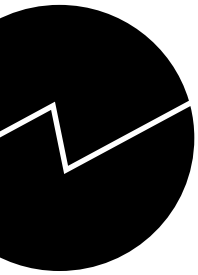
Statistics Norway



Statistisk sentralbyrå

Dina Rafat

Produksjonsopplegg for foreløpige tall i industristatistikken

Seksjon for Metoder og standarder

Innhold

1. Innledning

Industristatistikken er en skjemabasert undersøkelse av bedrifter innen industrinæringene. Tradisjonelt henter den informasjon for alle store bedrifter, mens verdiene for enheter utenfor utvalget predikeres ved hjelp av registervariable fra Bedrifts- og Foretaksregisteret (BoF). De endelige tallene publiseres ca.1,5 år etter referanseåret og inneholder ca. 90 poster. EUROSTAT krever rapportering av foreløpige tall 10 måneder etter referanseår på 5 hovedposter. Et omfatende revisjonsarbeid gjør det vanskelig å holde de fristene, men det satses på fremstilling av foreløpige tall på slutten av året (dvs. 12 måneder etter referanse år).

I dette notatet presenteres det et opplegg for produksjon av foreløpige tall. Modell og estimeringsopplegg beskrives, samt et program utviklet for beregning av foreløpige tall på mikronivå. Industristatistikken for år 2002 gis som illustrasjon av størrelsesforholdene og beregningene.

Dessuten viser vi hvordan S-KJR applikasjonen for estimering av variasjonskoeffisienter og konfidensintervall kan anvendes på industristatistikken. Gjennom dette opplegget satser man på å kunne drive bedre kvalitetsarbeid i statistikkproduksjonen og oppfylle kravene fra EUROSTAT med hensyn til rapportering av kvalitetsmål for tallene.

Kapitel 2 gir en kort oversikt over populasjonen og utvalget, kapitel 3 konsentrerer seg om prinsipper for produksjon av foreløpige tall. Modellen presenteres i detalj i kapitel 4. Kapitel 5 forklarer programmet og et opplegg for hvordan brukeren kan anvende programmet på en enklest mulig måte. I kapitel 6 sammenligner vi foreløpige tall for år 2002 med de endelige tallene for samme år. Kapitel 7 beskriver hvordan S-KJR applikasjonen kan anvendes.

2. Populasjon og utvalg

Populasjonen består av alle bedrifter innen næringene 10-37 med unntak av næring 11. I år 2002 besto den foreløpige populasjonen av 22 873 enheter, hvorav halvparten er svært små (hovedsakelig enkeltmannsforetak). Det er stor variasjon i populasjonen, hvor størrelse kan variere fra 0 til 1700 sysselsatte.

Utvalget består av en totaltelling av store bedrifter og et stratifisert utvalg av mindre bedrifter. Totaltellingen har med alle bedrifter med minst 20 sysselsatte i flerbedriftsforetak og minst 30 sysselsatte i enbedriftsforetak. Alle bedrifter i foretak som er trukket ut tas med. I tillegg kommer et utvalg av mindre bedrifter stratifisert etter næring og trukket enkelt tilfeldig innen strata. Antall bedrifter trukket i hvert strata er avhengig av tallet på sysselsetting og antall bedrifter i den næringen. Det er en cut-off grense, hvor bedrifter med under 10 sysselsatte ikke trekkes. I år 2002 utgjorde utvalget 4 236 bedrifter.

I tillegg til skjema er det tre andre kilder vi kan bruke:
- Næringsoppgave (NO)
- Regnskapsregister for aksjeselskap (BKF).
- MVA-tall

Dette er regnskapsvariable som gjelder foretak og derfor brukes de bare for bedrifter i enbedriftsforetak.

Siden utvalget består av de største bedriftene i populasjonen dekker bedriftene i utvalget en betydelig andel av populasjonenes totale omsetning. Tabell 1 viser at utvalget dekker 85% av den totale omsetningen. I tillegg er det et stort antall bedrifter med NO og BKF-tall som dekker til sammen ca. 10% av omsetningen.

Tabell 1. Antall bedrifter og omsetning fordelt på forskjellige type bedrifter.

