Statistics Norway
Statistisk sentralbyrå

Anna-Karin Mevik

Usikkerhet i ordrestatistikken

Seksjon for statistiske metoder og standarder

1. Innledning

I dette notatet skal vi presentere et usikkerhetsmål for den kvartalsvise ordrestatistikken. Selv om det i utvalgsundersøkelser er vanlig å beskrive usikkerheten ved hjelp av utvalgsvariansen, har vi valgt å bruke et modellbasert usikkerhetsmål. Begrunnelsen for dette er at utvalget som brukes i ordrestatistikken ble trukket i 1996. Etter 1996 er det kun gjort oppdateringer og suppleringer av utvalget.

Ordrestatistikken er en kvartalsvis statistikk der populasjonens samlede ordretilgang og ordrereserve, for inneværende kvartal, blir estimert. Basert på disse estimatene blir det beregnet verdiindekser. I dette notatet er det verdiindeksen for ordretilgangen som vi skal lage et usikkerhetsmål for.

Vi starter i avsnitt 2 med å gi en presentasjon av populasjonen. I avsnitt 3 gis en beskrivelse av utvalget, og hvordan det ble trukket i 1996. I avsnitt 4 presenteres estimatoren og verdiindeksen for ordretilgangen. I avsnitt 5 gjør vi en modellbasert analyse som leder frem til et usikkerhetsmål. Dette målet har vi så benyttet til å estimere usikkerheten i ordrestatistikken for 1. og 2. kvartal 2004. Resultatet av dette gir vi i avsnitt 6. Til slutt gir vi en oppsummering i avsnitt 7. (For en mer detaljert gjennomgang av ordrestatistikken enn det som gis i dette notatet, se Bakken og Osnes, 1998).

2. Populasjon

Populasjonen omfatter alle bedrifter innen de ordrebaserte næringene (unntatt enmannsbedrifter). En ordrebasert næring kjennetegnes ved at ferdigstillingen av en ordre vanligvis tar lengre tid enn ett kvartal. Per i dag er følgende næringer definert som ordrebasert: Tekstil- og bekledningsindustri, treforedlingsindustri, kjemisk industri, metall- og metallvareindustri, maskinindustri, elektroteknisk og optisk industri, oljeplattformer og moduler, og transportmiddelindustri.

Det er BoF (Bedrifts- og foretaksregisteret) som bestemmer hvilke bedrifter som til en hver tid er med i populasjonen. Populasjonen blir oppdatert mot BoF en gang hvert kvartal. I 2004 var det ca. 6800 bedrifter i populasjonen.

Populasjonen er delt inn i 12 delpopulasjoner på følgende måte:

Notasjon	Næring	Næringsstandard
SNN17	Tekstilindustri	17
SNN18	Bekledningsindustri	18
SNN21	Treforedlingsindustri	21
SNN24\241	Kjemikalier og kjemiske produkter unntatt kjemiske råvarer	24 unntatt 241
SNN241	Kjemiske råvarer	241
SNN27\274	Metallindustri unntatt ikke-jernholdige metaller	27 unntatt 274
SNN274	Ikke-jernholdige metaller	274
SNN28	Metallvareindustri	28
SNN29	Maskinindustri	29
SNN30_33	Elektronisk og optisk industri	30, 31, 32 og 33
TDM2	Oljeplattformer og moduler	35114, 35115
TDM1	Transportmiddelindustri	34, 35 unntatt 35114 og 35115

2.1. Ordretilgang

Ordretilgangen til en bedrift, for et gitt kvartal, er definert som verdien av alle ordre og bestillinger bedriften mottar i løpet av kvartalet. Verdien gis i løpende priser, og det er bare snakk om ordre og bestillinger på varer og tjenester som bedriften har eller skal produsere. Dvs. at handelsvarer ikke medregnes.

Som notasjon på ordretilgangen bruker vi y_i for en bedrift *i*. Den totale ordretilgangen for en delpopulasjon er dermed gitt ved

$$Y = \sum_{i \in U} y_i \ ,$$

der U er indeksmengden for delpopulasjonen.

2.2. Omsetning

Ved estimering av ordretilgangen brukes rate-estimatoren, hvor bedriftenes omsetning er tilleggsvariabel. Fordi vi på estimeringstidspunktet ikke kjenner bedriftenes omsetning for gjeldende kvartal, brukes gjennomsnittlige omsetningstall fra foregående år. Hovedsakelig er disse tallene hentet fra momsregisteret, men når disse mangler brukes tall fra BoF.

Vi benytter med andre ord omsetningstall fra to kilder:

$z_i =$ gj.sn. kvartalsvis omsetning for bedrift *i*, hentet fra BoF

$v_i =$ gj.sn. kvartalsvis omsetning for bedrift *i*, hentet fra momsregisteret .

(Med gjennomsnittlig kvartalsvis omsetning menes årlig omsetning delt på fire).

Både z_i og v_i er ukjent for en del bedrifter (men ikke nødvendigvis for de samme bedriftene). I 1. kvartal 2004 var f.eks. z_i ukjent for 789 av bedriftene i populasjonen, mens v_i var ukjent for 962 av bedriftene. Videre er ikke z_i og v_i like, og for noen bedrifter kan forskjellen være stor.

I estimeringen av ordretilgangen er det følgende omsetningstall som benyttes:

$$x_i = \begin{cases} v_i \text{ ,når denne er kjent} \\ z_i \text{ ,ellers} \end{cases} .$$

Dessverre vil også x_i være ukjent for en del bedrifter. I 1. kvartal 2004 var f.eks. x_i ukjent for 289 bedrifter.

3. Utvalg

I 1996 ble det trukket et stratifisert enkelt tilfeldig utvalg på ca. 750 bedrifter. Etter dette er det ikke trukket noe nytt utvalg. Det er kun gjort supplering og oppdatering av utvalget. I dag er ca. 800 bedrifter med i utvalget.

Som stratifiseringsvariabler brukes næring og sysselsetting. Næring er delt inn i ca. 60 grupper, mens sysselsetting er delt inn i intervallene 0-9, 10-19, 20-49, 50-99 og 100-. (Noen få bedrifter med færre enn 100 sysselsatte blir plassert sammen med bedriftene som har 100 eller flere sysselsatte. Dette er bedrifter som har, eller har hatt, veldig stor omsetning).

Når utvalget ble trukket i 1996 var det fulltelling av strataene der bedriftene hadde 100 eller flere sysselsatte. Det ble ikke trukket utvalg fra strataene der sysselsettingen var mindre enn 10. I de resterende strataene ble det trukket enkelt tilfeldig utvalg. (For en mer detaljert beskrivelse av utvalgsplanen, se Dokumentasjon av utvalgsplan).

