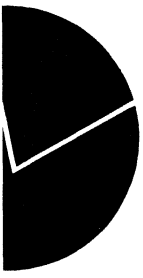


Li-Chun Zhang og Ole Klungsøyr

Med orden på data
- Estimering av terminvise omsetningstall

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder

1 Innledning

Seksjon for økonomiske indikatorer (240) planlegger å øke aktualiteten på omsetningsstatistikken ved bruk av månedlige skjematall fra et utvalg. Et nødvendig skritt på veien er å kunne estimere terminvis total omsetning i populasjonen (ca. 20 000 bedrifter) basert på et utvalg (ca. 2 000) og historien til tidligere termintall fra momsverdiregisteret.

Dette notatet fokuserer på det sistnevnte estimeringsproblemet. Først skal vi gi en kort oversikts beskrivelse av data, utvalg og populasjon. Deretter skal vi studere grunnlaget for analysen, og på bakgrunn av dette diskutere, nokså generelt, forskjellige alternative metoder. Mer detaljerte studier som støttemateriale finnes i flere appendiks for de interesserte.

2 Terminvise omsetningstall

2.1 Populasjonen og utvalget

Omsetningstall publiseres terminvis, 6 av dem i løpet av et år. Vi skriver 96b1 for den første terminen i 1996, og 96b2 for den andre, osv. Populasjonen omfatter i alt litt over 20 000 bedrifter, deriblant ca. 2 000 av dem i utvalget. I dette studiet har vi brukt datamateriale fom. 1995 tom. 1997, i alt 18 terminer.

Utvalgsbedriftene har vært plukket ut etter en standard utvalgsplan ved S240, der bedrifter med stor omsetning har større trekksannsynlighet enn bedrifter med mindre omsetning. Dekningsgraden er ca. 75%, dvs. total omsetning i utvalget utgjør ca. 75% av total omsetning i populasjonen. I tillegg er det slik at en bedrift, når den først er trukket, er med i utvalget så lenge den eksisterer.

Kommentar 1 *På den måten kan utvalget stort sett betraktes som et panel over flere terminer. Nye bedrifter trekkes for å beholde noenlunde konstant antall bedrifter i utvalget. Endring fra en termin til den neste er på under* 1%.

2.2 Data

I likhet med mange andre variabler definert på enheter som bedrift og foretak, hører omsetning til den type data som karakteristisk er mer 'normale' på log-skala enn på den opprinnelige skala. Eksempelvis har vi tatt med histogram for omsetning i 97b1 (Figur 1). Spesielt har vi brukt følgende transformasjon av data, nemlig

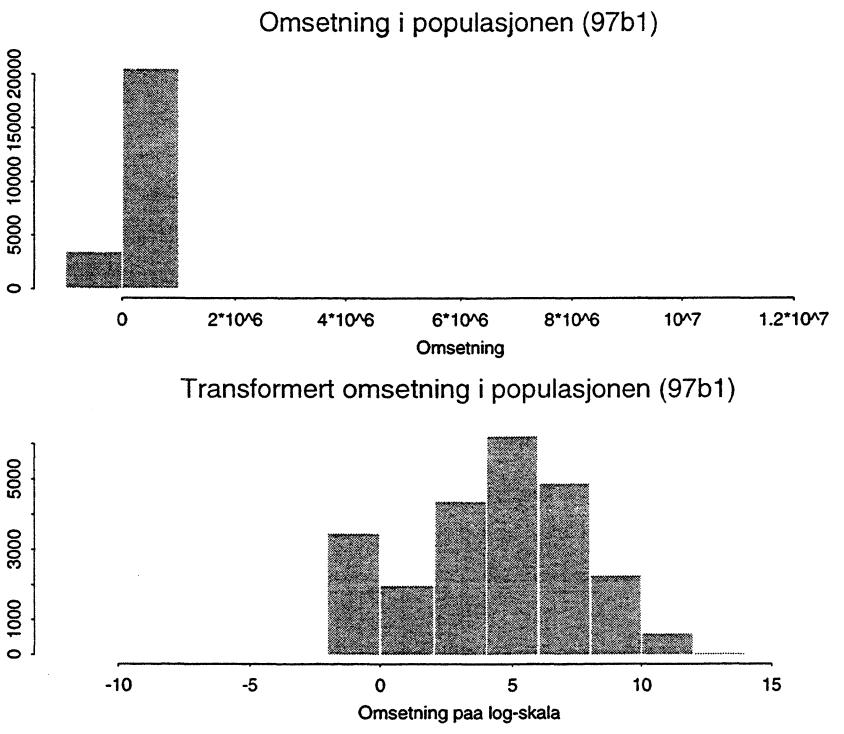
$$z = \log(x+1) \text{ for } x \geq 0 \qquad \text{og} \qquad z = -\log(-x+1) \text{ for } x < 0, \tag{1}$$

som (i) bevarer ordningen i data, og (ii) behandler verdi null på en elegant måte. Det er klart at data har nå fått en mye bedre spredning etter transformasjonen. Spesielt ser de positive omsetningstallene i populasjonen mer 'normale' ut.

Kommentar 2 *I tilfellet variabelen er strengt positiv, eller negativ, kan man også benytte direkte*

$$z = sgn(x) \cdot \log|x| \qquad sgn(x) = 1 \text{ for } x > 0 \text{ og } -1 \text{ for } x < 0.$$

Figur 1. Omsetningstall på forskjellige skala



3 Med orden på data

3.1 Tilleggsinformasjon

Fra skattedirektoratet kan man i momsverdiregisteret hente inn omsetning for hele populasjonen med ca. 5 mnd forsinkelse. Slike opplysninger danner over tid en omsetningshistorie for hver bedrift i landet. Dette er vår *tilleggsinformasjon*, dvs. i tillegg til utvalget ved den inneværende terminen.

For to terminer, betegnet med t_0 og t_1 og $t_0 < t_1$, finnes det en felles populasjon, betegnet med U, som består av alle bedriftene som eksisterte ved begge tidspunktene. For hver bedrift fra U, betegnet med $i \in U$, betegner vi dens omsetning ved t_0 med x_i og den ved t_1 med y_i. I tilfellet t_1 er den inneværende termin, vil $Y = \sum_{i \in U} y_i$ være total omsetning man ønsker å estimere, og x_i for $i \in U$ tilleggsinformasjon.

Eksempelvis har vi i Figur 2 plottet y_i ved 97b1 mot x_i for de 6 foregående terminene, etter at de er transformert ifølge (1). Samtidig har vi beregnet parvis korrelasjonskoeffisient, med og uten transformasjon (1):

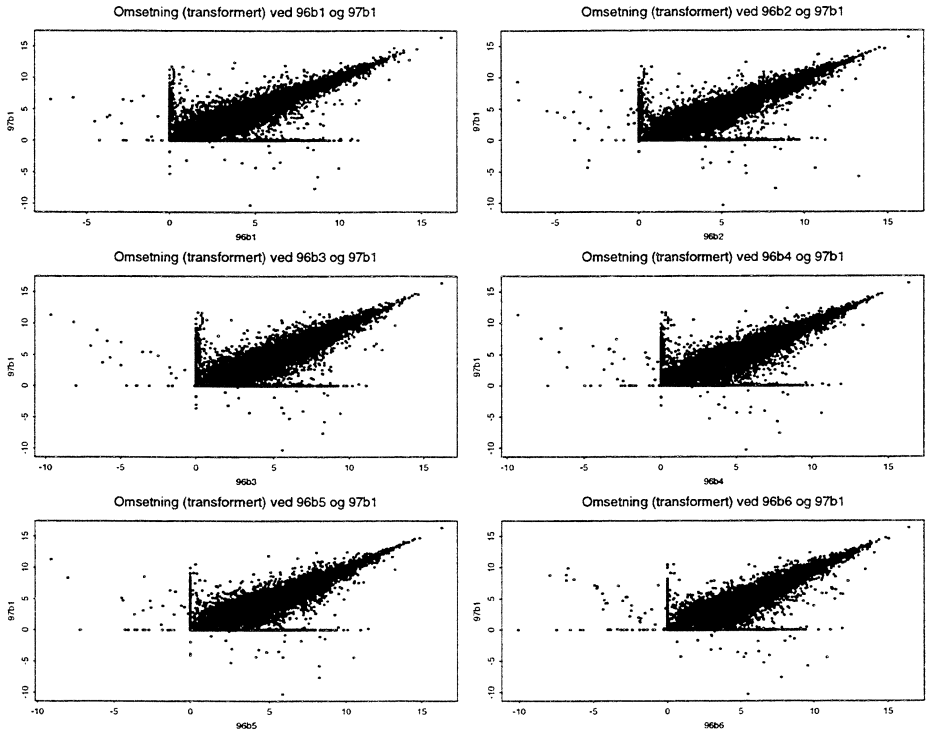
Tabell 1. Populasjons korrelasjonskoeffisient mellom omsetning i 97b1 og 96b1 - 96b6

	96b1	96b2	96b3	96b4	96b5	96b6
Omsetning	0.975	0.987	0.990	0.995	0.993	0.989
Transformert omsetning (log-skala)	0.865	0.863	0.858	0.858	0.891	0.885

Situasjonen er lignende i alle terminer vi har undersøkt, dvs. fra 1995 til 1997 og i alt 18.

Kommentar 3 *Til tross for mulige forstyrrelser fra extreme verdier i pupolasjonen, er det liten tvil om at korrelasjonen er høy. (Slike forstyrrelser er redusert på log-skala.) Viktigst er at den ikke*

Figur 2. Omsetning for alle bedrifter i 97b1 og 96b1 - 96b6



svekkes betydelig ettersom tiden går, i hvert fall hvis man ikke går altfor langt tilbake i historien. Dette kan man også se fra Figur 2 direkte.