Type	Antall	Omsetning	Omsetning, %
Hjelpebedrift	559	13 265 049	2,89
Utvalg	3 709	393 837 181	85,68
NO-tall	5 842	20 496 650	4,46
BKF-tall	3 547	20 676 433	4,50
Resten	9 216	11 371 202	2,47
I alt	22 873	459 646 500	100,00

Tabell A.1 i vedlegg A viser, for hver næring på 3-siffer nivå, antall bedrifter og prosent av total omsetning (dekning). Tabell A.2 presenterer prosent av omsetningen som dekkes av (i) utvalget, (ii) bedrifter med NO tall, (iii) bedrifter med BKF tall, og hvor mange prosent resterende enheter utgjør som vi må predikere tall for. Her ser vi at næringene som utgjør en stor del av total omsetning dekkes ganske bra av utvalget, mens næringene hvor resterende foretak dekker en stor andel er stort sett ubetydelige i den totale sammenhengen.

3. Prinsipper for produksjon av foreløpige tall

Revisjon av utvalget for foreløpige tall settes til ca. 1. desember. Følgende foretak prioriteres ved revisjon:

1. Alle foretak med minst 200 sysselsatte eller over 1 milliard i produksjonsverdi året før revideres ferdig;
2. Alle nye foretak med mange ansatte (hovedsakelig mer enn 200) og eventuelle viktige foretak revideres ferdig;
3. Foretak som har de største feilene på feillister.

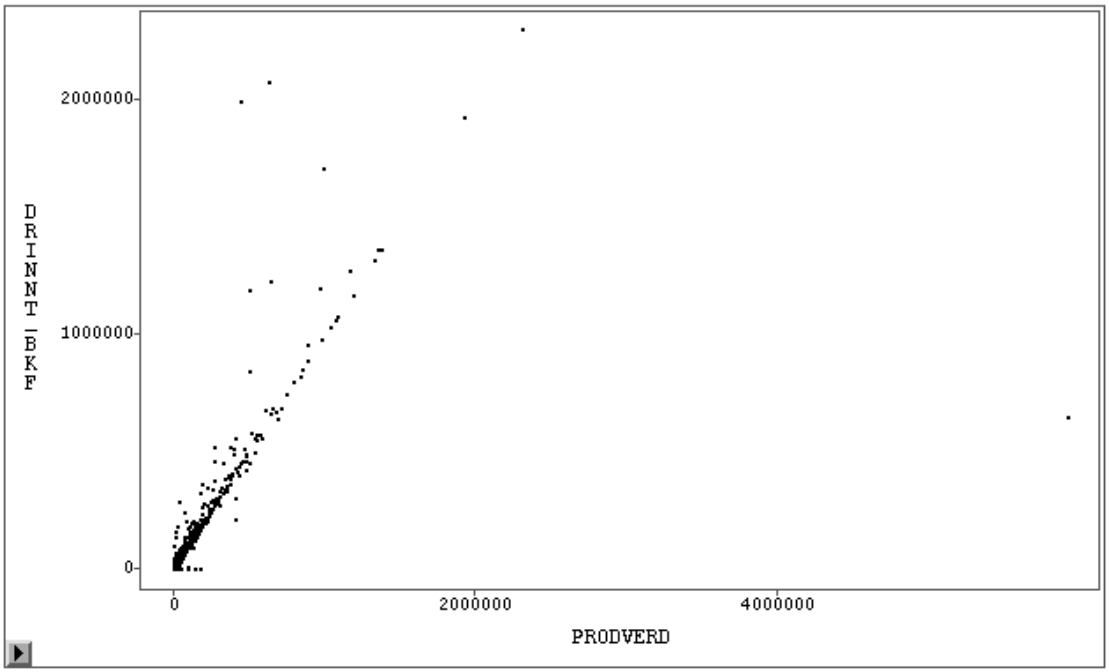
Det beregnes foreløpige tall for følgende variabler: produksjonsverdi, bearbeidingsverdi, produktinnsats og total lønn. Bearbeidingsverdi predikeres ikke direkte, men beregnes som differansen mellom produksjonsverdi og produktinnsats.

