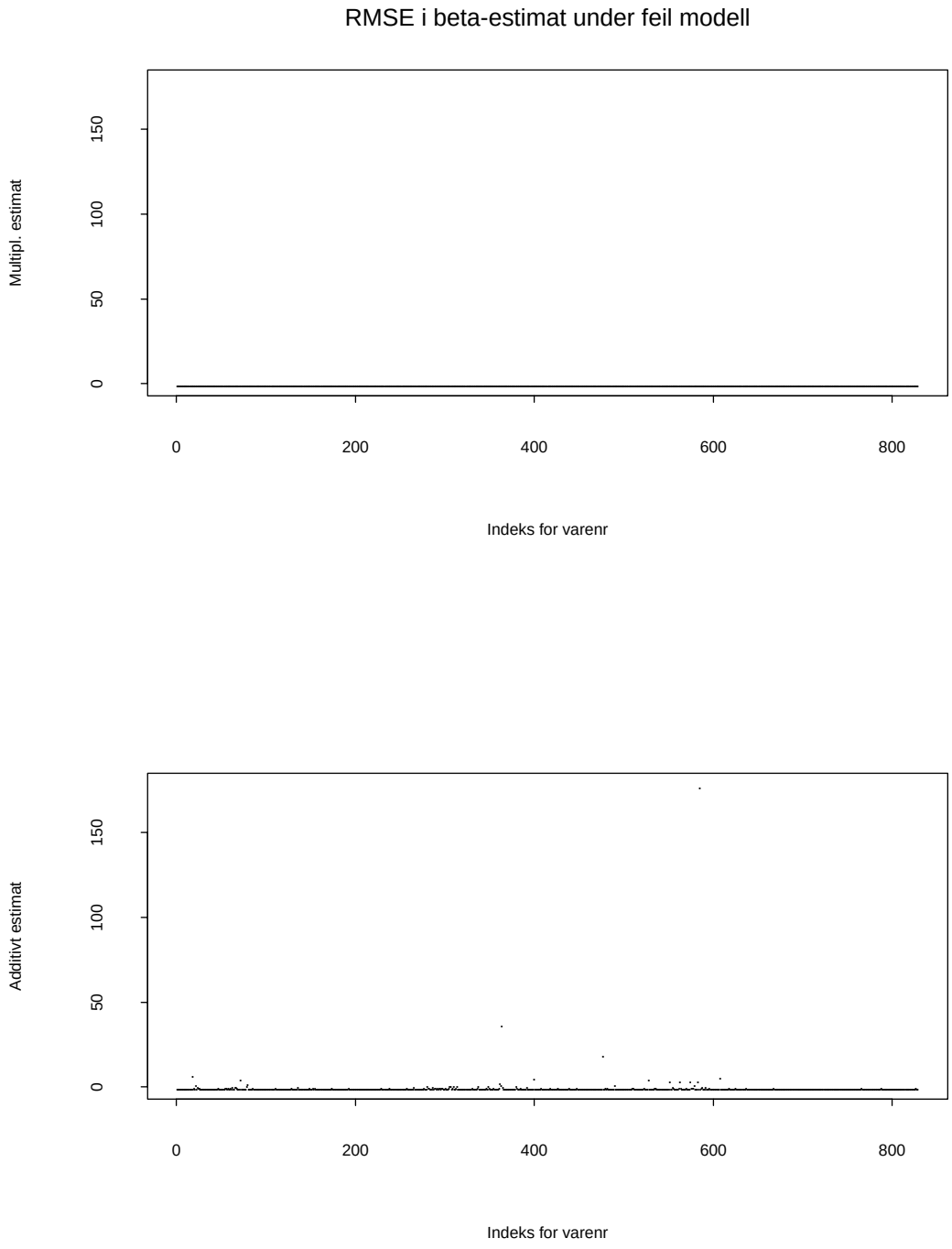
Figur 14: Sammenligning av standard-feil, feil modell og gamma- / loggamma-fordeling

Denne figuren viser det som figur 12 at for noen varenummer i det additive estimatet er standard-feilen dramatisk mye større enn i det multiplikative estimatet. Neste side viser samme figur med annen målestokk.