3.2 Ordnete data

Anta felles populasjonen U ved t_0 og t_1. Vi kan ta alle x_i for $i \in U$, og ordne dem til $x_o = (x_{(1)}, ..., x_{(n)})$, slik at $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. På den måten blir x_o den *den ordnete observasjon* av x. Det samme kan vi gjøre med y_i for $i \in U$, og som resultatet får vi y_o. Det er viktig å notere at $x_{(i)}$ og $y_{(i)}$ ikke lenger behøver å referere til den samme bedriften, i motsetning til x_i og y_i.

Vi har i Figur 3 plottet $y_{(i)}$ ved 97b1 mot $x_{(i)}$ for de 6 foregående terminene, etter at de er transformert ifølge (1), der vi har lagt inn linjen $y = x$ i tillegg. Samtidig har vi beregnet parvis korrelasjonskoeffisient mellom de ordnete populasjonene, med og uten transformasjon (1):

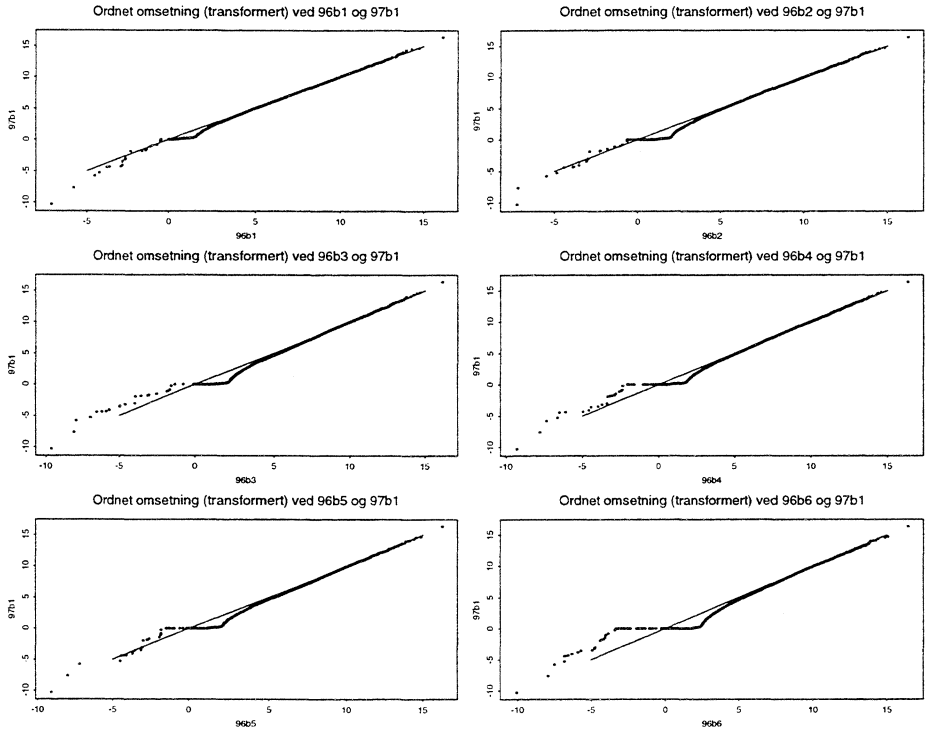
Tabell 2. Korrelasjonskoeffisient mellom ordnet omsetning i 97b1 og 96b1 - 96b6

	96b1	96b2	96b3	96b4	96b5	96b6
Omsetning	0.998	0.999	0.998	0.999	0.998	0.997
Transformert omsetning (log-skala)	0.998	0.995	0.994	0.996	0.995	0.989

Resultat 1 Vi kan bli kvitt en stor del av variasjon i data, *som er forårsaket av endring fra x_i til y_i for $i \in U$,* **ved å bruke ordnete observasjonen.**

Kommentar 4 *Så lenge spørsmålet dreier seg om totalen istedenfor prediksjon for enkelte bedrifter, kan vi tillate oss å permutere data på denne måten. Dette kan man forstå hvis man forestiller seg to bedrifter, betegnet med $i, j \in U$, slik at $x_i = y_j$ og $y_i = x_j$ med stor forskjell mellom x_i og y_i. Hver*

Figur 3. Ordnet omsetning for alle bedrifter i 97b1 og 96b1 - 96b6



for seg bidrar de to bedriftene stort til variasjon i data, til tross for at slik variasjon betyr ingenting for totalen, siden $x_i + x_j = y_i + y_j$.

3.3 Grunnlaget for analysen

Den *empiriske fordelingen* til x_i der $i \in U$ er definert ved

$$F_U(x) = \frac{\text{antall } x_i \leq x \text{ der } i \in U}{\text{antall bedrifter i populasjonen}}. \tag{2}$$

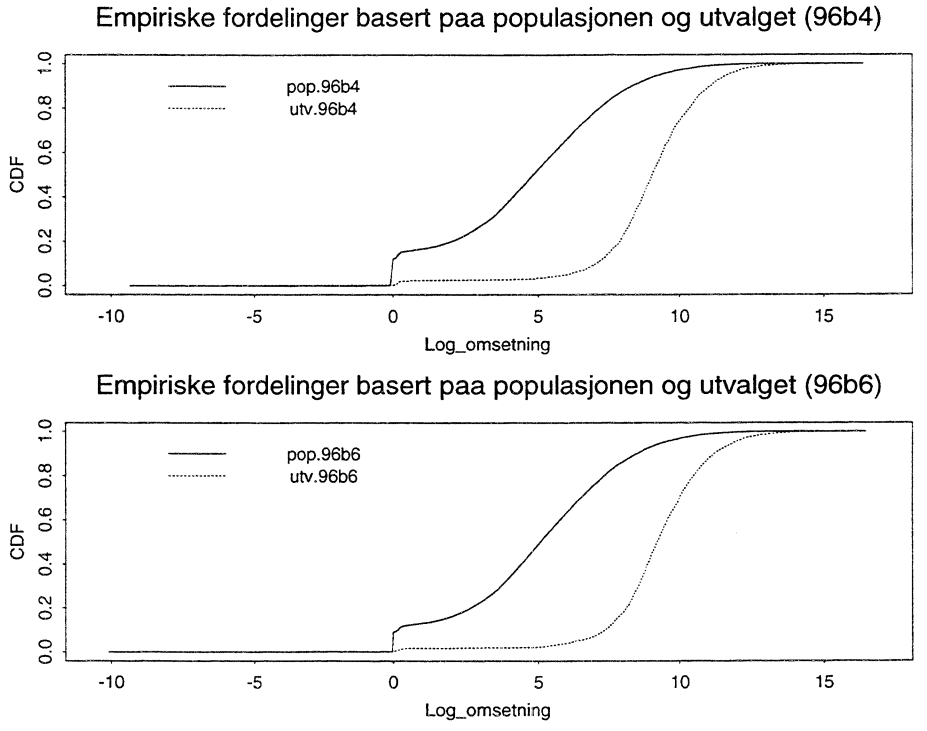
All statistikk om populasjonen, inkl. total/gjennomsnittlig omsetning, er en funksjon av den empiriske fordelingen. Mao. $F_U(x)$ inneholder all informasjon om omsetning i populasjonen ved t_0. Definer $F_U(y)$ på lignende vis for y_i der $i \in U$, som inneholder all informasjon om omsetning i populasjonen ved t_1. Betegn det felles utvalget ved t_0 og t_1 med s, og definer $F_s(x)$ og $F_s(y)$ for $i \in s$.

Ta den inneværende termin, betegnet med t_1*, og en tidligere termin, betegnet med* t_0*, vi kjenner til* $F_U(x)$*,* $F_s(x)$ *og* $F_s(y)$*, og ønsker å estimere den inneværende totale omsetningen, betegnet med* Y*, som er en funksjon av* $F_U(y)$.

Det er da instruktivt å se hvordan disse empiriske fordelingene ser ut. Eksempelvis har vi tatt med $F_U(x)$ og $F_s(x)$ fra 96b4 og $F_U(y)$ og $F_s(y)$ fra 96b6, og plottet dem i Figur 4. Igjen har vi transformert data etter (1), noe som flater F ut mer enn på den opprinnelige skala.

Man kan se en klar overtrekking blant de store bedrifter i utvalget, noe som gir utslag i at F_s er mye brattere enn F_U. Det er forholdsvis flere store bedrifter i utvalget enn i populasjonen, mens det er langt færre negative, eller null, omsetnings verdier i utvalget enn i populasjonen. Samtidig merker

Figur 4. Empiriske fordelinger i utvalget og populasjonen



man at F_U ved 96b4 ligner på F_U ved 96b6, og det samme gjør F_s ved 96b4 og 96b6. Dette ser vi bedre hvis vi stokker kurvene som i Figur 5.