Vi utvider estimeringsgrunnlaget med enbedriftsforetakene som har tall fra NO og BKF. NO-tall tilsvarer i stor grad hovedpostene i industristatistikken og brukes direkte som predikerte verdier mens BKF-regnskapstall brukes som stedfortreder på følgende måte:

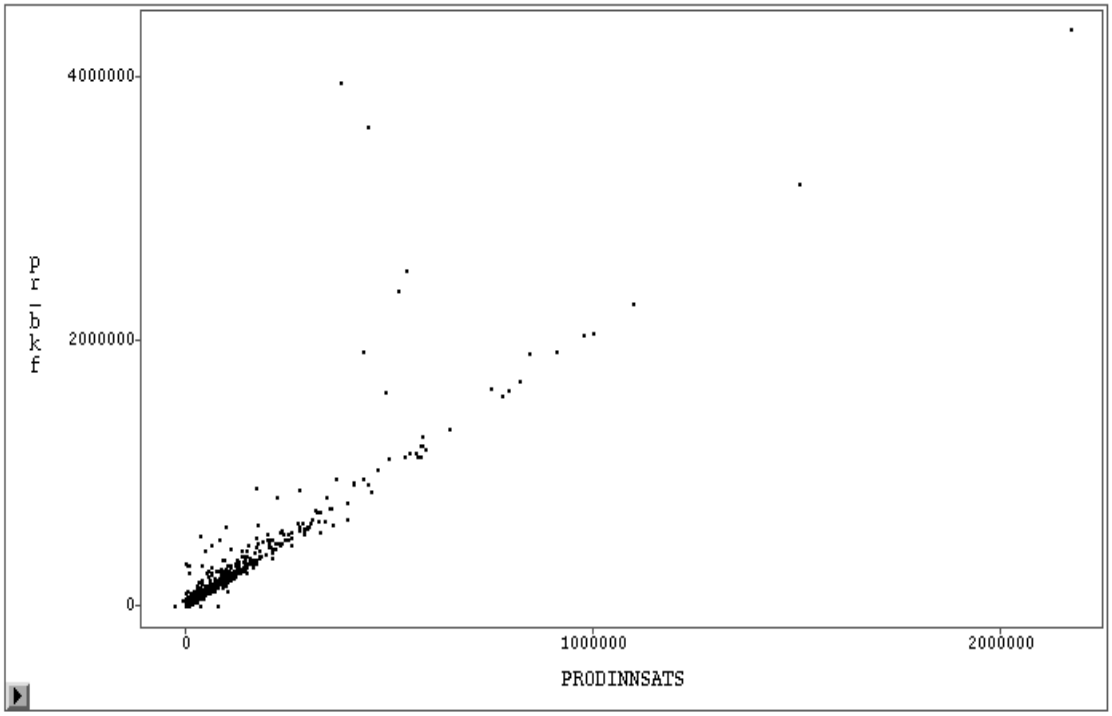
- Sum driftsinntekter fra BKF brukes som erstatning for produksjonsverdi;
- Sum varekostnader og andre driftskostnader fra BKF brukes som erstatning for produksjonsinnsats;
- Lønninger, hentet fra BKF-tall brukes som erstatning for totale lønnskostnader.

Figur 1, 2 og 3 viser samsvar mellom BKF-tall og verdier for hovedposter for enbedriftsforetakene fra utvalget. Her ser vi at det er bra overensstemmelse mellom produksjonsverdi og driftsinntekter fra BKF (bortsett fra få enkelte tilfeller) og mellom totall lønn og lønninger hentet fra BKF, mens sum av varekostnader og andre driftskostnader fra BKF gir høyere verdier enn produktinnsats fra skjema.