4. Estimering av ordretilgangen

Når ordretilgangen blir estimert i den kvartalsvise ordrestatistikken, blir noen få av bedriftene i nettoutvalget klassifisert som ekstreme (klassifiseringen skjer på grunnlag av forholdet mellom ordretilgangen y_i og omsetningen x_i til bedriftene i utvalget). Disse bedriftene blir så holdt utenfor ved estimeringen. Dessverre klarer vi ikke å lage et usikkerhetsmål som tar hensyn til dette. I stedet lager vi et usikkerhetsmål for estimatoren slik den er når man ikke klassifiserer og skiller ut bedrifter som ekstreme.

Når vi ikke klassifiserer og skiller ut bedrifter som ekstreme, blir ordretilgangen $Y = \sum_{i \in U} y_i$ for en delpopulasjon estimert med

$$(1) \qquad \hat{Y} = \frac{\sum_{i \in s_*^c} y_i}{\sum_{i \in s_*^c} x_i} \cdot X_V + \sum_{i \in s_*} y_i \,,$$

der

$s_* = \{ i \in s \text{ s.a. } syss_i \geq 100 \}$,

$syss_i$ er sysselsettingen til bedrift i ($syss_i$ er ukjent for de samme bedriftene hvor z_i er ukjent),

s er indeksmengden for bedriftene i nettoutvalget (pluss noen få bedrifter i frafallsgruppen hvor ordretilgangen blir imputert),

$s_*^c = s \setminus s_*$

$X_V = \sum_{i \in V} x_i$

og

$V = \{ i \in U \text{ s.a. } syss_i < 100 \text{ (ev. ukjent), og s.a. } x_i \text{ er kjent} \}$.[1]

[1] Estimatoren (1) forutsetter at x_i er kjent for alle bedriftene i s_*^c. I 1. og 2. kvartal 2004 er det kun en bedrift i utvalget hvor x_i er ukjent. For å komme frem til en mer generell formel for $\hat{Y}$ som også inkluderer denne situasjonen, må dataprogrammene som beregner estimatene gjennomgås veldig nøye. Dette vil kreve en del arbeid. I tillegg må vi regne

(Det kan finnes noen få bedrifter i s_* som har $syss_i < 100$. Dette er bedrifter som har/har hatt stor omsetning, og disse blir ikke inkludert i V).

(Hvis vi lar $s_* = \{i \in s \text{ s.a. } syss_i \geq 100, \text{ eller s.a. bedriften klassifiseres som ekstrem}\}$ og $V = \{i \in U \text{ s.a. } syss_i < 100 \text{ (ev. ukjent), og s.a. } x_i \text{ er kjent, og s.a. bed. ikke er klassifisert som ekstrem}\}$ får vi den estimatoren som benyttes i den kvartalsvise ordrestatistikken).

Merk at vi her begrenser oss til en delpopulasjon. Dvs. at notasjonen gjelder en delpopulasjon, slik at f.eks. s er indeksmengden for bedriftene i nettoutvalget som tilhører delpopulasjonen.

Verdiindeksen for ordretilgangen er gitt ved

$$\hat{d} = \frac{\hat{Y}}{Y95} \cdot 100,$$

der $Y95$ er den estimerte ordretilgangen for 1995.

Som vi ser av estimatoren (1), er det noen bedrifter som det ikke vektes for, dvs. som det ikke gjøres noe forsøk på å estimere ordretilgangen til. Dette er bedrifter utenom utvalget s, hvor $syss_i \geq 100$ eller x_i er ukjent. Fordi nesten alle delpopulasjonene har slike bedrifter, må vi regne med at $\hat{Y}$ som regel underestimerer Y. En alternativ estimator som ikke har denne skjevheten er

$$\begin{aligned}\tilde{Y} &= \hat{Y} + \frac{|U_* \setminus s_*|}{|s_*|}\sum_{i\in s_*} y_i + \frac{M}{m}\sum_{i\in s_m} y_i \\ &= \frac{\sum_{i\in s_*^c} y_i}{\sum_{i\in s_*^c} x_i} \cdot X_V + \frac{|U_*|}{|s_*|}\sum_{i\in s_*} y_i + \frac{M}{m}\sum_{i\in s_m} y_i ,\end{aligned}$$

der $U_* = \{i \in U \text{ s.a. } syss_i \geq 100\}$, M er antall bedrifter i populasjonen hvor x_i er ukjent, s_m er et utvagl av disse bedriftene, og $m = |s_m|$. (For å kunne bruke denne estimatoren kreves det med andre ord at man også trekker et utvalg av bedrifter hvor x_i er ukjent).

5. Modellbasert usikkerhetsmål

I dette avsittet skal vi lage et usikkerhetsmål for verdiindeksen $\hat{d} = \left(\hat{Y}/Y95\right)\cdot 100$. Som begrunnet i innledningen skal vi lage et modellbasert usikkerhetsmål. Dvs. at vi skal se på bedriftenes ordretilgang som stokastiske variabler, mens utvalget antas gitt. Vi vil se bort fra frafall, målefeil og registerfeil. Videre vil vi behandle $Y95$ som en konstant, og se på $\hat{d}$ som en estimator (eller prediktor) for $d = \left(Y/Y95\right)\cdot 100$.

---

med at det trengs en del mer notasjon for denne estimatoren. Av disse grunner velger vi å holde oss til estimatoren (1). Dvs. at vi lager et usikkerhetsmål under forutsetningen om at x_i er kjent for alle bedriftene i s_*^c.

Siden bedriftenes ordretilgang nå antas å være stokastiske variable, må vi ha en modell som kan beskrive variasjonen i y_i 'ene. Vi har prøvd med ulike modeller, og for de fleste delpopulasjonene har vi kommet frem til en modell som ser ut til å passe relativt godt.[2] Men for noen delpopulasjoner har vi hatt problem med å finne en passende modell, og for disse har vi måttet velge en modell som ikke klarer å beskrive variasjonen i y_i 'ene så veldig godt.

Valg av estimatoren $\hat{Y}$ er gjort ut fra en antagelse om at $y_i \sim \left(\beta x_i, \sigma^2 x_i\right)$, dvs. en regresjonsmodell med x_i som kovariat. For dataene fra 1. og 2. kvartal 2004 viser det seg at z_i egner seg bedre som forklaringsvariabel, og for de fleste delpopulasjonene har vi valgt å modellere y_i 'ene med en regresjonsmodell der z_i er kovariat. For noen av delpopulasjonene ser det derimot ikke ut til å passe med regresjonsmodell, verken med x_i eller z_i som kovariat (vi har også prøvd med andre forklaringsvariable). For disse har vi valgt å splitte opp delpopulasjonen med hensyn på sysselsetting, og anta at ordretilgangen er identisk fordelt innen hvert sysselsettingsintervall. For de resterende delpopulasjonene har vi endt opp med en kombinasjon av de to nevnte modellene. I underavsnittene 5.1, 5.2 og 5.3 skal vi gi en mer presis beskrivelse av modellene.