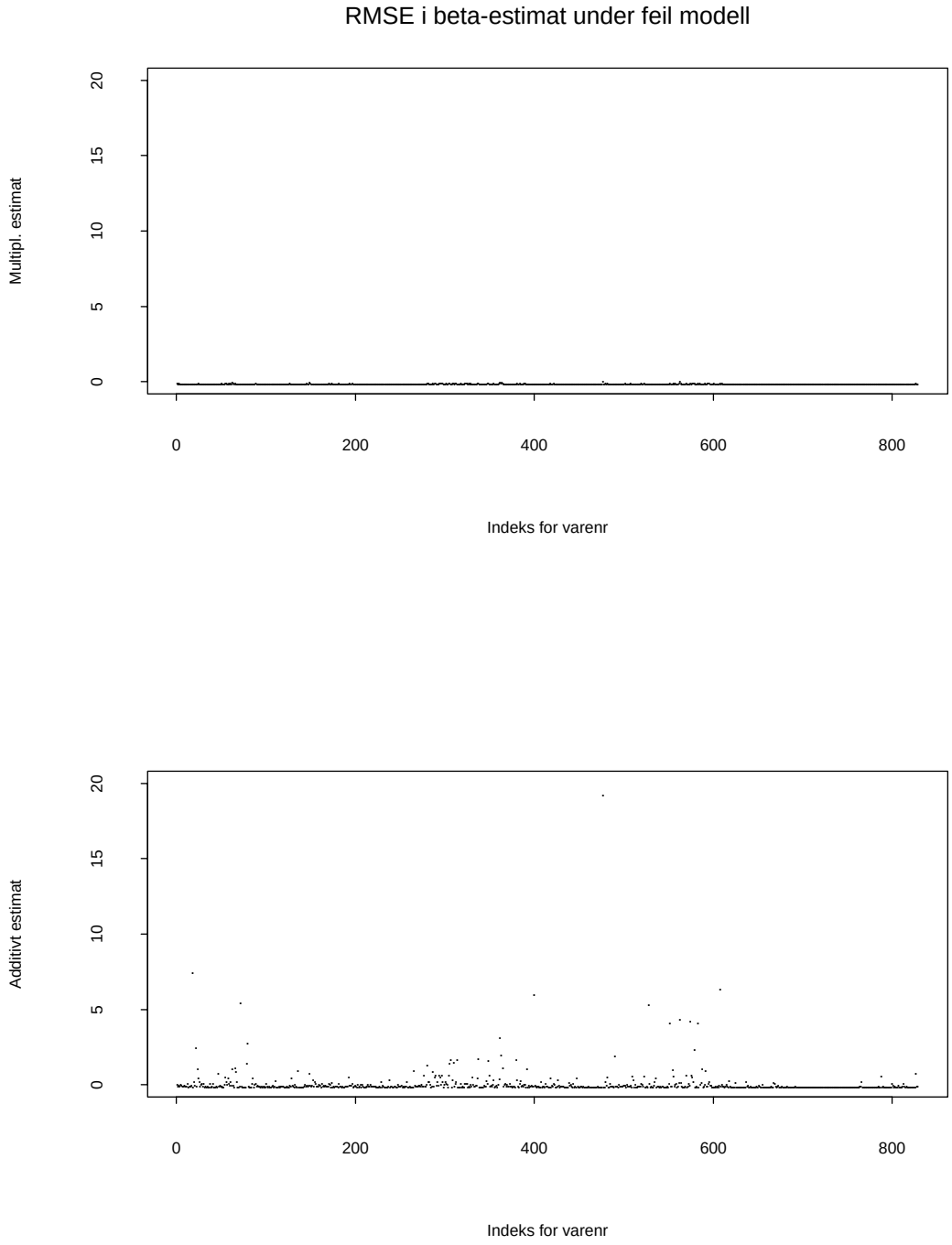
Figur 15: Som figur 14, annen målestokk

Med denne målestokken er det tydelig at standard-feilen for de fleste varenumrene er betydelig større i det additive estimatet enn i det multiplikative estimatet, og større enn i noen av de andre situasjonene.