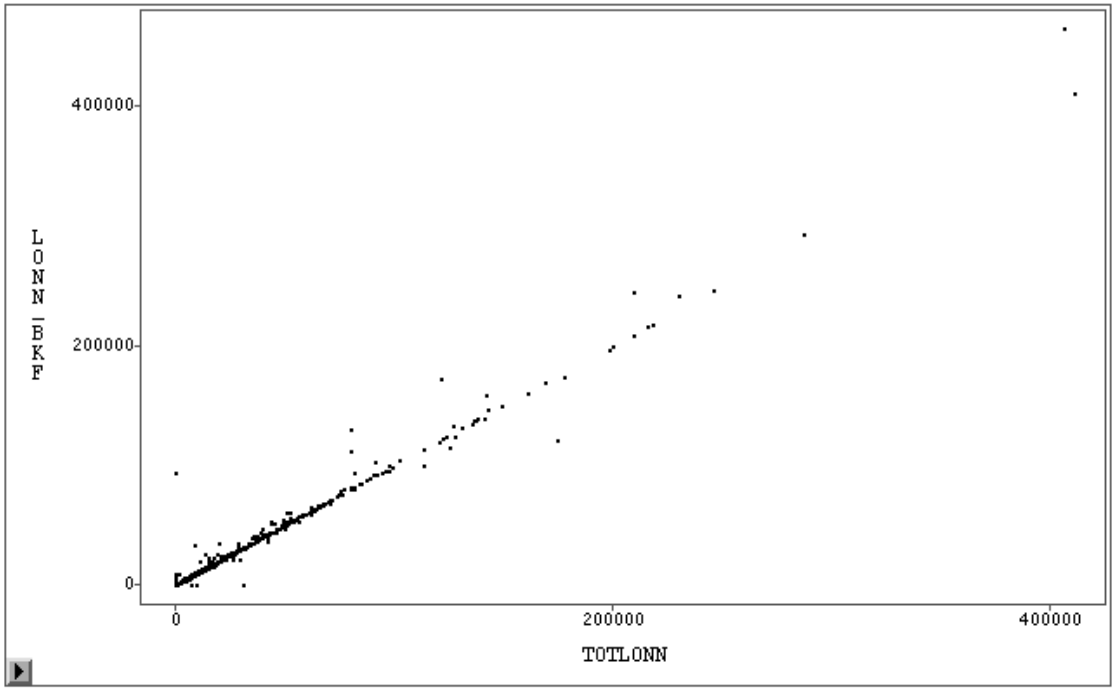
Figur 1. Samsvar mellom produksjonsverdi og driftsinntekter fra BKF for utvalgsbedrifter



Figur 2. Samsvar mellom produktinnsats og sum varekostnader og andre driftskostnader fra BKF for utvalgsbedrifter



Figur 3. Samsvar mellom totall lønn og lønn fra BKF for utvalgsbedrifter



I prioritert rekkefølge benyttes følgende tall for å beregne de enkelte variable:

- Produksjonsverdi, bearbeidingsverdi og produktinnsats:

1. Ferdig reviderte tall;
2. Tall laget på grunnlag av NO;
3. Tall laget på grunnlag av BKF-tall;
4. Tall laget på grunnlag av omsetning.

- Lønn:

1. Ferdig reviderte tall;
2. Tall laget på grunnlag av NO;
3. Tall laget på grunnlag av BKF-tall;
4. Tall laget på grunnlag av omsetning og sysselsetting.

Foreløpige tall brukes av Nasjonalregnskapet (NR) på 3-siffer næringsnivå. EUROSTAT bruker 2-siffer næring på foreløpige tall. Likevel var det et ønske fra fagseksjonen at tall, der det er mulig, lages på mikronivå for å kunne sjekke kvaliteten på andre nivå (for eksempel størrelsesgrupper). Vi gjør oppmerksom på at selv om dette opplegget beregner tall på mikronivå er disse verdiene ikke de faktiske tallene i bedriftene, men beregnete tall for en modell og derfor beheftet med usikkerhet. Slike beregnete tall for en enkelt bedrift må derfor brukes med varsomhet.

4. Modellbasert estimeringsmetode

4.1. Enkel ratemodell

Metoden som brukes for å beregne foreløpige tall er basert på en enkel ratemodell, hvor vi antar at det finnes en hjelpevariabel som bidrar til å forklare statistikkvariabelen. Den statistiske modellen kan beskrives på følgende måte:

(4.1.1) $Y_i = \beta X_i + \varepsilon_i$ der $E(\varepsilon_i)=0$
$Var(\varepsilon_i)= \sigma^2 * X_i$

Y_i er en statistikkvariabel: - produksjonsverdi
- produktinnsats
- total lønn

X_i er en forklaringsvariabel: - omsetning eller sysselsetting

ε_i er feilleddet i modellen med forventning lik null og varians proporsjonal med forklaringsvariabelen.

For eksempel, hvis Y_i er produksjonsverdi i bedrift i er det rimelig å måle dette tallet mot omsetningen, altså X i samme bedriften. Omsetning og sysselsetting er variable som er tilgjengelig for hele populasjonen og kan brukes som hjelpevariable ved prediksjon. Vi henter omsetningstall fra seksjon 240 sin korttidsstatistikk for å estimere produksjonsverdi, produktinnsats og sysselsetting fra registerbasert sysselsettingsstatistikk for å estimere total lønn. Sammenhengen for utvalgsbedriftene mellom hovedvariabel og forklaringsvariabel i modellen er vist i vedlegg B.