Uavhengig av modell kan forventningsskjevheten til $\hat{d}$ skrives som

$$\begin{aligned} \mathrm{B}\left(\hat{d}\right) &= \mathrm{E}\left[\hat{d} - d\right] \\ &= \frac{\mathrm{E}\left[\hat{Y} - Y\right]}{Y95} \cdot 100 \,. \end{aligned}$$

Dvs. at $\hat{d}$ er forventningsrett hvis, og bare hvis, $\hat{Y}$ er forventningsrett. Som vi skal se senere er dette ikke tilfelle under noen av modellene.

Fordi $\hat{d}$ ikke er forventningsrett er

$$\sqrt{\mathrm{E}\left[\left(\hat{d} - d\right)^2\right]} = \sqrt{\mathrm{V}\left(\hat{d} - d\right) + \left(\mathrm{B}\hat{d}\right)^2}$$

et fornuftig mål på usikkerheten. Men fordi vi ikke klarer å estimere forventningsskjevheten $\mathrm{B}\left(\hat{d}\right)$, velger vi å måle usikkerheten med standardavviket

$$\begin{aligned} \mathrm{st}\left(\hat{d} - d\right) &= \sqrt{\mathrm{V}\left(\hat{d} - d\right)} \\ &= \frac{\sqrt{\mathrm{V}\left(\hat{Y} - Y\right)}}{Y95} \cdot 100 \,. \end{aligned}$$

Vi trenger å estimere prediksjonsvariansen $\mathrm{V}\left(\hat{Y} - Y\right)$ under de ulike modellene. Generelt har vi at

[2] Modellene er vurdert på bakgrunn av dataene fra 1. og 2. kvartal 2004.

$$
\begin{aligned}
\hat{Y} - Y &= \frac{\sum_{i \in s_*^c} y_i}{\sum_{i \in s_*^c} x_i} \cdot \sum_{i \in V \setminus s} x_i - \sum_{i \in U \setminus s} y_i \\
&= \sum_{i \in s_*^c} \frac{\sum_{i \in V \setminus s} x_i}{\sum_{i \in s_*^c} x_i} y_i - \sum_{i \in U \setminus s} y_i \\
&= \sum_{i \in U} c_i y_i \ ,
\end{aligned} \tag{2}
$$

der

$$
c_i = \begin{cases} -1 & , i \notin s \\ 0 & , i \in s_* \\ \dfrac{\sum_{i \in V \setminus s} x_i}{\sum_{i \in s_*^c} x_i} & , i \in s_*^c \ . \end{cases}
$$

Dermed har vi at

$$
\mathrm{V}\left(\hat{Y} - Y\right) = \mathrm{V}\left(\sum_{i \in U} c_i y_i\right).
$$

5.1. SNN17, SNN21, SNN241, SNN27\274, SNN274 og SNN28

For delpopulasjonene SNN17 (tekstilindustri), SNN21 (treforedlingsindustri), SNN241 (kjemiske råvarer), SNN27\274 (metallindustri unntatt ikke-jernholdige metaller), SNN274 (ikke-jernholdige metaller) og SNN28 (metallvareindustri) har vi kommet frem til følgende modell:

$$
y_i \sim \left(\beta z_i, \sigma^2\right) \quad , i \in U_1 = \left\{ i \in U \text{ s.a. } z_i \text{ er kjent} \right\},
$$

$$
y_i \sim \left(\mu_i, \sigma^2\right) \quad , i \in U_2 = \left\{ i \in U \text{ s.a. } z_i \text{ er ukjent} \right\}.
$$

I tillegg antar vi at y_i'ene er uavhengige.

Fordi vi ikke har noe utvalg fra U_2, med unntak av SNN28 hvor to bedrifter er med i utvalget, kan vi ikke tilpasse noen modell for y_i, $i \in U_2$. Men vi må gjøre noen antagelser om variansen $\mathrm{V}\left(y_i\right)$ for å kunne estimere prediksjonsvariansen $\mathrm{V}\left(\hat{Y} - Y\right)$, og vi har valgt å anta at $\mathrm{V}\left(y_i\right) = \sigma^2$.

Vi starter med å se på forventingsskjevhetene til $\hat{Y}$. Fra ligning (2) får vi at

$$
\begin{aligned}
\mathrm{B}\left(\hat{Y}\right) &= \mathrm{E}\left[\hat{Y} - Y\right] \\
&= \mathrm{E}\left[\sum_{i \in U} c_i y_i\right] \\
&= \beta \sum_{i \in U_1} c_i z_i + \sum_{i \in U_2} c_i \mu_i \ .
\end{aligned}
$$

Denne er generelt ikke lik null. Som et eksempel kan vi se på SNN21. For 1. kvartal 2004 har vi at $\beta \cdot \sum_{i \in U_1} c_i z_i \approx -263566 \cdot \beta$. Hvis f.eks. $\mu_i = \beta x_i$ for $i \in U_2$ (i dette tilfellet er x_i kjent for alle bedriftene), får vi at $\sum_{i \in U_2} c_i \mu_i = \beta \sum_{i \in U_2} c_i x_i \approx -23492 \cdot \beta$. Dette gir $\beta \sum_{i \in U_1} c_i z_i + \sum_{i \in U_2} c_i \mu_i \approx -286058 \cdot \beta < 0$, dvs. at $\mathrm{B}\left(\hat{Y}\right)$ ikke er lik null. Hvis vi for det samme eksempelet beregner forventningsskjevheten til $\hat{d}$, får vi at $\mathrm{B}\left(\hat{d}\right) \approx -5.55 \cdot \beta$. Dvs. at $d = (Y / Y95) \cdot 100$ blir underestimert hvis $\mu_i = \beta x_i$ for $i \in U_2$ (andre verdier av μ_i vil gi andre verdier for $\mathrm{B}\left(\hat{d}\right)$).