Kommentar 5 *En god estimeringsmetode skal kunne fortolke og utnytte likhetstrekk mellom F_U ved t_0 og t_1, og F_s ved t_0 og t_1 på en fornuftig måte.*

4 Diskusjon om forskjellige metoder

4.1 To ratemodeller

Gitt den inneværende termin og en tidligere termin, er en ratemodell vanligvis definert som følgende, for hver bedrift $i \in U$,

$$y_i = \beta x_i + \epsilon_i \qquad \text{der } E[\epsilon_i] = 0 \quad \text{og} \quad Var(\epsilon_i) = \sigma_i^2. \tag{3}$$

Spesielt kan raten β og, dermed, total omsetning Y estimeres med

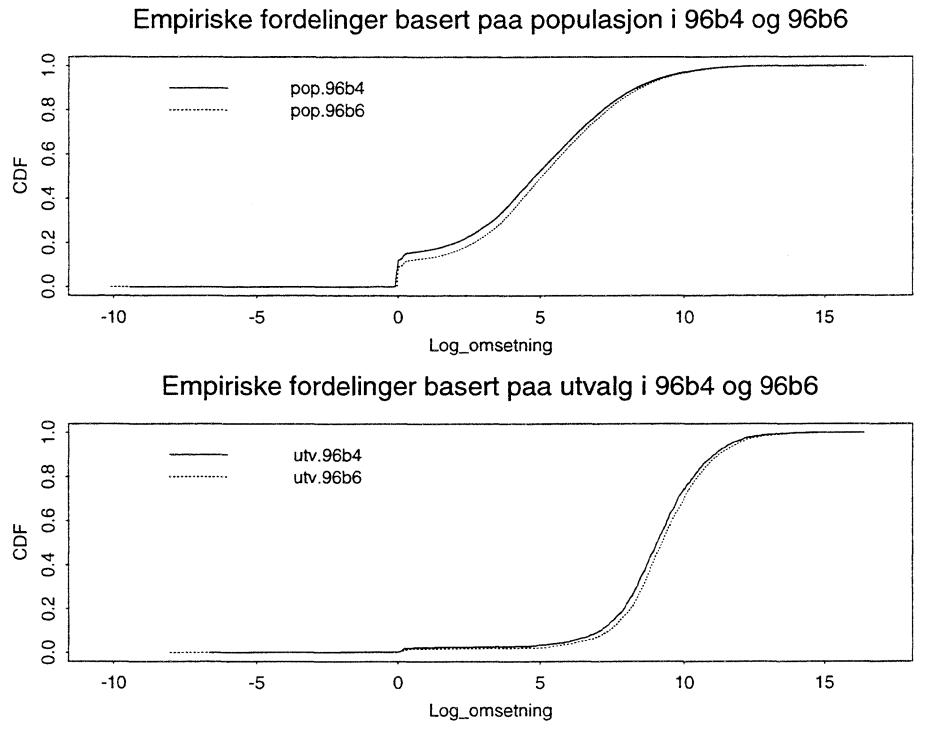
$$\hat{\beta} = (\sum_{i \in s} y_i)/(\sum_{i \in s} x_i) = Y_s/X_s \qquad \hat{Y} = \hat{\beta}(\sum_{i \in U} x_i) = \hat{\beta}X. \tag{4}$$

Kommentar 6 *Man har her benyttet informasjon i data, dvs. $F_U(x)$ og $F_s(x)$ og $F_s(y)$, i form av X som en funksjon av $F_U(x)$, og X_s av $F_s(x)$, og Y_s av $F_s(y)$.*

Men det finnes en ordnet versjon av modell (3) som gir det samme resultatet, nemlig

$$y_{(i)} = \beta x_{(i)} + \epsilon(i) \qquad \text{der } E[\epsilon(i)] = 0 \quad \text{og} \quad Var[\epsilon(i)] = \tau_i^2. \tag{5}$$

Figur 5. Re-arrangerte empiriske fordelinger i utvalget og populasjonen



Legg merke til at $\epsilon(i)$ er hverken ϵ_i, dvs. restledd til den i-te bedriften, eller $\epsilon_{(i)}$, dvs. det i-te minste restleddet. Mao. vi antar nå at raten holder for de to største omsetningene i populasjonen, og de to neste største, osv. istedenfor at raten holder for hver bedrift.

Kommentar 7 *Modell* (3) *impliserer modell* (5)*, ikke omvendt. Spesielt medfører modell* (5) *at* $\log E[y_{(i)}] = \log x_{(i)} + \log \beta$ *for* $x_{(i)} > 0$*. Mens modell* (3) *krever at* $\log E[y_i] = \log x_i + \log \beta$ *for* $x_i > 0$*, noe som er vanskeligere å tilfredsstille (Figur 2 og 3).*

Resultat 2 Under modell (5) **eller** (3)**, er den forventete empiriske fordelingen** *for den positive, eller negative, delen av populasjonen* **ved** t_1 **en parallel forskyvning av den empiriske fordelingen ved** t_0**.**

Begge ratemodellene fortolker likhetstrekk mellom de empiriske fordelingene i form av en parallel forskyvning av kurvene (Figur 5). Også estimatet for total omsetning er det samme. Forskjellen ligger i diagnostikk på hvor godt modellene fungerer i en bestemt situasjon — appendiks A.

4.2 Vekting

Vekting er en av de vanligste estimeringsmetodene i utvalgsundersøkelser. Hver bedrift i utvalget, betegnet med $i \in s$, får en vekt, betegnet med w_i, basert på tilleggsinformasjon, her f.eks. omsetning fra en tidligere termin, altså x_i for $i \in U$. Som regel krever man at de vektene skal være kalibrert i følgende forstand,

$$\sum_{i \in s} w_i = N \qquad \sum_{i \in s} w_i x_i = X = \sum_{i \in U} x_i,$$

om ikke enda flere. Estimatet for total omsetning i en delpopulasjon, betegnet med Q, er da gitt som $\sum_{i \in Q} w_i y_i$.

Kommentar 8 *Ser man bort fra trekksannsynlighet, vil kravet om $\sum_{i \in s} w_i = N$ føre til $w_i = N/n$ for $i \in s$, der n er antall bedrifter i utvalget, og katastrofale resultater. På den andre siden vil kravet om $\sum_{i \in s} w_i x_i = X$ føre til at $w_i = X/X_s = \hat{\beta}$ i (4) under ratemodellen.*

Vekting som tar hensyn til overtrekking blant de store bedriftene, hviler sterkt på antagelsen om at den samme bedrift skal ha omtrent det samme forholdet til hele populasjonen ved begge tidspunkter, noe som gjør at den er spesielt sårebar overfor stor, og til dels ekstrem, endring fra x_i til y_i, som slett ikke er uvanlig med bedriftsdata (Figur 2).

Kommentar 9 *Vekt w_i kan nemlig forstås som hvor mange 'kopier', av den i-te bedriften i utvalget, finnes i hele populasjonen. Men stor endring fra x_i til y_i er ofte heller et særtrekk enn et tegn på at alle bedrifter med omsetning rundt x_i har gjennomgått en like stor endring fra t_0 til t_1.*

Det er også mulig å bringe orden til vekting, som modell (5) gjør for (3). Mao. man kan lage vekt for $x_{(i)}$ og bruker den på $y_{(i)}$, uten å bry seg om de refererer til den samme bedriften. I appendiks B har vi beskrevet en slik orden-vektingsmetode. I de få tilfellene der vi prøvde ut denne *ad hoc* orden-vektingsmetoden, var resultatene slett ikke så verst. Det er allikevel en del betraktninger som gjør at vi ikke gikk videre med den (appendiks B)

Kommentar 10 *Det sentrale metodiske spørsmålet her dreier seg om, "hvordan definerer/estimerer man den empiriske fordelingen i populasjonen, basert på utvalget og vektene?" Altså hva er $\hat{F}_U(w, s)$?*

4.3 Autoregressiv modellering

Figur 2 og 3, samt Tabell 1 og 2, tyder på at det ikke nødvendigvis er mye å hente i en autoregressiv modell på omsetningstidsserie, siden slike modeller er best når korrelasjonen avtar betydelig etter et moderat antall tidspunkter. En autoregressiv modell vil også støte på betydelig vanskeligheter pga. store endringer fra x_i til y_i, dvs. fra tid til annen, siden en tidsserie er bundet fast til den samme bedriften.

4.4 Robustisering av rate (I): Ekstreme bedrifter?

Med den typen data som bedriftsomsetning, er det et kjent problem at enkelte/få ekstreme, eller veldig store, verdier kan ødelegge for rate-estimatet (4). Det finnes en del litteratur som handler om robustisering av rate-estimering — se Lee (1995) og referansene der. De forskjellige metodene koker ned til følgende to-trinns skjema: (i) identifiser såkalte ekstreme bedrifter basert på (x_i, y_i) , og (ii) modifiser estimat (4) enten ved å utelate de identifiserte ektremene eller nedveie dem på et eller annet vis.

Kommentar 11 *Slik robustisert rate-estimering er trolig nyttigere dersom man er mest interessert i det struktursmessige studiet, der man ofte med rett kan kategorisere noen ekstreme som 'ikke-representative' for strukturen. Men det finnes en del grunnleggende vanskeligheter med denne fremgangsmåten når man har en endelig populasjon å forholde seg til.*

I vårt tilfelle er målet å komme nærmest mulig den bestemte totale omsetningen til en hver tid; og det finnes kun en eneste riktig rate, nemlig Y/X, som slett ikke behøver å være fornuftig ut fra

enhver teoretisk synsvinkel. Identifikasjon av ekstreme bedrifter basert på forholdet mellom x_i og y_i fører til komplikasjon siden en ekstrem som ble utelatt/nedveiet i rate-estimering, vil nødvendigvis komme tilbake med 'full tyngde' når den estimerte raten etterpå skal brukes til å estimere totalen. Mao. det finnes ingen ekstreme som er "ikke-representative" for totalen, til tross for at den kan f.eks. være "ikke-representativ" for trenden fra t_0 til t_1. I appendiks C har vi beskrevet to slike rate-robustiseringsmetoder og en del erfaringer med dem.