4.2. Stratifisert ratemodell

Vi innfører betegnelsen ***h*** for stratum. Inndeling av bedrifter i strata bygger på at hvert stratum har noen felles trekk som gjør bedriftene sammenliknbare. I vårt tilfellet kan næring være opphav til en inndeling. Den stratifiserte ratemodellen ser slik ut:

(4.2.1) $Y_{i,h} = \beta_h x_{i,h} + \varepsilon_{i,h}$ for $i = 1,2,...,N_h$ og der $\text{var}(\varepsilon_{i,h}) = x_{i,h}\sigma_h^2$ for hvert stratum h

Modellen definert ved (4.2.1) forteller både at en antar at tall for bedrifter i stratum h har visse likheter, men også at de varierer i forhold til hverandre. Desto mindre de varierer innen en næring desto mer sikkert kan opplysninger om noen bedrifter i ett stratum fortelle hva tallene skal være for bedrifter utenfor utvalget i stratumet.

Prediksjonen i dette opplegget foretas etter mest mulig detaljert næring (5-siffer). Likevel finnes det næringer med få eller ingen bedrifter i utvalget og det fører til at prediksjonen må foretas på 3- siffer næringsnivå.

4.3. Estimater og usikkerhet

Vi skal nå se på hvordan en i ratemodellen estimerer den ukjente raten β_h og den ukjente populasjonsvariansen σ^2_h.

La oss anta at vi har for hvert stratum h samlet inn dataene

(4.3.1) $Y_{i,h}=y_{i,h}$, for $i \in s_h$ - (bedrift i i utvalget av bedrifter i stratum h)

Vi ser bort fra frafall og andre problemer med data og bruker hele utvalget til å estimere de ukjente parametrene, predikere den ukjente totalen og beregne usikkerheten til denne.

Estimeringen av parametrene bygger på minste kvadraters metode. Vi finner da følgende estimatorer for de to ukjente parametrene (Solheim, Faldmo og Sander, 2004):

$$\hat{\beta}_h = \frac{\sum_{i \in s_h} Y_{i,h}}{\sum_{i \in s_h} x_{i,h}} = \frac{Y_{s_h}}{x_{s_h}} \tag{4.3.2}$$

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \frac{(Y_{i,h} - \hat{\beta}_h x_{i,h})^2}{x_{i,h}} \tag{4.3.3}$$

Neste steg er å predikere en verdi for enhetene utenfor utvalget og det gjør en ved å multiplisere den estimerte raten med hjelpevariabelen.

$$\hat{Y}_{i,h} = \hat{\beta}_h x_{i,h} \tag{4.3.4}$$

Nå prediksjonen beregnes for den ukjente totalen i stratum h.

$$\hat{T}_{s_h} = \sum_{i \in s_h} Y_{i,h} + \sum_{i \notin s_h} \hat{Y}_{i,h} = \sum_{i \in s_h} Y_{i,h} + (X_h - x_{s_h})\hat{\mu}_h = \sum_{i \in s_h} \frac{X_h}{x_{s_h}} Y_{i,h} = X_h \cdot \hat{\beta}_h \tag{4.3.5}$$

Vi kan undersøke usikkerheten til den predikerte totalen i forhold til den ukjente totalen. Vi har nemlig at

$$\begin{aligned} \text{var}(\hat{T}_{s_h} - T_h \mid s_h) &= Var[(X_h - x_{s_h})\hat{\beta}_h - \sum_{i \notin s_h} Y_{i,h}] = (X_h - x_{s_h})^2 \frac{\sigma_h^2}{x_{s_h}} + (X_h - x_{s_h})\sigma_h^2 \\ &= X_h^2 \frac{X_h - x_{s_h}}{X_h} \frac{\sigma_h^2}{x_{s_h}} \end{aligned} \tag{4.3.6}$$

Ved å sette inn estimatet for variansen, den empiriske variansen gitt ved formel (4.3.3), i (4.3.6) gjelder følgende uttrykk for den empiriske variansen til avviket mellom den predikerte verdien og totalen selv.