Fordi y_i'ene er uavhengige får vi at

$$\begin{aligned} \mathrm{V}\left(\hat{Y} - Y\right) &= \mathrm{V}\left(\sum_{i \in U} c_i y_i\right) \\ &= \sigma^2 \sum_{i \in U} c_i^2 \ . \end{aligned}$$

Denne estimerer vi med

$$\hat{\mathrm{V}}\left(\hat{Y} - Y\right) = \hat{\sigma}^2 \sum_{i \in U} c_i^2 \ , \tag{3}$$

der

$$\hat{\sigma}^2 = \frac{1}{n_1 - 1} \sum_{i \in s_1} (y_i - \hat{\beta} z_i)^2 \ ,$$

$$s_1 = s \cap U_1 \ ,$$

$$n_1 = |s_1|$$

og

$$\hat{\beta} = \frac{\sum_{i \in s_1} z_i y_i}{\sum_{i \in s_1} z_i^2} \ .$$

Det kan vises at (3) er en forventningsrett estimator for $\mathrm{V}\left(\hat{Y} - Y\right)$.

Usikkerhetsmålet til $\hat{d}$ blir dermed

$$\hat{\mathrm{st}}\left(\hat{d} - d\right) = \frac{\sqrt{\hat{\mathrm{V}}\left(\hat{Y} - Y\right)}}{Y95} \cdot 100 \ ,$$

der $\hat{\mathrm{V}}\left(\hat{Y} - Y\right)$ er gitt ved (3).

5.2. SNN18, TDM1 og TDM2

For delpopulasjonene SNN18 (bekledningsindustri), TDM1 (transportmiddelindustri) og TDM2 (oljeplattformer og moduler) har vi kommet frem til følgende modell:

$$y_i \sim \left(\beta_k, \sigma_k^2\right) \quad , i \in U_k \quad , k = 1, ..., K-1$$

$$y_i \sim \left(\mu_i, \sigma_i^2\right) \quad , i \in U_K = \left\{i \in U \text{ s.a. } syss_i \text{ er ukjent}\right\}.$$

I tillegg antar vi at y_i'ene er uavhengige.

For SNN18 er

$$K = 4$$

$$U_1 = \left\{i \in U \text{ s.a. } syss_i \geq 50\right\}$$

$$U_2 = \left\{i \in U \text{ s.a. } 25 \leq syss_i < 50\right\}$$

$$U_3 = \left\{i \in U \text{ s.a. } 0 \leq syss_i < 25\right\}.$$

For TDM1 er

$$K = 5$$

$$U_1 = \left\{i \in U \text{ s.a. } syss_i \geq 100\right\}$$

$$U_2 = \left\{i \in U \text{ s.a. } 50 \leq syss_i < 100\right\}$$

$$U_3 = \left\{i \in U \text{ s.a. } 25 \leq syss_i < 50\right\}$$

$$U_4 = \left\{i \in U \text{ s.a. } 0 \leq syss_i < 25\right\}.$$

For TDM2 er

$$K = 4$$

$$U_1 = \left\{i \in U \text{ s.a. } syss_i \geq 200\right\}$$

$$U_2 = \left\{i \in U \text{ s.a. } 100 \leq syss_i < 200\right\}$$

$$U_3 = \left\{i \in U \text{ s.a. } 0 \leq syss_i < 100\right\}.$$

Fordi vi ikke har noe utvalg fra U_K, med unntak av TDM1 hvor en bedrift er med i utvalget, kan vi ikke tilpasse noen modell for y_i, $i \in U_K$. Men vi skal gjøre to antagelser. Alle bedrifter har en sysselsetting (selv om den er ukjent for bedriftene i U_K). Vi skal anta at sysselsettingen fordeler seg

likt i U_K som i resten av U. Denne antagelsen medfører at det fins en partisjon W_k, $k=1,...,K-1$, av U_K, slik at

$$\frac{|W_k|}{|U_K|} \approx \frac{|U_k|}{|U \setminus U_K|}, \quad k=1,...,K-1.$$

Den andre antagelsen er at ordretilgangen i W_k kan modelleres med samme modell som ordretilgangen i U_k, dvs. at $\mu_i = \beta_k$ og $\sigma_i^2 = \sigma_k^2$ for $i \in W_k$.

Vi starter med å se på forventningsskjevheten til $\hat{Y}$. Fra ligning (2) får vi at denne kan skrives som

$$\begin{aligned} \mathrm{B}(\hat{Y}) &= \mathrm{E}\left[\hat{Y} - Y\right] \\ &= \mathrm{E}\left[\sum_{i \in U} c_i y_i\right] \\ &= \sum_{k=1}^{K-1} \beta_k \sum_{i \in U_k} c_i + \sum_{i \in U_K} c_i \mu_i \\ &= \sum_{k=1}^{K-1} \beta_k \left(\sum_{i \in U_k} c_i + \sum_{i \in W_k} c_i \right) \end{aligned}$$

Ved å sette inn ulike verdier for β_k, $k=1,....,K-1$, og velge forskjellige partisjoner av U_K, kan det vises at $\mathrm{B}(\hat{Y})$ generelt ikke er lik 0. Dvs. at $\hat{Y}$ ikke er forventningsrett, og dermed er heller ikke $\hat{d}$ forventningsrett.

For prediksjonsvariansen $\mathrm{V}(\hat{Y} - Y)$ får vi

$$\begin{aligned} \mathrm{V}(\hat{Y} - Y) &= \mathrm{V}\left(\sum_{i \in U} c_i y_i\right) \\ &= \sum_{k=1}^{K-1} \sigma_k^2 \sum_{i \in U_k} c_i^2 + \sum_{i \in U_K} c_i^2 \sigma_i^2 \\ &= \sum_{k=1}^{K-1} \sigma_k^2 \left(\sum_{i \in U_k} c_i^2 + \sum_{i \in W_k} c_i^2 \right). \end{aligned}$$

Denne estimerer vi med

$$\hat{\mathrm{V}}(\hat{Y} - Y) = \sum_{k=1}^{K-1} \hat{\sigma}_k^2 \left(\sum_{i \in U_k} c_i^2 + \sum_{i \in \widetilde{W}_k} c_i^2 \right), \tag{4}$$

der $\widetilde{W}_k$, $k=1,...,K-1$, er en partisjon av U_K slik at

$$\frac{|\widetilde{W}_k|}{|U_K|} \approx \frac{|U_k|}{|U \setminus U_K|},$$

og

$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1}\sum_{i \in s_k}(y_i - \bar{y}_{s_k})^2 ,$$

$s_k = s \cap U_k$, $n_k = |s_k|$ og $\bar{y}_{s_k} = \frac{1}{n_k}\sum_{i \in s_k} y_i$.

Usikkerhetsmålet til $\hat{d}$ blir dermed

$$\hat{\text{st}}\left(\hat{d} - d\right) = \frac{\sqrt{\hat{\text{V}}\left(\hat{Y} - Y\right)}}{Y95} \cdot 100 ,$$

der $\hat{\text{V}}\left(\hat{Y} - Y\right)$ er gitt ved (4).