4.5 Robustisering av rate (II): Ekstrem omsetning

Studiet av ratemodellene (appendiks A) støttet til den ordnete versjon (5), der restledd, dvs. $\epsilon(i) = y_{(i)} - \hat{\beta}x_{(i)}$, helst skal (i) summere opp til null, og (ii) fordele seg balansert rundt linje $y = \hat{\beta}x$, uansett hvilken del av det ordnete utvalget man ser på. Spesielt tyder detaljert analyse av restledd på at $\epsilon(i)$ blant de største i'ene er mest avgjørende for resultatet. For å minske denne følsomheten i rateestimering, tenker man seg nå følgende *dempet rate-estimat*, betegnet med $\hat{\beta}^*$, og tilsvarende justert estimat for total omsetning, betegnet med $\hat{Y}^*$, nemlig

$$\hat{\beta}^* = \frac{\sum_{i=1}^{n-k_y} y_{(i)} + k_y \cdot t_y}{\sum_{i=1}^{n-k_x} x_{(i)} + k_x \cdot t_x} = \frac{Y_s^*}{X_s^*}, \qquad \hat{Y}^* = Y_s + \hat{\beta}^*(X - X_s), \tag{6}$$

der t_y er et på forhånd bestemt snittpunkt ("cut-off") til y_i, og t_x tilsvarende for x_i. Typisk er $t_y < y_{(n)}$, slik at k_y er antall y_i som er større eller lik t_y. Lignende gjelder for t_x og k_x. Mao. man omgjør de største k_x omsetningstallene ved t_0 til t_x, og de k_y ved t_1 til t_y, og beregner først deretter X_s^* og Y_s^* som vanlig. Legg merke til at generelt vil denne trunkeringen bli brukt på forskjellige bedrifter ved forskjellige tidspunkter, dvs. man identifiserer ekstrem omsetning istedenfor ekstreme bedrifter. Manipulering av snittpunkt (t_x, t_y) kan bidra til å redusere både $|\epsilon(i)|$ og $Var\{\epsilon(i)\}$ for $i > \min(n - k_x, n - k_y)$. I det tilfellet der man kan finne

$$t_y = t_x\{(\sum_{i=1}^{n-k} y_{(i)})/(\sum_{i=1}^{n-k} x_{(i)})\},$$

slik at $k_x = k_y = k$, blir $\epsilon(i) = 0$ for all $i > n - k$.

For å minske skjevheten i $\hat{\beta}^*$ (6), dvs. når den skal brukes til å estimere total omsetning, kan man benytte tidligere data for å eksperimentere seg fram til et fornuftig valg på snittpunkt (t_x, t_y). I appendiks D har vi beskrevet noen erfaringer og resultater i denne sammenheng.

Kommentar 12 *F.eks. kan man bruke empiriske materialer fra (95b1, 96b1), (95b2, 96b2), ..., og (96b6, 97b6) til å finne (t_x, t_y) i $\hat{\beta}^*$ for 98b1 basert på 97b1.*

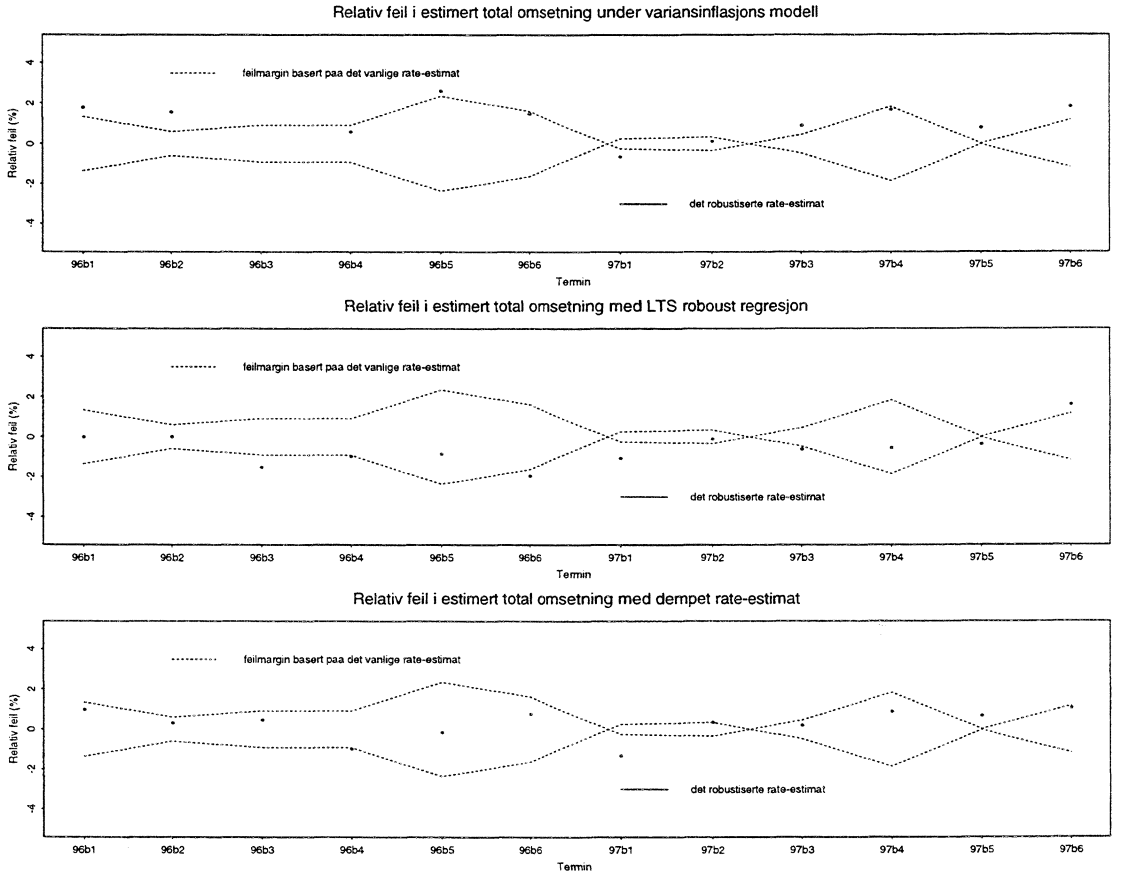
5 Oppsummering og videre arbeid

Med orden på data har vi funnet et ankepunkt for metodisk utvikling i forbindelse med estimering for terminvise omsetningstall, der standard metoder som vekting og autoregressiv modellering møter betydelige vanskeligheter pga. store enhetsvise endringer. Idéen kan ha generelle implikasjoner for analyse av data knyttet til enheter som bedrifter og foretak, der det er vanlig med slik stor variasjon i enhetenes historie.

Vi har studert rateestimering i detalj, og kommet fram til et robust rate-estimat (6), basert på analyse av restledd under ratemodell (5). I Figur 6 har vi plottet relativ feil i tre forskjellige robustiserte rate-estimater, sammenlignet med den vanlige $\hat{\beta}$ (4). (Et robustisert rate-estimat er

bedre enn $\hat{\beta}$ hvis feilen faller mellom den feilmarginen basert på $\hat{\beta}$.) Det dempete rate-estimatet (6) viser seg å være bedre sammenlignet med de to robuste rate-estimatene man finner i litteraturen. (Se Tabell 5 i appendiks C for de numeriske detaljene). Begrunnelsen ligger igjen i den ordnete tankegang.

Figur 6. Sammenligning av tre forskjellige robustiserte rate-estimater



Rateestimering for terminvis total omsetning kan videre stabiliseres dersom vi kombinerer flere rate-estimater på en fornuftig måte. La 98b5 være den inneværende termin, hver foregående termin gir oss et estimat, slik at man kan snakke om $\hat{\beta}^*$(98b4,98b5), $\hat{\beta}^*$(98b3,98b5), osv.. Legg merke til at disse estimerte ratene generelt refererer til forskjellige deler av den inneværende populasjonen pga. endring i populasjonen over tid, selv om forskjellene er små.

I tillegg bør man huske i det videre arbeidet at publisering foregår på et lavere nivå istedenfor hele popualsjonen i ett. Man kan anvende estimat (6) innen hver delpopulasjon man lager statistikk for. Men det kan også være nyttig å prøve ut et syntetisk opplegg, der man kombinerer estimatet for hele populasjonen med estimatet for delpopulasjonen på en balansert måte.

A Ratemodell og diagnostikk

A.1 Diagnostikk

Vi kan kalle en tidligere termin misvisende for den inneværende terminen, dersom det enkle rate-estimatet (4) har et forholdsvis stort avvik fra den sanne totalen. Generelt betyr dette at informasjonen i $F_s(x)$ og $F_s(y)$ ikke lar seg så godt sammenfattes av ratemodellen. *Diagnostikk* er en observator basert på utvalget, som danner grunnlag for vurderingen om nettopp dette.