Figur 16: Sammenligning av RMSE, feil modell og gamma- / loggamma-fordeling

Som for skjevheten og standard-feilen er det også noen varenummer som har dramatisk mye større RMSE i det additive estimatet enn i det multiplikative estimatet. Nivået blir dominert av standard-feilen. Neste side er figuren vist med annen målestokk.



Figur 17: Som figur 16, annen målestokk

Denne figuren viser at bildet av RMSE ligner på bildet av standard-feilen. Den viser at nivået i RMSE er dominert av standard-feilen og at jevnt over har det additive estimatet betydelig høyere RMSE enn det multiplikative estimatet.

6.3. Konklusjon

Modellene i (2.1) og (3.1) er identiske hvis man fjerner støy-leddet. Ved å legge til støy med forventning 0 i den additive modellen og multiplisere med støy som har forventning nær 1 i den multiplikative modellen fås to modeller hvor den tilfeldige komponentens innvirkning er det som skiller dem. Ligning (3.2) viser at den multiplikative modellen er en additiv modell på log-skala. Ved å sammenligne den additive modellen på formen $\frac{P_{t,i}}{P_{0,i}} = \beta_t + \frac{1}{P_{0,i}}\varepsilon_i$ og den multiplikative på formen $\log(\frac{P_{t,i}}{P_{0,i}}) = \log(\beta_t) + \log(\varepsilon_i)$ ser man at med forutsetningene fra (2.2) og (3.3) samt våre valg av vekter er de begge lineære regresjons-modeller med støy ledd ($\frac{1}{P_{0,i}}\varepsilon_i$ og $\log(\varepsilon_i)$) som har samme uttrykk for forventning og varians. Den multiplikative modellen kan altså betraktes som en lineær (additiv) regresjons-modell etter å ha transformert dataene med logaritmen og ved å gjøre støyfordelings-antagelser på de log-transformerte dataene. Estimeringen ("minste-kvadrater") utføres på de log-transformerte dataene, hvor variansen er mindre, og gir estimat av $\log(\beta_t)$ (egentlig $\log(\beta_t)$ og ikke $\log(\hat{\beta}_t)$ som i teksten) som så kan tilbake-transformeres for tolkning på opprinnelig skala.

I tilfeller med liten varians og symmetri i støyen (på additiv skala) er modellene i praksis like og estimatene $\hat{\beta}_a$ og $\hat{\beta}_m$ som utledes fra modellene blir derfor også like. Dette ser vi i figur 9 som viser at differansen mellom $\hat{\beta}_a$ og $\hat{\beta}_m$ er ubetydelig (under begge modellene) og like liten som skjevheten under sann modell i figur 3. Skjevheten under sann modell i figur 3 er liten i begge modellene, som viser at estimatene er gode. Når variansen er større vil den additive modellen generere negative pris-forhold som er urealistisk.

Når støyen er usymmetrisk oppstår forskjeller mellom estimatene. Figurene 6 og 12 viser dette. $\hat{\beta}_m$ har omtrent like liten skjevhet uansett om modellen er multiplikativ (figur 6), eller modellen er additiv (figur 12), og nivået er dessuten ikke noe særlig større enn for symmetrisk støy som bekrefter god estimering. $\hat{\beta}_a$ derimot, oppfører seg anderledes. Under additiv modell (figur 6) fungerer den bra (litt bedre enn $\hat{\beta}_m$), mens under multiplikativ modell (figur 12) bryter den sammen og gir stor skjevhet. Dette kan oppsummeres i flg. tabell, basert på figurene.