Delmengdene $\widetilde{W}_k$, $k = 1,...,K-1$, kan velges på mange måter. For SNN18 og TDM2, hvor det ikke er noe utvalg fra U_K, er $c_i = -1$ for alle $i \in U_K$ slik at $\sum_{i \in \widetilde{W}_k} c_i^2 = \sum_{i \in \widetilde{W}_k} 1 = |\widetilde{W}_k| \approx \frac{|U_k|}{|U \setminus U_K|} \cdot |U_K|$. Dermed får vi tilnærmet det samme estimatet uansett hvordan delmengdene velges. For TDM1 er $c_i = 2.38$ for den bedriften i U_K som er i utvalget, og -1 for resten. For denne delpopulasjonen velger vi $\widetilde{W}_k$ 'ene slik at bedriften med $c_i = 2.38$ er med i den delmengden som har størst $\hat{\sigma}_k^2$.

5.3. SNN24\241, SNN29 og SNN30_33

For delpopulasjonene SNN24\241 (kjemikalier og kjemiske produkter unntatt kjemiske råvarer), SNN29 (maskinindustri) og SNN30_33 (elektronisk og optisk industri) har vi kommet frem til følgende modell:

$$y_i \sim \left(\beta_1, \sigma_1^2\right) \quad , i \in U_1$$

$$y_i \sim \left(\beta_2 z_i, \sigma_2^2\right) \quad , i \in U_2$$

$$y_i \sim \left(\mu_i, \sigma_i^2\right) \quad , i \in U_3 = \{i \in U \text{ s.a. } syss_i \text{ er ukjent}\}$$

I tillegg antar vi at y_i 'ene er uavhengige.

For SNN24\241 og SNN30_33 er

$$U_1 = \{i \in U \text{ s.a. } syss_i \geq 200\} \quad \text{og} \quad U_2 = \{i \in U \text{ s.a. } 0 \leq syss_i < 200\} ,$$

og for SNN29 er

$$U_1 = \{i \in U \text{ s.a. } syss_i \geq 100\} \quad \text{og} \quad U_2 = \{i \in U \text{ s.a. } 0 \leq syss_i < 100\}.^3$$

Fordi vi ikke har noe utvalg fra U_3, med unntak av SNN29 hvor en bedrift er med i utvalget, kan vi ikke tilpasse noen modell for y_i, $i \in U_3$. Men vi skal gjøre tilsvarende antagelser som vi gjorde i avsnitt 5.1 og 5.2. Dvs. vi antar at sysselsettingen fordeler seg likt i U_3 som i $U_1 \cup U_2$. Denne antagelsen medfører at U_3 kan splittes opp i W_1 og W_2 på en slik måte at

$$\frac{|W_1|}{|U_3|} \approx \frac{|U_1|}{|U_1 \cup U_2|} \text{ og } \frac{|W_2|}{|U_3|} \approx \frac{|U_2|}{|U_1 \cup U_2|}.$$

Videre antar vi at ordretilgangen i W_1 kan modelleres med samme modell som ordretilgangen i U_1 (dvs. at $\mu_i = \beta_1$ og $\sigma_i^2 = \sigma_1^2$ for $i \in W_1$), og at $\sigma_i^2 = \sigma_2^2$ for $i \in W_2$.

Vi starter med å se på forventingsskjevhetene til $\hat{Y}$. Fra ligning (2) får vi at

$$\begin{aligned}
\mathrm{B}(\hat{Y}) &= \mathrm{E}\left[\hat{Y} - Y\right] \\
&= \mathrm{E}\left[\sum_{i \in U} c_i y_i\right] \\
&= \beta_1 \sum_{i \in U_1} c_i + \beta_2 \sum_{i \in U_2} c_i z_i + \sum_{i \in U_3} c_i \mu_i \\
&= \beta_1 \left(\sum_{i \in U_1} c_i + \sum_{i \in W_1} c_i\right) + \beta_2 \sum_{i \in U_2} c_i z_i + \sum_{i \in W_2} c_i \mu_i \ .
\end{aligned}$$

Ved å sette inne ulike verdier for β_1, β_2 og μ_i'ene, og velge forskjellige partisjoner av U_3, kan det vises at $\mathrm{B}(\hat{Y})$ generelt ikke er lik 0. Dvs. at $\hat{Y}$ ikke er forventningsrett, og dermed er heller ikke $\hat{d}$ forventningsrett.

For prediksjonsvariansen $\mathrm{V}(\hat{Y} - Y)$ får vi

$$\begin{aligned}
\mathrm{V}(\hat{Y} - Y) &= \mathrm{V}\left(\sum_{i \in U} c_i y_i\right) \\
&= \sigma_1^2 \sum_{i \in U_1} c_i^2 + \sigma_2^2 \sum_{i \in U_2} c_i^2 + \sum_{i \in U_3} c_i^2 \sigma_i^2 \\
&= \sigma_1^2 \left(\sum_{i \in U_1} c_i^2 + \sum_{i \in W_1} c_i^2\right) + \sigma_2^2 \left(\sum_{i \in U_2} c_i^2 + \sum_{i \in W_2} c_i^2\right).
\end{aligned}$$

Vi estimerer σ_1^2 og σ_2^2 med henholdsvis

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - 1} \sum_{i \in s_1} (y_i - \bar{y}_{s_1})^2$$

[3] Når sysselsettingen er kjent er også z_i kjent. Dermed er z_i kjent for $i \in U_2$.

og

$$\hat{\sigma}_2^2 = \frac{1}{n_2 - 1}\sum_{i \in s_2}(y_i - \hat{\beta}_2 z_i)^2 ,$$

der

$$\bar{y}_{s_1} = \frac{1}{n_1}\sum_{i \in s_1} y_i , \qquad \hat{\beta}_2 = \frac{\sum_{i \in s_2} z_i y_i}{\sum_{i \in s_2} z_i^2},$$

$s_k = s \cap U_k$ og $n_k = |s_k|$, $k = 1,2$.

Prediksjonsvariansen estimeres nå med

$$\text{(5)} \qquad \hat{\mathrm{V}}\left(\hat{Y} - Y\right) = \hat{\sigma}_1^2\left(\sum_{i \in U_1} c_i^2 + \sum_{i \in \widetilde{W}_1} c_i^2\right) + \hat{\sigma}_2^2\left(\sum_{i \in U_2} c_i^2 + \sum_{i \in \widetilde{W}_2} c_i^2\right) ,$$

der $\widetilde{W}_1$ og $\widetilde{W}_2$ er en partisjon av U_3 slik at

$$\frac{|\widetilde{W}_1|}{|U_3|} \approx \frac{|U_1|}{|U_1 \cup U_2|} \text{ og } \frac{|\widetilde{W}_2|}{|U_3|} \approx \frac{|U_2|}{|U_1 \cup U_2|} .$$

Usikkerhetsmålet til $\hat{d}$ blir dermed

$$\hat{\mathrm{st}}\left(\hat{d} - d\right) = \frac{\sqrt{\hat{\mathrm{V}}\left(\hat{Y} - Y\right)}}{Y95} \cdot 100 ,$$

der $\hat{\mathrm{V}}\left(\hat{Y} - Y\right)$ er gitt ved (5).