Under modell (3) er det vanlig å bruke diagnostikk basert på summen til de kvadratiske restleddene. Da variasjonen i omsetning er størst blant de store bedriftene, antar man ofte en inflasjonsmodell på variansen σ_i^2 i (3), mao.

$$\sigma_i^2 = x_i^\gamma \sigma^2.$$

Ved å variere verdien til γ, kan man kontrollere størrelsen på "inflasjon". Vi skal sette $\gamma = 1$, noe som gir oss følgende diagnostikk for raten under modell (3),

$$D_1 = \sum_{i \in s} (y_i - \hat{\beta} x_i)^2 / x_i = \sum_{i \in s} \hat{\epsilon}_i^2 / x_i. \tag{7}$$

Under modell (5), derimot, beregner man restledd $\hat{\epsilon}(i)$ basert på $(x_{(i)}, y_{(i)})$ i stedet. På lignende vis som for D_1 i (7) har vi

$$D_{2,1} = \sum_{i \in s} (y_{(i)} - \hat{\beta} x_{(i)})^2 / x_{(i)} = \sum_{i \in s} \hat{\epsilon}(i)^2 / x_{(i)}. \tag{8}$$

Diagnostikk av type D_1 og $D_{2,1}$ skildrer 'profilen' til modelltilpasning fra en bestemt vinkel. Flere diagnostikker kan lages fra andre vinkler. Modell (5) kan betraktes som god dersom $F_s(x)$ og $F_s(y)$ står parallelt med hverandre. Dette medfører at, på den opprinnelige skala til data, sakl de positive og negative restleddene $\hat{\epsilon}(i)$ omtrent slå hverandre ut uansett hvilken del av den ordnete populasjonen man ser på, eller mer presist, for positive heltall a og b s.a. $a + b \leq n$, $\sum_{a \leq i \leq a+b} \hat{\epsilon}(i) \approx 0$. Spesielt viktig er det at dette holder i de øverste kvantilene siden, som vi skal vise, det er der $\hat{\epsilon}(i)$ absolutt er størst. Vi definerer derfor følgende diagnostikk under modell (5), for $0 < \alpha < 1$,

$$D_{2,2} = \sum_{[n(1-\alpha)] \leq i \leq n} \hat{\epsilon}(i), \tag{9}$$

der $[n(1-\alpha)]$ betegner det nærmeste heltall til $n(1-\alpha)$. Videre og på log-skala, bør differansen mellom kvantiler til de to empiriske fordelingene ligger nærmest mulig rundt en konstant, nemlig $\log \beta$. Vi definerer en tredje diagnostikk under modell (5) som, for $0 < \xi < 1$,

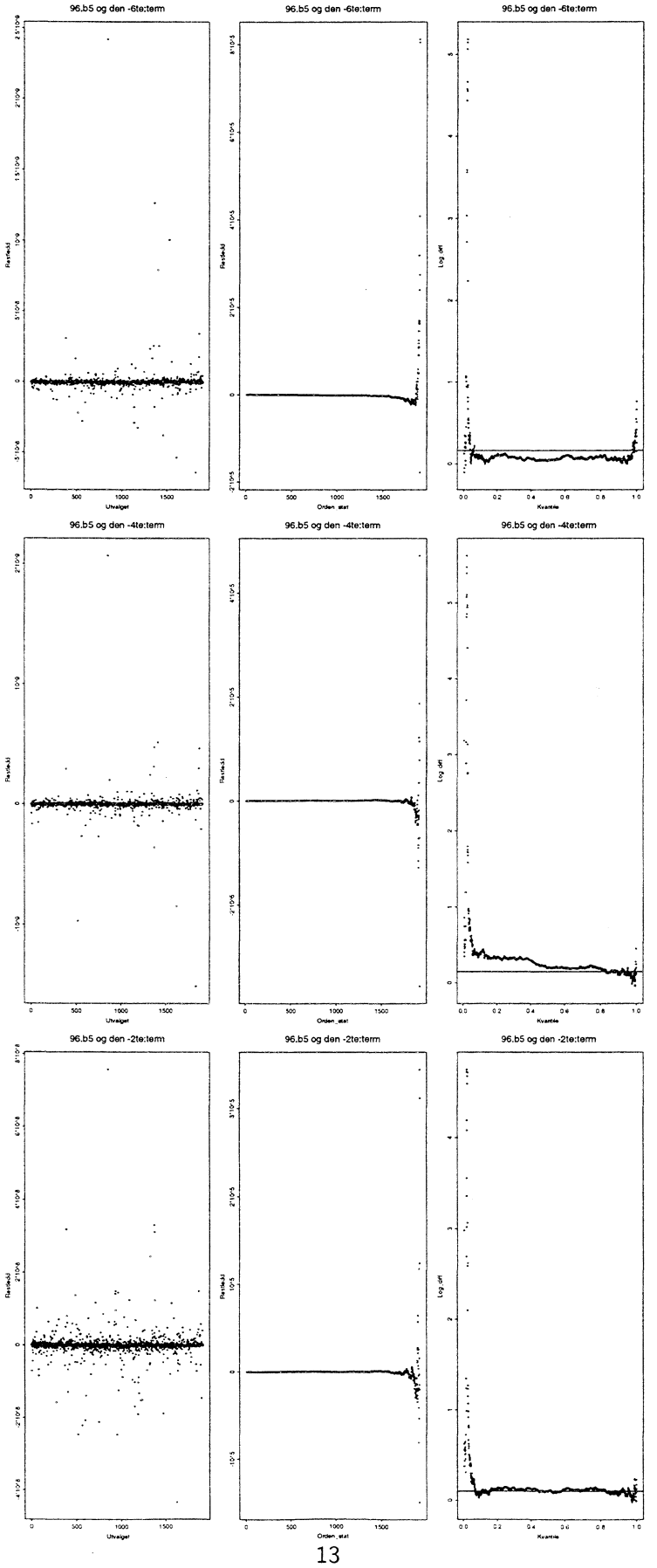
$$D_{2,3} = Var_{[n\xi] \leq i \leq n} (\log y_{(i)} - \log x_{(i)}) \qquad \text{der} \quad x_{(i)},\, y_{(i)} > 0. \tag{10}$$

Ved å velge passe stor ξ får man en mer stabil diagnostikk — de minste kvantilene er ustabile pga. undertrekking samtidig som de har lite å si om estimatet (4).

A.2 Ek eksempel: Diagnostikk for 96b5

Vi bruker 96b5 for en enkel illustrasjon. Ta først rate-estimatet basert på 95b5 og 96b5, nemlig den 6te termin før den inneværende. I Figur 7 har vi plottet, fra venstre til høyre, (i) ϵ_i i utvalget under modell (3), (ii) $\epsilon(i)$ basert på $x_{(i)}$ og $y_{(i)}$ under modell (5), og (iii) $\log y_{(i)} - \log x_{(i)}$ under

Figur 7. Noen eksempler på diagnostikk



modell (5) — linjen der angir $\log \hat{\beta}$ ifølge (4). Videre tar vi 96b5 og 96b1, nemlig 4 terminer før den inneværende; og til slutt, 96b5 og 96b3, nemlig 2 terminer før.

Felles for alle tre terminene legger vi merke til at $\hat{\epsilon}(i)$ er absolutt størst i de få øverste kvantilene, slik at disse kommer til å dominere diagnostikk (8). Samtidig blir det klart at man bare behøver å se på diagnostikk (9) med en liten α, som f.eks. $\alpha = 0.05$. Mens ξ for diagnostikk (10) skal helst være såpass stor at man unngår den store variasjonen i log-differansen blant de minste kvantilene. Når man betrakter de tre terminene hver for seg, synes det at, (i) $\hat{\epsilon}(i)$ er ganske skjevt fordelt i de øverste kvantilene for 95b5, (ii) $\log y_{(i)} - \log x_{(i)}$ holder seg liten til en bestemt verdi for 96b1, og (iii) uten noen oppsiktsvekkende trekk for 96b3.

A.3 Resultater

Vi skal studere disse diagnostikkene i detalj. For hver termin mellom 1996 og 1997, i alt 12 av dem, estimerer man ratene ved (4) basert på de 6 foregående terminer til den inneværende. På bakgrunn av hver av de 4 diagnostikkene (7) - (10), klassifiserer man den mest misvisende foregående termin, samtidig som den 'beste'. Resultatene er blitt samlet i Tabell 3.1 og 3.2, der de misvisende terminene er markert med "-", og de 'gode' med "+" — spesielt skal vi bruke "(-)" og "(+)" der skillet synes å være knapt. I tillegg har vi tatt med $\hat{\beta}$ (4), fasiten $\beta = Y/X$ og relativ feil i $\hat{\beta}$, nemlig $\hat{\beta}/\beta - 1$ i prosent. (Den siste kolonnen skal vi komme tilbake til.) Forsøksvis har vi satt $\alpha = 0.05$ i $D_{2,2}$ og $\xi = 0.25$ i $D_{2,3}$.

Generelt ser vi at diagnostikkene $D_{2,1}$ - $D_{2,3}$ stemmer bedre overens seg imellom enn i forhold til D_1. Mer detaljert skal vi forsøke å oppsummere resultatene her i flere retninger. Definer $\eta_t(D)$ av diagnostikk D slik at, for en bestemt termin t, $\eta_t(D) = 1$ dersom den 'beste' raten ifølge D faktisk har en mindre feil, dvs. absolutt verdi til $\hat{\beta}/\beta - 1$, enn den 'verste' raten ifølge D; og $S_t(D) = -1$ ellers. Eksempelvis har vi nå at $\eta_{96b1}(D_1) = 1$ og $\eta_{96b2}(D_1) = -1$. Summen av $\eta_t(D)$ over alle 12 terminer i 1996 og 1997 gir oss skåren S_1 for diagnostikk D, nemlig $S_1 = \sum_t \eta_t(D)$. Vi har da

$$S_1(D_1) = -2 \qquad S_1(D_{2,1}) = 4 \qquad S_1(D_{2,2}) = 8 \qquad S_1(D_{2,3}) = 7,$$

der $\eta_{96b3}(D_{2,3}) = 0$. Spesielt er en diagnostikk irrelevant dersom S_1 ligger rundt null, siden da er det som å kaste kron med den.