	ADDITIVT ESTIMAT $\hat{\beta}_a$		MULTIPLIKATIVT ESTIMAT $\hat{\beta}_m$	
	Sann modell	Feil modell (*)	Sann modell	Feil modell (*)
$\lvert$Skjevhet$\rvert$	≤ 0.009	≤ 8.428 (37.39 %)	≤ 0.046	≤ 0.103 (1.09 %)
Standard-feil	≤ 0.168	≤ 177.5 (21.35 %)	≤ 0.473	≤ 0.175 (0 %)
RMSE	≤ 0.168	≤ 177.7 (21.59 %)	≤ 0.476	≤ 0.188 (0 %)

(*): Andel varenummer (av 829) som har nivå større enn max-nivået i sann modell

Tabell 1: Usymmetrisk støy gir forskjell på $\hat{\beta}_a$ og $\hat{\beta}_m$

Dette viser at det veiede geometriske gjennomsnittet er mer robust overfor modell og fordeling på støy enn det veiede aritmetiske gjennomsnittet. Hvorvidt usymmetrisk støy er realistisk er usikkert.

6.4. Referanse

[1]: Johnson, N. L., Kotz, S., "Continuous Univariate Distributions - I" John Wiley & Sons 1970.

[2]: Ghangurde, P.D. (1989). Outliers in sample surveys. In "Proceedings for the Survey Research Methods Section", American Statistical Association, pp. 736-739.

[3]: Johanessen, Randi. "Valg av mikroindeksformel i konsumprisindeksen", kommer i SSB-serien Notater

7. Appendix

7.1. Normal- / Lognormal-fordeling

En tilfeldig variabel U er standard normal-fordelt hvis den har tetthet $p_U(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. Forventning og varians er gitt ved $E(U) = 0$ og $Var(U) = 1$. La $Z = \sigma U + \mu$. Da vil Z ha tetthet $q_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2}$ med betegnelsen $Z \sim N(\mu, \sigma^2)$. Forventning og varians til Z er $E(Z) = \mu$ og $Var(Z) = \sigma^2$.

En tilfeldig variabel er lognormal-fordelt hvis logaritmen til den er normal-fordelt. Med utgangspunkt i normalfordelingen fremkommer lognormal-fordelingen ved transformasjonen $Y = e^Z$. Y har tetthet $f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma y} e^{-\frac{1}{2}(\frac{\log(y)-\mu}{\sigma})^2}$, $y > 0$. Eventuelt kan y byttes ut med $y - \lambda$, for en ekstra lokasjonsparameter og slik at $y > \lambda$ er definisjonsområdet. Forventning og varians til Y (med $\lambda = 0$) er $E(Y) = e^{\mu + \frac{1}{2}\sigma^2}$ og $Var(Y) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$. Lognormal-fordelingen har moderat tung hale, tyngre enn normal-fordelingen og gamma-fordelingen (se nedenfor), men lettere enn loggamma-fordelingen. Vi ser at forventning og varians eksisterer for alle endelige verdier av parametrene.

7.2. Gamma- / Loggamma-fordeling

En tilfeldig variabel U er standard gamma-fordelt med parameter α hvis den har tetthet $p_U(u) = \frac{u^{\alpha-1}e^{-u}}{\Gamma(\alpha)}$, med $\alpha > 0, u \geq 0$. Forventning og varians er gitt ved $E(U) = Var(U) = \alpha$. La $Z = \theta U + \gamma$. Da vil Z være tre-parameter gamma-fordelt med tetthet $q_Z(z) = \frac{(z-\gamma)^{\alpha-1} e^{-\frac{1}{\theta}(z-\gamma)}}{\theta^\alpha \Gamma(\alpha)}$, der $\alpha > 0, \theta > 0, z > \gamma$, med betegnelsen $Z \sim G(\alpha, \theta, \gamma)$. Forventning og varians blir da $E(Z) = \theta\alpha + \gamma$ og $Var(Z) = \theta^2\alpha$ (se for øvrig [1]).