$$\hat{V}(\hat{T}_{s_h} - T_h \mid s_h) = X_h^2 \frac{X_h - x_{s_h}}{X_h} \frac{\hat{\sigma}_h^2}{x_{s_h}} \tag{4.3.7}$$

Videre kan vi nå skrive opp standardfeilen, variasjonskoeffisienten og et 95 prosent konfidensintervall for den ukjente totalen.

$$STE(\hat{T}_{s_h} - T_h \mid s_h) = X_h \sqrt{\frac{X_h - x_{s_h}}{X_h}} \frac{\hat{\sigma}_h}{\sqrt{x_{s_h}}} \tag{4.3.8}$$

(4.3.9) $$CV(\hat{T}_{s_h} - T_h \mid s_h) = \frac{STE(\hat{T}_{s_h} - T_h \mid s_h)}{\hat{T}_{s_h}} \sqrt{\frac{X_h - x_{s_h}}{X_h}} \frac{\hat{\sigma}_h}{\hat{\beta}_h \sqrt{x_{s_h}}}$$

(.4.3.10) $$[\hat{T}_{s_h} - 1{,}96 \cdot STE(\hat{T}_{s_h} - T_h \mid s_h)\ ,\ \hat{T}_{s_h} + 1{,}96 \cdot STE(\hat{T}_{s_h} - T_h \mid s_h)]$$

4.4. Kontroll av ekstremverdier

Før vi tar i bruk modellen må vi undersøke om noen observasjoner i utvalget er ekstreme. I litteraturen brukes det studentiserte residualet for å sjekke avviket for observasjonen til enhet *j*:

$$r_{j(j),h} = \frac{\hat{Y}_{j(j),h} - Y_{j,h}}{SD(\hat{Y}_{j(j),h} - Y_{j,h})} = \frac{\hat{Y}_{j(j),h} - Y_{j,h}}{\hat{\sigma}_{h(j)}} \sqrt{\frac{x_h - x_{1,h}}{x_{j,h} x_h}}$$

Kriteriet $\left|r_{j(j),h}\right| > 2$ anbefales for å plukke ut enheter som avviker kraftig fra resten av utvalget.

Vi kan også sjekke i hvilken grad en observasjon påvirker resultatet. Da sammenliknes parameterestimatet med og uten observasjonen:

$$\hat{\beta}_{h(j)} - \hat{\beta}_h = \frac{\hat{Y}_{j(j),h} - Y_{j,h}}{x_h}$$

Følgende standardiserte uttrykk

$$DFBETAS_{h(j)} = \frac{\hat{\beta}_{h(j)} - \hat{\beta}_h}{\hat{\sigma}_{h(j)}} \sqrt{x_h} = r_{j(j),h} \sqrt{\frac{x_{j,h}}{x_h - x_{j,h}}}$$

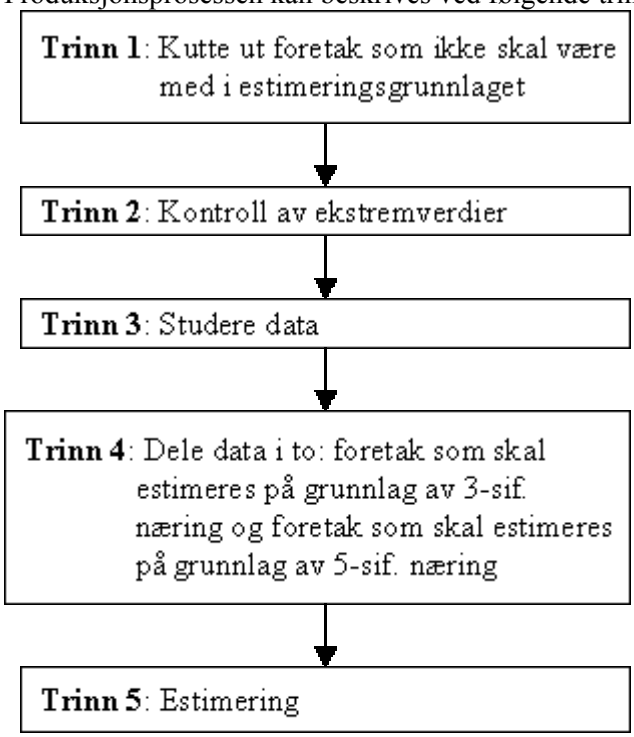
anbefales som indikator for å undersøke påvirkning på resultatet. Et vanlig kriterium for å påstå at en enhet har stor påvirkning er $\left|DFBETAS_{h(j)}\right| > 2$.