Definer $\psi_t(D)$ av diagnostikk D slik at, for en bestemt termin t, $\psi_t(D) = 6$ dersom den 'beste' raten ifølge D faktisk har den minste feilen, og $\psi_t(D) = 5$ dersom den 'beste' raten har den nest minste feilen, ..., og $\psi_t(D) = 1$ dersom den 'beste' raten har den største feilen. Gjennomsnittet av $\psi_t(D)$ over alle 12 terminer gir oss skåren S_2 for diagnostikk D, nemlig $S_2 = \sum_t \psi_t(D)/12$. Vi har

$$S_2(D_1) = 3.6 \qquad S_2(D_{2,1}) = 3.7 \qquad S_2(D_{2,2}) = 4.5 \qquad S_2(D_{2,3}) = 4.6,$$

der $\psi_{97b4}(D_{2,3}) = (2+6)/2 = 4$. Mens D_1, $D_{2,1}$ og $D_{2,3}$ fungerer best i 1996, er $D_{2,2}$ best i 1997.

På et lignende vis definerer vi $\phi_t(D)$ av diagnostikk D slik at, for en bestemt termin t, $\phi_t(D) = 6$ dersom den 'verste' raten ifølge D faktisk har den minste feilen, og $\phi_t(D) = 5$ dersom den 'verste' raten har den neste minste feilen, ..., og $\phi_t(D) = 1$ dersom den 'verste' raten har den største feilen. Gjennomsnittet av $\phi_t(D)$ over alle 12 terminer gir oss skåren S_3 for diagnostikk D, nemlig $S_3 = \sum_t \phi_t(D)/12$. Vi har da

$$S_3(D_1) = 4.3 \qquad S_3(D_{2,1}) = 2.3 \qquad S_3(D_{2,2}) = 1.9 \qquad S_3(D_{2,3}) = 2.4,$$

der $\phi_{96b3}(D_{2,3}) = 4$ og $\phi_{97b2}(D_{2,3}) = 5$. Spesielt er $\phi_t(D_{2,1})$ og $\phi_t(D_{2,2})$ konstant 1 i 1996.

Tabell 3. Rate-estimering for terminvis total omsetning i 1996

	96b1											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b1	2091.8		2322.2		4191.2	+	0.0013		1.111	1.126	1.35	1.7
95b2	2344.3		2784.4		-5756.8		0.0008	+	1.009	1.024	1.42	2.3
95b3	2500.8	-	2615.6		12129.4		0.0039		1.001	1.031	3.01	4.0
95b4	2228.0		1986.3	+	10319.8		0.0015		1.082	1.112	2.79	2.3
95b5	1803.8	(+)	4797.9	-	44981.6	-	0.0080	-	0.975	1.016	4.29	6.0
95b6	1733.1	+	3371.1		14888.9		0.0032		0.929	0.954	2.68	4.7
	96b2											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b2	2659.2	-	1446.8	+	-7391.4		0.0009	+	1.092	1.098	0.60	1.3
95b3	2595.6	(-)	2919.9		11883.5		0.0021		1.083	1.106	2.16	4.3
95b4	2590.2	(-)	2035.8		9285.0		0.0013		1.172	1.195	1.92	3.0
95b5	2354.9		5505.0	-	46418.5	-	0.0049	-	1.056	1.092	3.44	6.0
95b6	2132.0		3963.8		14113.3		0.0015		1.006	1.025	1.84	4.7
96b1	2016.2	+	1649.0		-1088.5	+	0.0012		1.083	1.073	-0.94	1.7
	96b3											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b3	2555.3	-	4325.2		2195.9		0.0012	+	1.067	1.077	0.91	2.3
95b4	2370.6		3964.9		-63.9	+	0.0033	-	1.154	1.162	0.68	3.0
95b5	2000.2		7254.5	-	36069.4	-	0.0028		1.039	1.063	2.26	5.3
95b6	1888.3	+	6518.3		4614.5		0.0013	(+)	0.991	0.997	0.68	3.3
96b1	1945.5		1850.1		-10103.5		0.0033	-	1.066	1.044	-2.04	4.3
96b2	1949.0		1661.8	+	-9092.9		0.0015		0.984	0.973	-1.15	2.7
	96b4											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b4	2324.6	-	5637.3		17825.6		0.0011	(+)	1.126	1.136	0.91	3.3
95b5	2103.5		9770.2	-	53233.2	-	0.0060	-	1.013	1.039	2.51	6.0
95b6	2023.8		7171.1		22462.2		0.0028		0.967	0.975	0.87	5.0
96b1	1987.8		4529.8		7184.7	+	0.0014		1.039	1.022	-1.62	2.3
96b2	1860.5		2883.5	+	8221.2		0.0010	+	0.959	0.953	-0.67	1.3
96b3	1570.1	+	4268.3		17116.9		0.0021		0.977	0.979	0.24	3.0
	96b5											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b5	2227.4	(-)	8595.1	-	39377.9	-	0.0024		1.149	1.176	2.36	5.3
95b6	2243.8	-	7318.2		4900.6		0.0020		1.095	1.103	0.68	3.0
96b1	2225.2	(-)	1851.4		-12052.8		0.0046	-	1.177	1.156	-1.86	4.3
96b2	2098.0		1336.6		-10318.6		0.0024		1.087	1.076	-0.98	2.7
96b3	1790.9		1001.1	+	65.5	+	0.0006	+	1.104	1.105	0.05	1.0
96b4	1491.3	+	4377.5		-20192.2		0.0030		1.131	1.131	-0.07	4.7
	96b6											
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
95b6	2301.0		9394.6	-	31210.7	-	0.0023		1.161	1.18	1.63	5.7
96b1	2483.0	-	3713.4		13648.0		0.0018		1.247	1.235	-0.94	3.0
96b2	2136.5		1823.6	+	15110.1		0.0008	+	1.151	1.151	-0.08	1.7
96b3	2135.7		4565.8		25690.2		0.0017		1.171	1.183	0.98	3.7
96b4	2099.1		4834.8		4655.6	+	0.0011		1.198	1.209	0.89	2.7
96b5	1881.6	+	2760.9		26158.9		0.0027	-	1.059	1.069	0.95	4.3

Tabell 4. Rate-estimering for terminvis total omsetning i 1997

	97b1										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b1	2094.4	-	3748.5	-	6605.1		0.0015		1.091	1.088	-0.25	4.3
96b2	2021.0	(-)	2151.1		8107.0		0.0007	(+)	1.007	1.014	0.62	3.0
96b3	1952.1		2939.0		17638.3	(-)	0.0022		1.023	1.041	1.84	3.5
96b4	1749.8		829.2	+	-1032.6	+	0.0006	+	1.047	1.065	1.69	1.0
96b5	1583.4	+	2101.9		18089.1	-	0.0028	-	0.925	0.941	1.79	4.7
96b6	1733.1		2423.1		-5468.9		0.0009		0.873	0.881	1.02	3.0
	97b2										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b2	2090.9	(-)	3912.6		-10921.2		0.0016		1.036	1.033	-0.34	3.7
96b3	2158.0	-	3719.2		-1300.7		0.0008	+	1.052	1.061	0.87	2.0
96b4	2057.3		3126.5		-19869.5		0.0021	-	1.077	1.084	0.64	4.0
96b5	1876.3		3803.6		-304.2	+	0.0009	(+)	0.951	0.958	0.78	2.3
96b6	1977.8		6540.2	-	-24274.8	-	0.0021	-	0.897	0.897	0.03	6.0
97b1	1588.9	+	2301.7	+	-18902.5		0.0019	(-)	1.028	1.017	-1.14	3.0
	97b3										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b3	2074.7		3282.8		-7399.7		0.0009	(+)	1.074	1.079	0.46	2.7
96b4	2209.4	-	4330.7		-26258.5		0.0029	-	1.099	1.102	0.26	5.3
96b5	2064.1		3840.7		-6778.7		0.0007	+	0.971	0.975	0.42	2.3
96b6	2010.9		6044.7	-	-30811.5	-	0.0025		0.918	0.912	-0.59	5.3
97b1	1881.0	(+)	3726.3		-25484.5		0.0028	(-)	1.052	1.034	-1.71	4.0
97b2	1800.4	+	1966.2	+	-6119.5	+	0.0008	(+)	1.022	1.016	-0.62	1.3
	97b4										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b4	2057.2		3326.5	+	7434.0		0.0008	+	1.080	1.100	1.85	1.3
96b5	2181.1	(-)	11264.2	-	27178.7		0.0042	-	0.953	0.973	2.04	5.3
96b6	2227.5	-	9050.9		3059.1	+	0.0014		0.900	0.910	1.16	3.0
97b1	1865.0	(+)	5391.1		8349.0		0.0008	+	1.032	1.032	0.01	2.3
97b2	1823.0	(+)	6117.2		27511.2		0.0027		1.002	1.014	1.16	4.3
97b3	1820.8	+	5865.3		33256.2	-	0.0033		0.981	0.998	1.77	4.7
	97b5										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b5	2242.6		6276.7		-18552.5		0.0011		1.061	1.061	0.02	3.0
96b6	2442.0	-	10229.1	(-)	-44997.8		0.0037		1.000	0.993	-0.71	4.7
97b1	2257.1		7345.6		-39075.0		0.0039	(-)	1.148	1.125	-2.01	4.3
97b2	2090.4		2846.7		-17982.4		0.0011		1.114	1.105	-0.82	2.0
97b3	1702.2	+	1084.2	+	-12052.9	+	0.0007	+	1.089	1.087	-0.21	1.0
97b4	2044.2		10401.7	-	-47398.6	-	0.0044	-	1.112	1.090	-2.00	6.0
	97b6										
Termin	D_1		$D_{2,1}$		$D_{2,2}$		$D_{2,3}$		$\hat{\beta}$	β	(%)	
96b6	2511.1	(-)	9889.1	-	-6607.7		0.0011	+	1.030	1.042	1.17	3.0
97b1	2548.4	-	5359.1		-676.2	+	0.0014		1.180	1.18	-0.03	3.0
97b2	2439.1		3648.3		21554.8		0.0019		1.145	1.159	1.22	3.7
97b3	2280.2		2061.9	+	28100.3		0.0022		1.121	1.142	1.83	3.7
97b4	2410.4		4887.8		-9765.9		0.0012	(+)	1.144	1.143	-0.06	3.0
97b5	2168.5	+	3538.7		40775.5	-	0.0032	-	1.029	1.050	2.01	4.7

Diagnostikk D_1, dermed modell (3), har den laveste S_1. Spesielt ille er at S_3 er enda høyere enn S_2. Mao. modell (3) passer ikke til den type data vi behandler her. Mens $D_{2,2}$ virker som den beste diagnostikk under modell (5), når det gjelder å skille den mest misvisende foregående termin fra den 'beste'; $D_{2,1}$ er muligens den dårligste der.