Med utgangspunkt i $Z \sim G(\alpha, \theta, \gamma)$ ovenfra får man den loggamma-fordelte variabelen Y ved transformasjon $Y = e^Z$. Y har da tettheten $f_Y(y) = \frac{1}{\Gamma(\alpha)}(\frac{1}{\theta})^\alpha \left[\log(y) - \gamma\right]^{\alpha-1} e^{\frac{\gamma}{\theta}} y^{-\frac{1}{\theta}-1}$, der $\alpha > 0, \theta > 0, y \geq e^\gamma$. Eventuelt kan y i uttrykket for tettheten erstattes med $y - \lambda$ (for en ekstra lokasjonsparameter slik at $y \geq \lambda + e^\gamma$). Momentene utledes ved å omforme uttrykket for tettheten, slik at $E(Y) = \int_{e^\gamma}^{\infty} \frac{1}{\Gamma(\alpha)}(\frac{1}{\theta})^\alpha \left[\log(y) - \gamma\right]^{\alpha-1} e^{\frac{\gamma}{\theta}} y^{-\frac{1}{\theta}-1} y \, dy = \left(\frac{\frac{1}{\theta}}{\frac{1}{\theta}-1}\right)^\alpha e^\gamma \int_{e^\gamma}^{\infty} \frac{(\frac{1}{\theta}-1)^\alpha}{\Gamma(\alpha)} \left[\log(y) - \gamma\right]^{\alpha-1} e^{(\frac{1}{\theta}-1)\gamma} y^{-(\frac{1}{\theta}-1)-1} \, dy$ der integralet blir 1 (siden det er en tetthet) og man får $E(Y) = \left(\frac{\frac{1}{\theta}}{\frac{1}{\theta}-1}\right)^\alpha e^\gamma = \left(\frac{1}{1-\theta}\right)^\alpha e^\gamma, \quad \theta < 1$ og på tilsvarende måte kan utlede at $Var(Y) = \left[(\frac{1}{1-2\theta})^\alpha - (\frac{1}{1-\theta})^{2\alpha}\right] e^{2\gamma}, \quad \theta < \frac{1}{2}$. Denne fordelingen har så tung hale at max-verdien i sampler fra denne fordelingen tilhører den klassen (av tre mulige) som har tyngst hale, med sin potens-avtagende hale (veldig sein i forhold til eksponentielt avtagende). Av uttrykket for forventningen ser man at den ikke eksisterer for $\theta > 1$, nettopp fordi halen er for tung. Tilsvarende for variansen for $\theta > \frac{1}{2}$.

7.3. Sammenligning av aritmetisk og geometrisk gjennomsnitt

Figur 9 i resultatavsnittet viser skjevhet i estimatet under feil modell. Det er tydelig forskjell på fortegnet i de to forskjellige tilfellene og dette illustrerer at det aritmetiske gjennomsnittet i nesten alle tilfeller er det største. Men figuren viser også enkelte tilfeller på det motsatte. Denne sammenhengen undersøkes her.

Teorem 1 (like vekter):

La $u_1,\cdots,u_n$ og $x_1,\cdots,x_n$ være slik at $0<u_i<1,\quad x_i>0$ for $i=1,\cdots,n$, $\sum_{i=1}^{n}u_i=1$ og $x_i\neq x_j$ for minst en $i\neq j$. Da gjelder:

$$\sum_{i=1}^{n}u_i x_i > \prod_{i=1}^{n}x_i^{u_i} \tag{7.1}$$

(Hvis alle x_i-ene er like er det likhet).

Bevis:
Anta først at $n=2$. Da gjelder det å vise at $u_1x_1+u_2x_2>x_1^{u_1}x_2^{u_2}$ når $x_1,x_2>0$ og $x_1\neq x_2$, $u_1,u_2\in(0,1)$ og $u_1+u_2=1$. Siden x-ene er forskjellige finnes en $c>0,\neq 1$ s.a. $x_2=cx_1$. Det som skal vises er altså at: $(1-u_2)x_1+u_2cx_1=(1-u_2+u_2c)x_1>x_1^{1-u_2}(cx_1)^{u_2}=c^{u_2}x_1$ som igjen er ekvivalent med å vise at :

$$1-u_2+u_2c-c^{u_2}>0 \quad \text{for} \quad u_2\in(0,1) \quad \text{og} \quad c>0,\neq 1 \tag{7.2}$$