Dersom x-verdien er stor øker sjansen for at enheten både avviker sterkt og har stor innflytelse samtidig. For nærmere beskrivelse av metoden se Solheim, Faldmo og Sander, 2004.

De to kriteriene er brukt i opplegget for å identifisere utliggere. Sterkt avvikende observasjoner i forhold til resten av data i utvalget som har stor innflytelse på resultatet lot vi kun telle for seg selv, dvs. vi ekskluderte de avvikende observasjoner fra oppblåsingen.

5. Programmet for beregning av foreløpige tall

Produksjonsprosessen kan beskrives ved følgende trinn:



Filen må inneholde følgende variabler:

bnr - bedriftsnummer;
nace1 - næring;
reg_type - registerenhetstype;
mrk_utvalg - bedriften er med på utvalgsfilen;
mrk_no - foretaket er med på NO-filen;
mrk_bkf - foretaket er med på BKF-filen;
syss - sysselsetting
oms_valgt - omsetningstall hentet fra seksjon 240 sin kortidsstatistikk;
oms_bof - omsetningstall hentet fra bedriftsregisteret;
prodverd - produksjonsverdi, beregningsgrunnlag;
prodverd_no - brutto produksjonsverdi, hentet fra NO-fil;
prodinnsats - produksjonsinnsats, beregningsgrunnlag;
prodinnsats_no - produksjonsinnsats, hentet fra NO-fil;
bearbvm - bearbeidingsverdi til markedspriser, beregningsgrunnlag;
bearbvm_no - bearbeidingsverdi til markedspriser, hentet fra NO-fil;
totlonn - totale lønnskostnader, beregningsgunnlag;
totlonn_no - totale lønnskostnader, hentet fra NO-fil;
lonn_bkf - lønninger, hentet fra BKF-tall;
drinnt_bkf - sum driftsinntekter, hentet fra BKF-tall;
varef_bkf - varekostnader, hentet fra BKF-tall;
drkost_bkf - sum driftskostnader, hentet fra BKF-tall;

Vi skal nå se litt nærmere på hvert trinn. Programmene for hvert trinn finnes i vedlegg C.

Trinn 1: Spesifiserer filen og kutter ut bedrifter som ikke skal være med i estimeringsgrunnlaget:

- bedrifter med registertype 04 (hjelpebedrifter) holdes utenfor prediksjonsgrunnlaget fordi de har en annen profil enn vanlige industribedrifter;
- utvalgsbedrifter som mangler tall i hovedpostene (597 enheter);
- bedrifter med negativ omsetning eller sysselsetting må sees nærmere på (6 enheter).

På dette trinnet retter vi også på formateringen og supplerer utvalget med bedrifter som har NO-tall og BKF-tall.

Trinn 2: Kontroll av ekstremverdier.
Her sjekkes det for avvikene enheter og enheter med stor innflytelse, som tas ut av beregningene. Det kjøres to "innflytelses"modeller: en med omsetning mot produksjonsverdi og en med sysselsetting mot totall lønn. For nærmere forklaring se kapittel 4.4.
I tillegg holdes alle bedrifter med sysselsetting over 200 utenfor prediksjonsgrunnlaget (157 enheter). Til sammen var antall utliggere 273.

Trinn 3: På dette trinnet studeres data. Vi kontrollerer utvalg mot populasjonen med å lage en tabell med antall bedrifter i hver enkelt gruppe i hver næring:
ut - antall bedrifter i utvalget;
no - antall bedrifter med NO-tall;
bk - antall bedrifter med BKF-tall;
rs - antall bedrifter, som vi skal estimere tall for;
total - total antall bedrifter i populasjonen i denne næringen.

I noen tilfeller har vi ikke nok bedrifter i prediksjonsgrunnlaget sammenlignet med antall bedrifter i populasjonen. Da må man koble sammen den tomme 5-siffer næring med en annen (mest mulig lik) 5-siffer næring, eller slå den sammen i en 3-siffer næring. Fra eksempel i tabell 2 ser vi at næring 14.400 har en bedrift i populasjonen og vi har ikke tall på den bedriften fra noen av kildene. Da kobler vi sammen denne næringen med næring 14.300 hvor vi har tilstrekkelig tall i prediksjonsgrunnlaget.