For å kombinere alle tre diagnostikker under modell (5) skal vi, for hver av de 12 terminene, prøve å rangere de foregående terminene som følgende: (i) for hver foregående termin t, la $\kappa(t;D) = 1$ hvis den gir den 'beste' raten ifølge diagnostikk D, og $\kappa(t;D) = 2$ hvis den gir den 'nest beste', ... og $\kappa(t;D) = 6$ hvis den gir den 'verste' raten, (ii) gjennomsnittet til $\kappa(t;D)$ over alle tre diagnostikker gir oss "kreditt" til termin t, nemlig $\kappa(t) = \sum_D \kappa(t;D)/3$. Denne er oppgitt i den siste kolonnen til de to tabellene ovenfor. Spesielt har vi beregnet S_1 - S_3 for κ, nemlig

$$S_1(\kappa) = 8 \qquad S_2(\kappa) = 4.3 \qquad S_3(\kappa) = 1.9.$$

Det synes at κ stort sett har klart å kombinere diagnostikk $D_{2,1}$ til $D_{2,3}$, muligens på en liten bekostning mht. S_2.

B En ad hoc ordenvektingsmetode

Vi betegner som vanlig populasjonen med U og utvalget med s. Vi ordner omsetning i utvalget som $x_s = \{x_{(1)}, ..., x_{(n)}\}$, slik at $x_{(1)} \leq \cdots \leq x_{(n)}$. Anta at $x_{(1)}$ også er den minste og $x_{(n)}$ den største i hele populasjonen. Sett $w(i) = 0$ for $i \in s$, unntatt $w(i) = 1$ for $x_{(1)}$. For hver $x_{(1)} < x_i \leq x_{(2)}$ der $i \in U$, oppdaterer man $w(1)$ med å legge til $(x_{(2)} - x_i)/(x_{(2)} - x_{(1)})$, og $w(2)$ med $(x_i - x_{(1)})/(x_{(2)} - x_{(1)})$. Gå gjennom alle x_i der $x_{(2)} < x_i \leq x_{(3)}$, og oppdater $w(2)$ og $w(3)$ på samme måten. Gjenta dette til man er ferdig $x_{(n-1)}$ og $x_{(n)}$.

Metoden gir oss vekter som er kalibrert i den forstand at $\sum_{i\in s} w_i x_i = X$ og $\sum_{i\in s} w_i = N$, siden $\forall i \in U$ s.a. $x_{(j)} < x_i \leq x_{(j+1)}$ for viss $1 \leq j \leq n-1$, har man at

$$x_i \equiv \frac{x_{(j+1)} - x_i}{x_{(j+1)} - x_{(j)}} x_{(j)} + \frac{x_i - x_{(j)}}{x_{(j+1)} - x_{(j)}} x_{(j+1)} \qquad \text{og} \qquad \frac{x_{(j+1)} - x_i}{x_{(j+1)} - x_{(j)}} + \frac{x_i - x_{(j)}}{x_{(j+1)} - x_{(j)}} \equiv 1.$$

Det er allikevel en del betraktninger som er viktig her: (a) det er ikke opplagt hva man skal gjøre dersom den minste og den største bedriften *ikke* er med i utvalget; (b) man kan lett ende opp med ekstremt store vekter for små bedrifter i utvalget — da de er undertrukket, dermed ustabilitet i estimering; (c) det er vanskelig å utvide metoden dersom multivariat tilleggsinformasjon er tilgjengelig — man kjenner jo her til omsetning fra flere tidligere terminer, siden orden i utvalget endrer seg fra gang til gang.

C To robuste rate-estimater i litteraturen

C.1 Varians-inflasjonsmodell og nedveiing av ekstreme enheter

Ghangurde (1989) beskrev en varians-inflasjons regresjons modell. I tilfellet ratemodell blir den, $\forall i \in s$, $y_i = \beta x_i + \epsilon_i$, for $\epsilon_i \sim N(0, \sigma_i^2)$, der $\sigma_i^2 = \sigma^2 w_i$ for ikke-ekstreme i, og $\sigma_i^2 = \sigma^2 w_i / w$ for ekstreme i, der $0 < w \leq 1$ og w_i typisk en funksjon av x_i. Den beste lineære forventningsrette estimatoren for β, betegnet med $\tilde{\beta}$, der $w_i = x_i$, er gitt ved

$$\tilde{\beta} = \frac{wk\bar{y}_k + (n-k)\bar{y}_{n-k}}{wk\bar{x}_k + (n-k)\bar{x}_{n-k}} \qquad \text{og} \qquad w = \frac{\sigma_{2x}^2\mu_y^2 + \sigma_{2y}^2\mu_x^2 - 2\sigma_{2xy}\mu_x\mu_y}{\sigma_{1x}^2\mu_y^2 + \sigma_{1y}^2\mu_x^2 - 2\sigma_{1xy}\mu_x\mu_y},$$

der $(\bar{y}_k, \bar{x}_k)$ er gjennomsnittet blant ekstremene og $(\bar{y}_{n-k}, \bar{x}_{n-k})$ blant ikke-ekstremene, og σ_1. for ekstremene og σ_2. for ikke-ekstremene. Spesielt brukte Ghangurde (1989) den såkalte "Cook's distance (CD)" (Cook, 1979) til å identifisere de k ektreme enhetene. Akkurat hvor grensen til CD skal ligger i en bestemt situasjon, finnes det ikke noe generelt svar på.

C.2 Robust regresjon

Rousseeuw and Leroy (1987) arbeidet ut ifra følgende kriterium for robustheten. I tilfellet ratemodell, la $Z = \{(x_1, y_1), ..., (x_n, y_n)\}$ være parvis observasjoner. Anta nå at m av disse er "forgiftet", dvs. byttet ut med vikårlige verdier, betegnet med Z'. La $\hat{\beta}(Z)$ være estimatoren basert på Z. Definer forgiftningseffekt som

$$b(m; T, Z) = \sup_{Z'} \|T(Z') - T(Z)\|,$$

dvs. over alle mulige Z'. Uendelig b betyr at m ekstremer kan ha vikårlig stor innflytelse på T. Man kan derfor definere et sammenbruddspunkt til T som

$$\gamma^*(T, Z) = \min\{m/n; b(m; T, Z) = \infty\},$$

dvs. den minste forgiftningsandelen som skal til for at T kan bryte sammen.

Det viser seg for at det minste kvadraters estimatet (LS), gitt ved $\min_\beta(\sum_{i=1}^n \epsilon_i^2)$, er $\gamma^* = 1/n$, dvs. det er nok med en ekstrem enhet, og 0% asymptotisk når $n \to \infty$. Derimot har *det minste kvadraters median (LMS)*, gitt ved $\min_\beta\{Median_i(\epsilon_i^2)\}$, asymptotisk et sammenbruddspunkt på 50%, som også er det maksimale. Geometrisk går LMS-metoden ut på å finne den smaleste belte som inneholder halvparten av observasjonene. Metoden har veldig sakte konvergens, så Rousseeuw and Leroy (1987) har foreslått et alternativ, nemlig *det minste trimmede kvadratet (LTS)*, gitt ved

$$\min_\beta\{\sum_{i=1}^{h}(r^2)_{(i)}\} \qquad 1 \leq h \leq n,$$

der $(r^2)_{(i)}$ er det i-te ordnete kvadratiske restleddet. Dermed unngår LTS-metoden de største $n - h$ kvadratiske restleddene. Det omtrent optimale valget på h er $n/2$. Både LMS og LTS er implementert i S-Plus, og Venables and Ripley (1994) anbfalte LTS framfor LMS.

C.3 En liten sammenligning

En liten sammenligning foregikk på følgende måte: For hver termin fra 96b1 til 97b6, i alt 12 av dem, har vi beregnet tre rate-estimater, nemlig (i) det vanlige $\hat{\beta}$ (4), (ii) $\tilde{\beta}$ av Ghangurde (1989), og (iii) $\hat{\beta}_{LTS}$ av Rousseeuw and Leroy (1987), basert på det samme termin fra fjoråret, dvs. 95b1 for 96b1, 95b2 for 96b2, osv.. Hvert rate-estimat β_{est} gir oss et estimat for total omsetning ved $Y_{est} = Y_s + \beta_{est}(X - X_s)$, og en relativt feil ved $Y_{est}/Y - 1$.