La $f(y)=ay-y^a,\quad$ for $\quad a\in(0,1)\quad$ og $\quad y\geq 0$. Da er

$$f(y)=\begin{cases} 0 & \text{for} \quad y=0 \quad \text{og} \quad y=(\frac{1}{a})^{\frac{1}{1-a}} \\ <0 & \text{for} \quad y\in\left(0,(\frac{1}{a})^{\frac{1}{1-a}}\right) \\ >0 & \text{for} \quad y>(\frac{1}{a})^{\frac{1}{1-a}} \end{cases}$$

Videre har $f(y)$ minimumspunkt for $y=1$ med $f(1)=a-1$. For alle $y\neq 1$ er $f(y)>a-1$.
Tilsvarende med positiv $c\neq 1$ vil $u_2c-c^{u_2}>u_2-1$ og dermed er (7.2) oppfylt.

For generell n brukes induksjon. Vi antar at

$$u_1x_1+\cdots+u_kx_k>x_1^{u_1}\cdots x_k^{u_k},\quad k>2 \tag{7.3}$$

Vil vise at da må (7.3) også gjelde for k+1. Siden vi vet at (7.3) gjelder for k=2, kan vi skrive:

$$u_1x_1+\cdots+u_kx_k+u_{k+1}x_{k+1}=(u_1+\cdots+u_k)\tilde{x}+u_{k+1}x_{k+1}>\tilde{x}^{(u_1+\cdots+u_k)}x_{k+1}^{u_{k+1}} \tag{7.4}$$

med $\tilde{x}=\frac{u_1}{u_1+\cdots+u_k}x_1+\cdots+\frac{u_k}{u_1+\cdots+u_k}x_k$ og $u_1+\cdots+u_{k+1}=1$. Ved å sette inn for $\tilde{x}$ på høyre side og legge merke til at koeffisientene i $\tilde{x}$ summeres til 1 oppnås ved (7.3):

$$\tilde{x}^{(u_1+\cdots+u_k)}=(\frac{u_1}{u_1+\cdots+u_k}x_1+\cdots+\frac{u_k}{u_1+\cdots+u_k}x_k)^{(u_1+\cdots+u_k)}>\left(x_1^{\frac{u_1}{u_1+\cdots+u_k}}x_2^{\frac{u_1}{u_1+\cdots+u_k}}\cdots x_k^{\frac{u_1}{u_1+\cdots+u_k}}\right)^{u_1+\cdots+u_k}=x_1^{u_1}x_2^{u_2}\cdots x_k^{u_k}$$

Innsatt i (7.4) gir dette: $u_1x_1+\cdots+u_{k+1}x_{k+1}>x_1^{u_1}\cdots x_{k+1}^{u_{k+1}}$ som var det vi skulle vise. □

Teorem 2 (ulike vekter):
La $0 < u_i, v_i < 1$ og $x_i > 0$, for $i = 1,2$ og slik at $u_1 + u_2 = 1$, $v_1 + v_2 = 1$ og $x_1 \neq x_2$. Flg. kriterium:

(7.5) $$\left(\frac{v_2}{u_2}\right)^{\frac{v_2}{v_1}} < \frac{u_1}{v_1}$$

er ekvivalent med:

(7.6) $$u_1x_1 + u_2x_2 > x_1^{v_1}x_2^{v_2}, \text{ for alle } x_1 \text{ og } x_2$$

Hvis (7.5) ikke er oppfylt betyr det at ulikheten i (7.6) ikke er oppfylt for alle x_1 og x_2. Dersom forskjellen mellom x_1 og x_2 er stor nok er likevel ulikheten i (7.6) oppfylt.