6. Talleksempel

Vi skal nå estimere usikkerheten til verdiindeksen $\hat{d}$, for 1. og 2. kvartal 2004, med

$$\hat{\mathrm{st}}\left(\hat{d} - d\right) = \frac{\sqrt{\hat{\mathrm{V}}\left(\hat{Y} - Y\right)}}{Y95} \cdot 100 ,$$

der $\hat{\mathrm{V}}\left(\hat{Y} - Y\right)$ er gitt ved (3) for SNN17, SNN21, SNN241, SNN27\274, SNN274 og SNN28, (4) for SNN18, TDM1 og TDM2, og (5) for SNN24\241, SNN29 og SNN30_33.

I 2004 var det ca. 6800 bedrifter i populasjonen, og utvalget for 1. og 2. kvartal var på henholdsvis 806 og 796 bedrifter. I Tabell 1 ser vi hvordan populasjonen og utvalget fordelte seg mellom delpopulasjonene i 1. kvartal.

Tabell 1

Delpopulasjon	Ant. bedrifter i populasjonen	Ant. bedrifter i bruttoutvalget	Frafall i %
SNN17	426	58	1.72
SNN18	218	20	5.00
SNN21	122	45	0
SNN24\241	205	34	0
SNN241	139	40	5.00
SNN27\274	143	32	3.13
SNN274	47	18	5.56
SNN28	1613	155	0
SNN29	1678	123	7.32
SNN30_33	1167	131	5.34
TDM2	154	44	4.55
TDM1	916	106	0.94

I ordrestatistikken blir verdiindeksen beregnet ikke bare for hver av de 12 delpopulasjonene, men også for følgende grupper:

Hele populasjonen
Tekstil- og bekledingsindustri (tilsvarer SNN17 og SNN18)
Kjemikalier og kjemiske produkter (tilsvarer SNN24\241 og SNN241)
Metallindustri (tilsvarer SNN27\274 og SNN274)

Vi estimerer derfor usikkerheten for disse gruppene også. (Som notasjon for de tre siste gruppene bruker vi henholdsvis SNN17_18, SNN24 og SNN27).

Estimatene vi har fått er vist i Tabell 2 og Tabell 3. (Vi har ikke tatt med estimatene for SNN24\241 og SNN27\274, fordi disse ikke blir presentert i ordrestatistikken). Verdiindeksen er her beregnet slik den beregnes i ordrestatistikken, dvs. når ekstreme observasjoner blir skilt ut ved estimeringen. Den siste kolonnen i tabellene viser variasjonskoeffisienten (coefficient of variation), som er gitt ved

$$\widehat{c.v.}\left(\hat{d}\right) = \hat{\mathrm{st}}\left(\hat{d} - d\right) \Big/ \hat{d} \, .$$

Dvs. at variasjonskoeffisienten måler hvor stor den estimerte usikkerheten er i forhold til verdiindeksen.

Tabell 2: Verdiindeks og estimert usikkerhet for 1. kvartal 2004

Publiseringsnivå	Verdiindeks	Estimert usikkerhet	Variasjons-koeffisienten (%)
Hele pop.	141.95	9.66	6.81
SNN17_18	117.62	17.59	14.96
SNN17	115.95	17.88	15.42
SNN18	122.00	43.25	35.45
SNN21	71.74	5.56	7.75
SNN24	229.13	34.49	15.05
SNN241	249.45	51.02	20.45
SNN27	134.99	17.05	12.63
SNN274	139.98	22.79	16.28
SNN28	191.98	92.90	48.39
SNN29	191.08	23.97	12.54
SNN30_33	97.04	12.36	12.74
TDM2	124.40	29.09	23.39
TDM1	130.13	19.08	14.66

Tabell 3: Verdiindeks og estimert usikkerhet for 2. kvartal 2004

Publiseringsnivå	Verdiindeks	Estimert usikkerhet	Variasjons-koeffisienten (%)
Hele pop.	140.43	13.49	9.61
SNN17_18	94.17	13.46	14.30
SNN17	110.80	18.19	16.41
SNN18	50.29	9.87	19.62
SNN21	85.33	12.55	14.71
SNN24	180.04	26.38	14.65
SNN241	159.13	33.30	20.93
SNN27	141.45	16.29	11.52
SNN274	145.94	17.57	12.04
SNN28	236.75	176.90	74.72
SNN29	190.47	22.70	11.92
SNN30_33	107.07	11.23	10.49
TDM2	125.46	24.53	19.55
TDM1	107.48	11.05	10.28

Hvis vi sammenligner verdiindeksen for de to kvartalene, ser vi at den endrer seg ganske mye for noen av gruppene. Størst endring har vi for SNN241, der $\hat{d}$ er 249.45 i 1. kvartal mot 159.13 i 2. kvartal.

De fleste publiseringsnivåene har fått en variasjonskoeffisient som ligger mellom 10% og 20%. Dvs. at den estimerte usikkerheten utgjør mellom 10 og 20 prosent av verdiindeksen, og det er ganske stor usikkerhet (men akseptabelt). Det er bare for hele populasjonen at variasjonskoeffisienten er blitt mindre enn 10% for begge kvartalene. For 1. kvartal er den 6.81% og for 2. kvartal er den 9.61%. Dette tyder på at det er mindre usikkerhet i verdiindeksen som gjelder hele populasjonen, enn i verdiindeksen for de andre gruppene.

Utenom SNN18 og SNN28, som vi skal omtale senere, er det bare SNN241 og TDM2 som har fått en variasjonskoeffisient større enn 20%. For SNN241 er $\widehat{c.v.}\left(\hat{d}\right) = 20.45\%$ og 20.93%, for henholdsvis 1. og 2. kvartal, og for TDM2 er $\widehat{c.v.}\left(\hat{d}\right) = 23.39\%$ for 1. kvartal (for 2. kvartal er $\widehat{c.v.}\left(\hat{d}\right) = 19.55\%$).