Resultatene er samlet i Tabell 5. Spesielt har man forsøksvis satt grensen til CD lik 1 for $\tilde{\beta}$, der det for 96b3 var for få ekstremer til å regne ut estimatet. Det virker ikke som om $\tilde{\beta}$ og $\hat{\beta}_{LTS}$ representerer betydelig forbedring overfor det vanlige $\hat{\beta}$ — $\tilde{\beta}$ er kanskje enda dåligere enn $\hat{\beta}$. Heller ikke er LTS-metoden spesielt robust, der den relative feilen svinger fra -1.96% til 1.63%, sammenlignet med $\hat{\beta}$ der feilen svinger fra -0.34% til 2.36%.

Tabell 5. Sammenligning av relativ feil i Y_{est} (%) med forskjellige rateestimering

Metode	96b1	96b2	96b3	96b4	96b5	96b6	97b1	97b2	97b3	97b4	97b5	97b6
$\hat{\beta}$	1.35	0.60	0.91	0.91	2.36	1.63	-0.25	-0.34	0.46	1.85	0.02	1.17
$\tilde{\beta}$	1.78	1.56	-	0.58	2.62	1.47	-0.66	0.12	0.90	1.68	0.79	1.85
$\hat{\beta}_{LTS}$	-0.03	-0.01	-1.52	-0.99	-0.85	-1.96	-1.08	-0.12	-0.61	-0.56	-0.38	1.63
$\hat{\beta}^*$	0.97	0.31	0.46	-0.98	-0.15	0.76	-1.32	0.35	0.20	0.89	0.66	1.04

D Dempet rateestimering

Vi har beregnet det dempete rate-estimat (6) for alle 12 terminene fra 1996 til 1997, basert på den samme termin fra fjoråret, dvs. 95b1 for 96b1, 95b2 for 96b2, osv.. (Bruk av andre foregående terminer ga ikke bedre resultater.) For enkelthets skyld har vi satt $t_y = t_x = t$ i (6). Med $t = 4^5$ fikk vi den relative feilen i Y_{est} til å svinge fra -1.32% til 0.97%, som representerer en forbedring over det vanlige $\hat{\beta}$ (4), og $\tilde{\beta}$ (Ghangurde, 1989), og $\hat{\beta}_{LTS}$ (Rousseeuw and Leroy, 1987).

Vi har også beregnet

$$\hat{Y}_X^* = Y_s + \hat{\beta}(X - X_s)^*,$$

med $\hat{\beta}$ fra (4), men demping i populasjonen utenfor utvalget. Mao. vi vil sjekke om demping i tidligere omsetning kunne gi den samme robustisering som demping i rate-estimatet. Med snittpunkt ("cut-off") på 5.5×10^5 fikk vi de beste resultatene, der den relative feilen svinger fra -0.49% til 2.18%. Dette ligner svært på resultatene man fikk hos det vanlige $\hat{\beta}$, noe nedover justert naturligvis. Dette kan tyde på at de forbedrete resultatene hos (6) virkelig kommer av en forbedret $\hat{\beta}^*$, overfor $\hat{\beta}$.

Til slutt har vi forsøkt med et rate-estimat basert på medianene, nemlig

$$\hat{\beta}_{Med} = \text{Median}_{i \in s}(y_i)/\text{Median}_{i \in s}(x_i).$$

Dette resulterte i nærmest dobbelt så store svingninger i den relative feilen sammenlignet med den vanlige $\hat{\beta}$. Men dersom man anvender $\hat{\beta}_{Med}$ på publikasjonsnivå, dvs. innen de forskjellige sektorene, og etterpå estimerer totalen i populasjonen ved å summere alle disse sammen, så svinger den relative feilen bare fra -0.91% til 1.00%. Om dette skyldes tilfeldigheter trenger man flere studier for å svare på. Det er allikevel noen betraktninger som er viktige, nemlig (i) selv med en total telling, dvs. $s = U$, vil $\hat{\beta}_{Med}$ generelt ikke være lik Y/X, i motsetning til $\hat{\beta}$, og (ii) pga. overtrekking blant de store bedriftene, er medianen i utvalget langt fra medianen i populasjonen, dermed er det slett ikke klart at medianen i utvalget er mer robust enn gjennomsnittet i utvalget.

References

Cook, R.D. (1979). Influential observations in linear regression. *J. Am. Statist. Assoc.*, **74**, 169–74.

Ghangurde, P.D. (1989). Ouliers in sample surveys. In *Proceedings for the Survey Research Methods Section, American Statistical Association*, pp. 736–9.

Lee, H. (1995). Ouliers in business surveys. In *Business Survey methods*, chap. 26, pp. 503–26. John Wiley & Sons, Inc.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-Plus*. Sringler.

De sist utgitte publikasjonene i serien Notater

98/42 M.V. Dysterud og P. Schøning: Etterprøvbare miljømål for byer og tettsteder: Et metodeprosjekt for utvikling og prøving av miljøindikatorer. 40s.

98/43 J. Epland: Inntekt etter skatt: Revisjon av inntektsregnskapet i inntekts- og formuesundersøkelsen for husholdninger. 40s.

98/44 E. Sørensen: Produksjonsindeks for industrien. 48s.

98/45 L. Aaram og Ø. Skullerud: Statistikk over emballasjeavfall: Utprøving av metode og foreløpige resultater. 32s.

98/46 L-C. Zhang: Empirisk imputering: En ny metode for å behandle tilfeldig partielt frafall. 20s.

98/47 L. Dalen, P.M. Bergh, J-A. Sigstad Lie og A. Vedø: Energibruk i næringsbygg 1995-1997. 69s.

98/48 B. Strand og H. Utne: FoB2000: Rapport fra seminar 12. mars 1998 om arbeidsmarkedsdelen i Folke- og boligtellingen 2000. 33s.

98/49 N.Ø. Mæhle og K. Nyborg: Energibruk og utslipp til luft i norsk produksjon: Direkte og indirekte virkninger. 23s.

98/50 T. Eidem og J. Lajord: FD-Trygd. Dokumentasjonsrapport: Utdanning 1992-1993. 87s.

98/51 A. Bjerkestrand og S. Fjeld: Regnskapsstatistikk for aksjeselskaper 1996: Dokumentasjon. 34s.

98/52 G. Haakonsen, S. Holtskog og B. Tornsjø: Energibruk og utslipp til luft i Oslo, Drammen, Bergen og Trondheim 1995. 57s.

98/53 E. Holmøy: Hvordan generelle likevektseffekter bidrar til prisfølsomheten i den norske el-etterspørselen: Dokumentasjon av beregningsrutiner. 33s.

98/54 F.R. Aune, T. Bye, M.I. Hansen og T.A. Johnsen: Kraftpris og skyggepris på CO_2 - utslipp i Norge til 2027. 13s.

98/55 S. Blom: Holdning til innvandrere og innvandringspolitikk: Spørsmål i SSBs omnibus i mai/juni 1998. 34s.

98/56 K. Bjønnes og B.R. Joneid: FD - Trygd. Dokumentasjonsrapport: Foreløpig uførestønad, 1992-1993. 34s.

98/57 T. Bye: Fleksibel gjennomføring av en klimaavtale. 27s.

98/58 K.J. Einarsen (red.): Arbeidsutvalgets evaluering av faktaark for FylkesKOSTRA-utdanning: 1. tertial 1998. Sør-Trøndelag fylkeskommune. 33s.

98/59 I. Øyangen: Inntekts- og formueundersøkelsen 1997: Dokumentasjonsrapport. 23s.

98/60 B. Olsen og I. Tuveng: Utvalgsundersøkelsen om sykefravær, 1-3 dager for 3. kvartal 1997: Dokumentasjon. 19s.

98/61 E. Rønning: Barnefamiliers tilsynsordninger, yrkesdeltakelse og økonomi før innføring av kontantstøtte: Hovedresultater og dokumentasjon. 138s.

98/62 A.G. Hustoft: Forslag til ny regional inndeling: Etablering av publiseringsnivå mellom fylke og kommune. 61s.

98/63 H.M. Edvardsen: Fylkesfordelt nasjonalregnskap 1993: Resultater og metoder. 30s.

98/65 T. Vogt: Næringslivets kostnader ved lover og regelverk: Dokumentasjonsrapport. 34s.

98/66 M. Søberg: Omsetjelege kvotar og internasjonale miljøavtalar. 15s.

98/67 J. Lindstrøm: Dokumentasjon: Kvartalsvis kraftprisstatistikk. 44s.

98/68 P. Schøning: Oppsummering av høring angående metode for tettstedavgrensing 1998. 53s.

98/69 J. I. Røstadsand: Husholdningssektoren i nasjonalregnskapet: Sektorer og undergrupper. 18s.

98/70 E. Skaansar: Nasjonalregnskap: Beregning av næringene for elektrisitet og fjernvarme. 32s.

Notater

PORTO BETALT
VED
INNLEVERINGA
P.P.
NORGE/NOREG

Tillatelse nr.
159 000/502

B *Returadresse:*
Statistisk sentralbyrå
Postboks 8131 Dep.
N-0033 Oslo

Statistisk sentralbyrå

Oslo:
Postboks 8131 Dep.
0033 Oslo

Telefon: 22 86 45 00
Telefaks: 22 86 49 73

Kongsvinger:
Postboks 1260
2201 Kongsvinger

Telefon: 62 88 50 00
Telefaks: 62 88 50 30

ISSN 0806-3745

Statistisk sentralbyrå
Statistics Norway