Bevis:
Som i Teorem 1 kan (7.6) formuleres som flg. ulikhet:

(7.7) $$1 - u_2 + u_2c - c^{v_2} > 0 \quad \text{for} \quad u_2, v_2 \in (0,1) \quad \text{og} \quad c > 0, \neq 1$$

La $a = u_2$, $b = v_2$, $f(y) = ay - y^b$ for $a, b \in (0,1)$ og $y \geq 0$. Da har $f(y)$ minimumspunkt for $y^* = \left(\frac{b}{a}\right)^{\frac{1}{1-b}}$ med $f(y^*) = \left(\frac{b}{a}\right)^{\frac{b}{1-b}}(b-1)$. Ved å kreve $f(y^*) > a - 1$ følger (7.5) direkte. Hvis $f(y^*) < a - 1$ er likevel (7.7) oppfylt når $c \to \infty$. □

Vi er interessert i å sammenligne $\sum_{i=1}^{n} u_i x_i$ og $\prod_{i=1}^{n} x_i^{v_i}$ generelt. Tilfellet der $u_i = v_i$ er dekket av Teorem 1. For å undersøke $u_i \neq v_i$ begrenser vi mulighetene til å gjelde kun våre vekter, slik at vektene må velges blant de tre alternativene $u_i, v_i \in \left\{\frac{p_i^2}{\sum_i p_i^2}, \frac{p_i}{\sum_i p_i}, \frac{1}{n}\right\}$ for $i = 1, \cdots, n$ (det samme alternativet for alle i-ene). Hvis vi videre antar at $n = 2$ og $p_2 = cp_1$, for $c > 0, \neq 1$ er det lett å vise at (7.5) ikke er oppfylt for våre vekter. Med omformingen $\left(\frac{p_1^2}{p_1^2+p_2^2}, \frac{p_2^2}{p_1^2+p_2^2}\right) = \left(\frac{p_1^2}{p_1^2+c^2p_1^2}, \frac{c^2p_1^2}{p_1^2+c^2p_1^2}\right) = \left(\frac{1}{1+c^2}, \frac{c^2}{1+c^2}\right)$ og $\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) = \left(\frac{1}{1+c}, \frac{c}{1+c}\right)$ får vi flg. for de tre forskjellige kombinasjonene:

(7.8) $$(u_1, u_2) \text{ og } (v_1, v_2) = \begin{cases} (\frac{1}{1+c^2}, \frac{c^2}{1+c^2}) \text{ og } (\frac{1}{1+c}, \frac{c}{1+c}) & \Rightarrow \ (\frac{v_2}{u_2})^{\frac{v_2}{v_1}} = (\frac{1+c^2}{c+c^2})^c \text{ og } \frac{u_1}{v_1} = \frac{1+c}{1+c^2} \\ (\frac{1}{1+c}, \frac{c}{1+c}) \text{ og } (\frac{1}{2}, \frac{1}{2}) & \Rightarrow \ (\frac{v_2}{u_2})^{\frac{v_2}{v_1}} = \frac{1+c}{2c} \text{ og } \frac{u_1}{v_1} = \frac{2}{1+c} \\ (\frac{1}{2}, \frac{1}{2}) \text{ og } (\frac{1}{1+c^2}, \frac{c^2}{1+c^2}) & \Rightarrow \ (\frac{v_2}{u_2})^{\frac{v_2}{v_1}} = (\frac{2c^2}{1+c^2})^{c^2} \text{ og } \frac{u_1}{v_1} = \frac{1+c^2}{2} \end{cases}$$

Enkel drøfting av funksjonene viser at i alle tre tilfeller er $(\frac{v_2}{u_2})^{\frac{v_2}{v_1}} > \frac{u_1}{v_1}$ for alle $c > 0, \neq 1$, og altså er ikke (7.5) oppfylt. De tre resterende kombinasjonene (bytter om u-ene og v-ene) gir det samme resultatet. Teorem 2 gir at ulikheten i (7.6) dermed ikke gjelder for alle x_1 og x_2. Siden ulikheten i (7.6) ikke nødvendigvis er oppfylt for $n = 2$, er den tilsvarende ulikheten for vilkårlig n heller ikke nødvendigvis oppfylt. Vi kan dermed konkludere med at når vektene er forskjellige ($u_i \neq v_i$) og blant de vi har sett spesielt på, er ikke nødvendigvis det aritmetiske gjennomsnittet større enn det geometriske.