SNN18 og SNN28 skiller seg ut fra de andre ved at variasjonskoeffisienten er blitt ekstremt stor. Verst er SNN28 med $\widehat{c.v.}\left(\hat{d}\right) = 48.39\%$ og 74.72% for henholdsvis 1. og 2. kvartal. For SNN18 er variasjonskoeffisienten blitt 35.45% i 1. kvartal og 19.62% i 2. kvartal. Grunnen til de store verdiene er at vi ikke har klart å komme frem til en bra nok modell for bedriftenes ordretilgang. Modellen vi bruker passer bra for alle bedriftene i utvalget, med unntak av et par bedrifter som skiller seg ut som ekstreme (se Vedlegg for illustrerende figurer for SNN28). Selv om det bare er snakk om en eller to bedrifter som modellen ikke passer for, forårsaker disse at $\widehat{\text{st}}\left(\hat{d} - d\right)$, og dermed også $\widehat{c.v.}\left(\hat{d}\right)$, blir veldig stor.

Grunnen til at variasjonskoeffisienten ikke er blitt ekstremt stor for SNN18 i 2. kvartal, er at det for dette kvartalet ikke er noen bedrifter i utvalget som skiller seg ut som ekstreme i forhold til den modellen vi bruker. (Den bedriften som skilte seg ut i 1. kvartal, er ikke med i populasjonen i 2. kvartal).

Hvis det ikke er målefeil som er årsaken til at noen få av bedriftene i SNN18 og SNN28 skiller seg fra resten, betyr det at vi bruker feil modell for en del av bedriftene. Dette betyr igjen at usikkerhetsmålet vi har lagd for SNN18 og SNN28, ikke er så bra. Hadde vi klart å skille ut bedriftene som modellen ikke passer for, og tilpasset en egen modell for disse, ville vi fått et mer riktig usikkerhetsmål (og da ville vi antakelig unngått de ekstremt store variasjonskoeffisientene).

7. Oppsummering

I dette notatet har vi lagd et usikkerhetsmål for verdiindeksen til ordretilgangen i den kvartalsvise ordrestatistikken. Usikkerhetsmålet er modellbasert, dvs. at vi ser på bedriftenes ordretilgang som stokastiske variable, mens utvalget antas gitt.

Vi har brukt dette usikkerhetsmålet til å estimere usikkerheten til verdiindeksen for 1. og 2. kvartal 2004. For de aller fleste publiseringsnivåene fikk vi en variasjonskoeffisient som lå mellom 10% og 20%, hvilket betyr at den estimerte usikkerheten utgjør mellom 10 og 20 prosent av verdiindeksen. Dette er en stor usikkerhet, men den er akseptabel. Det ser ut til å være noe mindre usikkerhet for verdiindeksen som gjelder hele populasjonen. Variasjonskoeffisienten ble for hele populasjonen 6.81% i 1. kvartal og 9.61% i 2. kvartal.

For publiseringsnivåene SNN18 og SNN28 fikk vi ekstremt store variasjonskoeffisienter. Årsaken til dette er at det er noen få bedrifter i utvalget som ikke passer til den modellen vi har lagt til grunn for usikkerhetsmålet. Det betyr at usikkerhetsmålet for disse to publiseringsnivåene ikke er så bra.

8. Vedlegg

For SNN28 (metallvareindustri) har vi valgt følgende modell for $i \in U_1 = \{i \in U \text{ s.a. } z_i \text{ er kjent}\}$:

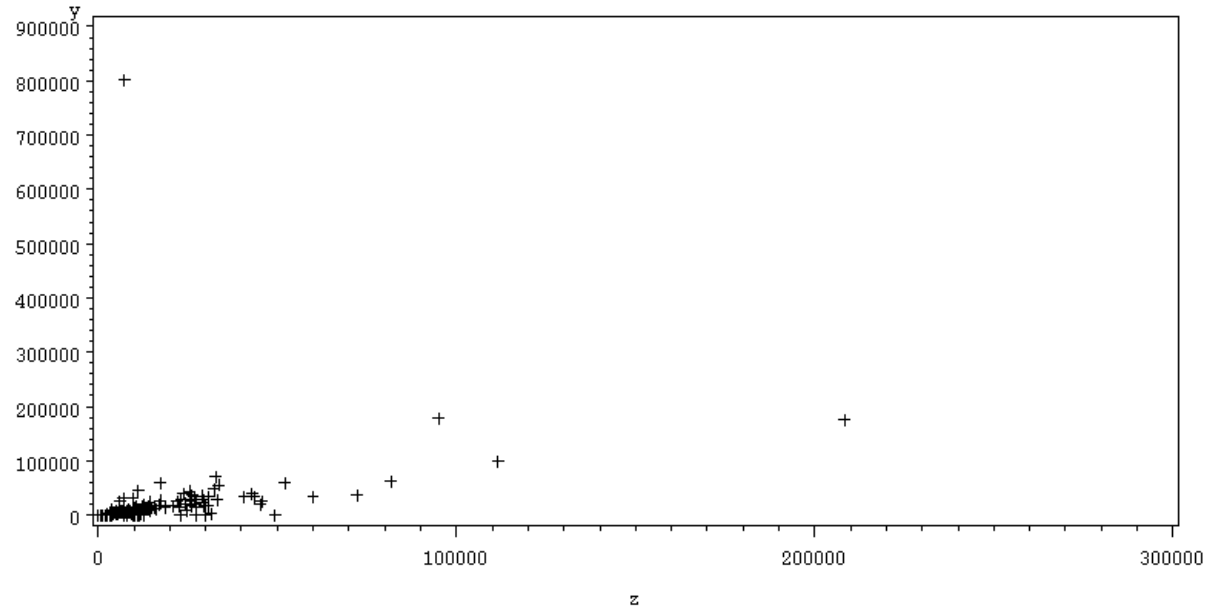
$$y_i \sim (\beta z_i, \sigma^2).$$

Denne modellen ser ut til å passe noenlunde bra for alle bedriftene i utvalget, med unntak av følgende to bedrifter:

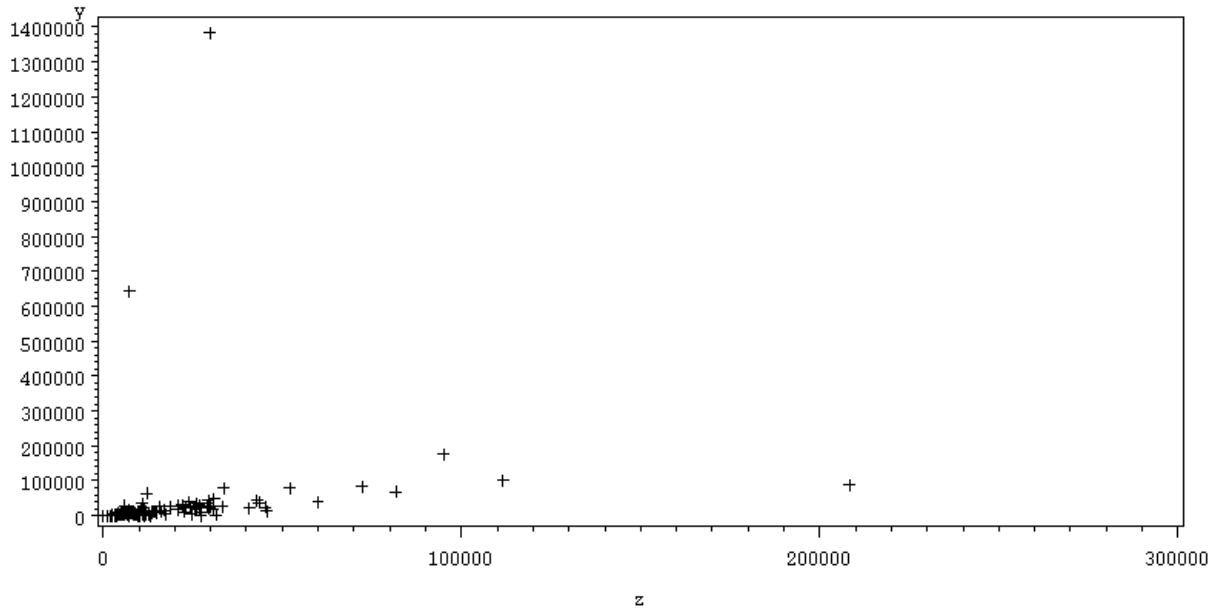
Bedrift	sysselsetting	z_i	y_i (1. kvartal)	y_i (2. kvartal)
A	53	7328.5	801926	643270
B	13	29891.5	0	1383388

(Bedrift B skiller seg ikke så tydelig ut i 1. kvartal).

Plott av y_i mot z_i (1. kvartal)



Plott av y_i mot z_i (2. kvartal)



Vi har brukt SAS/INSIGHT til å estimere β og beregne R^2, både med og uten bedrift A og B. (R^2 er et mål på hvor stor andel av den totale variasjonen i dataene som forklares av regresjonsmodellen).

For 1. kvartal fikk vi følgende tall:

	$\hat{\beta}$	R^2
Med bedrift A og B	0.95	0.149
Uten bedrift A og B	0.91	0.804

For 2. kvartal fikk vi tallene:

	$\hat{\beta}$	R^2
Med bedrift A og B	1.18	0.070
Uten bedrift A og B	0.81	0.713

9. Referanser

Bakken, P. og Osnes, J.A. (1998): *Kvartalsvis ordrestatistikk*, Notater 98/36, Statistisk sentralbyrå.

Statistisk sentralbyrå (1997): Dokumentasjon av utvalgsplan.

De sist utgitte publikasjonene i serien Notater

2004/74 M. Åamodt: Kvalitetsprosjekt for videregående opplæring Utført på oppdrag fra Utdannings- og forskningsdepartementet i perioden mars 2003-september 2004. 188s.

2004/75 S. Blom: Holdninger til innvandrere og innvandring 2004. 54s.

2004/76 A. Rolland: En inspeksjon av Elevinspektørene. 51s.

2004/77 A. Rolland: KOSTRA og kvaliteten på de kommunale tjenester. 32.

2004/78 J.A. Osnes: Beregningsutvalget. Dokumentasjon av SAS-systemet. 98s.

2004/79 T. Eika og T. Skjerpen. Hvitevarer 2005. Modell og prognose. 18s.

2004/80 A.K. Johnsen og T. Nøtnes: Biblioteket i forkus? Rapport fra fokusgrupper for bibliotek og informassjonssenteret i Statistisk sentralbyrå. 26s.

2004/81 H. Tønseth: Årsrapport 2003. Kontaktutvalget for helse- sosialstatistikk. 12s.

2004/82 I. Håland og G. Næringsrud: Kontantstøtte og Arbeidskraftundersøkelsen (AKU). 28s.

2004/83 L. Vågane: Omnibusundersøkelsen juli /august 2004. Dokumentasjonsrapport. 45s.

2004/84 D. Spilde: Statistikk over energibruk i industrien. Dokumentasjon og brukerveiledning. 53s.

2004/85 L. Haakonsen: KVARTS i paksis III. Systemer og rutiner i den daglige driften. 72s.

2004/86 L-C. Zhang og A. Vedø: Omlegging av utvalgsplan for (AKU). 15s.

2004/87 F. Strøm: Personer uten registrert inntekt eller formue. En gjennomgang av SSBs datagrunnlag for registerbasert inntekts- og formustatistikk 30s.

2004/88 G. Daugstad og B. Lie: Kvalitativ forstudie til levekårsundersøkelse blant ikkevestlige innvandrere. 138s.

2004/89 S. Lien og Ø. Sivertstøl: Langtidsmottakere av sosialhjelp 1997-1999. 64s. ISSN 0806-3745

2005/1 S. Hansen og T. Skoglund: Sysslsetting og lønn i historisk nasjonalregnskap. Beregninger for 1949-1969. 36s.

2004/2 FoU og innovasjonstatistikk 2001 og 2002-dokumentasjon. 82s.

2005/3 M. Steinnes, J. Monsrud, E. Engelien og V.V. Holst Bloch: Samferdsel og miljø. Utvikling av et norsk indikatorsett tilpasset et felles europeisk sammenligningsgrunnlag. 80s.

2005/4 E. Falnes-Dalheim og A. Falnes-Dalheim: Dokumentasjon av FoB2001. Spesifikasjoner, bearbeiding, flytdiagram for spørreskjemadelen av tellingen. Del I. 117s.

2005/5 E. Falnes-Dalheim, A. Falnes-Dalheim: J. Sjørbotten og B. Østvedt: Dokumentasjon av FoB2001. Spesifikasjoner, bearbeiding, flytdiagram for spørreskjemadelen av tellingen. Del II Vedlegg. 146s.

2005/6 E. Falnes-Dalheim: Bearbeiding av prøvetellingen i Stange 2000. Folke- og boligtellingen 2001. 126s.

2005/7 S. Kwesi Baateng og S. Ferstad: Dokumentasjonsnotat for FylkesKOSTRA vidregående opplæring. Publisering av 2003-tallene. 221s.

2005/8 Ø. Linnestad og O.K. Lien: SM08 Prisindekser. Fraktindeks på utenriks sjøfart. 56s.

2005/9 E. Cometa Rauan og R. Johannessen: Forventningsindikator - konsumprisene. November 2004 - mai 2005. 18s.

2005/10 A.S. Abrahamsen: Analyse av revisjon - Feilkoder og endringer i utenrikshandelstatistikken. 71s.