De sist utgitte publikasjonene i serien Notater

2000/23 T. Risberg. G. Rogdaberg og R.M. Rosvold: Sykepleiernes tilpasning i arbeidsmarkedet: En kort beskrivelse av teorier og dataregistre. 46s.

2000/24 A.S. Brørs, K. Dybendal, A.H. Foss og T. Jakobsen: Dokumentasjon av BESYS - befolkningsstatistikksystemet: Befolkningsendringer i 1998 og befolkningsbasen (BEBAS) 1. januar 2000. 43s.

2000/25 E. Høydahl: FoB2001: Kommunenes innspill om kommunehefter. 18s.

2000/26 T. Kalve og J. Sørøy: Revisjon av barnevernsdata. 30s.

2000/27 A. Skoglund: Publikasjoner fra forskningsvirksomheten 1991-1999. 72s.

2000/28 H. Hungnes: Omregning av KVARTS-relasjoner til MODAG-relasjoner. 12s.

2000/29 R.N. Johnsen: Undersøking om foreldrebetaling i barnehagar, januar 2000. 36s.

2000/30 O. Rognstad: Plan for landbruksstatistikken etter 1999. 23s.

2000/31 Ø. Kleven: Levekårsundersøkelsen i Longyearbyen 2000: Dokumentasjon og tabellrapport. 188s.

2000/32 E. Rønning: Omnibusundersøkelse - mars 2000: Dokumentasjonsrapport. 34s.

2000/33 J. Johansen og J. Lajord: FD-trygd. Dokumentasjonsrapport. Utdanning. 1992-1997. 119s.

2000/34 A.L. Brathaug, J. Holmøy og H. Tønseth: Årsrapport: Kontaktutvalget for helse- og sosialstatistikk 1999. 24s.

2000/35 N. Barrabés: Norsk standard for utdanningsgruppering: Høringsnotat. 110s.

2000/36 D. Roll-Hansen og C.M. Hansen: En evaluering av datainnsamling gjennom KOSTRA: Rapportering av data fra 1999. 94s.

2000/37 B.R. Joneid og Ø. Sivertstøl: FD - trygd: Dokumentasjonsrapport: Foreløpig uførestønad, 1992-1998. 30s.

2000/38 R.N. Johnsen: Kommunale gebyrer knyttet til bolig. Januar 2000. 27s.

2000/39 J-A.S. Lie: Revisjon av data til Pleie- og omsorgsstatistikken i 1997 og 1998. 83s.

2000/40 Y. Holm, A.H. Tangen og O.M. Tidemann: Forprosjektrapport om etablering av IMFs internasjonale investeringsposisjon (IIP) for Norge. 97s.

2000/41 K.O. Olsen: Forsikring i nasjonalregnskapet. 42s.

2000/42 J. Johansen og J. Lajord: FD - Trygd: Dokumentasjonsrapport: Arbeidssøkere. 1992-1998. 74s.

2000/43 H.V. Sæbø: Til statistikkens pris: Prispolitikk i statistikkbyråene med hovedvekt på elektronisk formidling. 9s.

2000/44 E. Rønning: Omnibusundersøkelse - mai/juni 2000. Dokumentasjonsrapport. 32s.

2000/45 A. Holmøy og M. Høstmark: Undersøkelse om omfanget av utgifter til helse- og sosialtjenester: Dokumentasjon og tabellrapport. 116s.

2000/46 Fagseminar om arealpolitikk og arealstatistikk i opptakten til et nytt årtusen. Seminarrapport 30. mars 2000. 167s.

2000/47 Publikasjoner fra forskningsvirksomheten 1991-1999: Revidert versjon. 82s.

2000/48 A.-K. H. Grorud: Bedrifts- og foretaksregisteret: Regler og rutiner for ajourhold. 121s.

2000/49 T. Hoel, B.R. Joneid og G.E. Wangen: Trekkbas: Brukerdokumentasjon. 35s.

2000/50 J.F. Bjørnstad: En innføring i utvalgsundersøkelser. 91s.