Fagseksjonen kjenner næringene best og har kompetanse til å vurdere hvilke næringene som passer sammen. Derfor er det trinnet ikke automatisert og må kontrolleres for hver gang programmet kjøres.

Tabell 2. Kontroll av utvalg mot populasjonen

NACE1(NACE1) mrk

Frequency | bk	no	rs	ut	Total
10.100	0	0	2	2
10.300	3	1	9	0	13
13.100	0	0	2	1	3
13.200	0	1	0	2	3
14.110	30	24	25	8	87
14.120	5	4	10	12	31
14.130	4	13	42	4	63
14.210	86	109	215	58	468
14.300	2	2	8	4	16
14.400	0	0	1	0	1
14.500	9	3	11	4	27

Trinn 4: På dette trinnet deles filen i to: delen hvor prediksjonen foretas på 5-siffers næringsnivå og delen hvor prediksjonen foretas på 3 siffers nivå. Kriterium for beregning på 3-siffer nivå er hvis antall bedrifter i populasjonen er mer enn 0 og antall bedrifter i utvidet utvalg er lik 0.

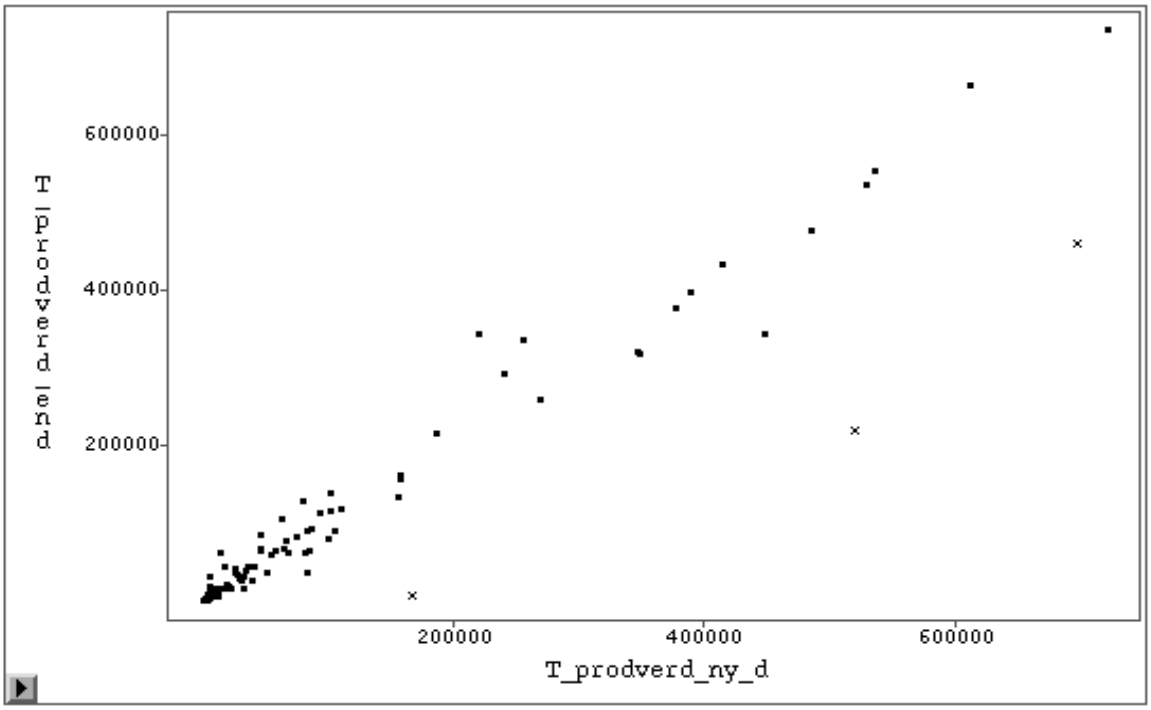
Trinn 5: Den delen av programmet er selve prediksjonen av hovedpostene for bedrifter på 5- og 3-siffer næring (se kapittel 4.1 for nærmere forklaring av modellen).

6. Resultater: foreløpige mot endelige tall

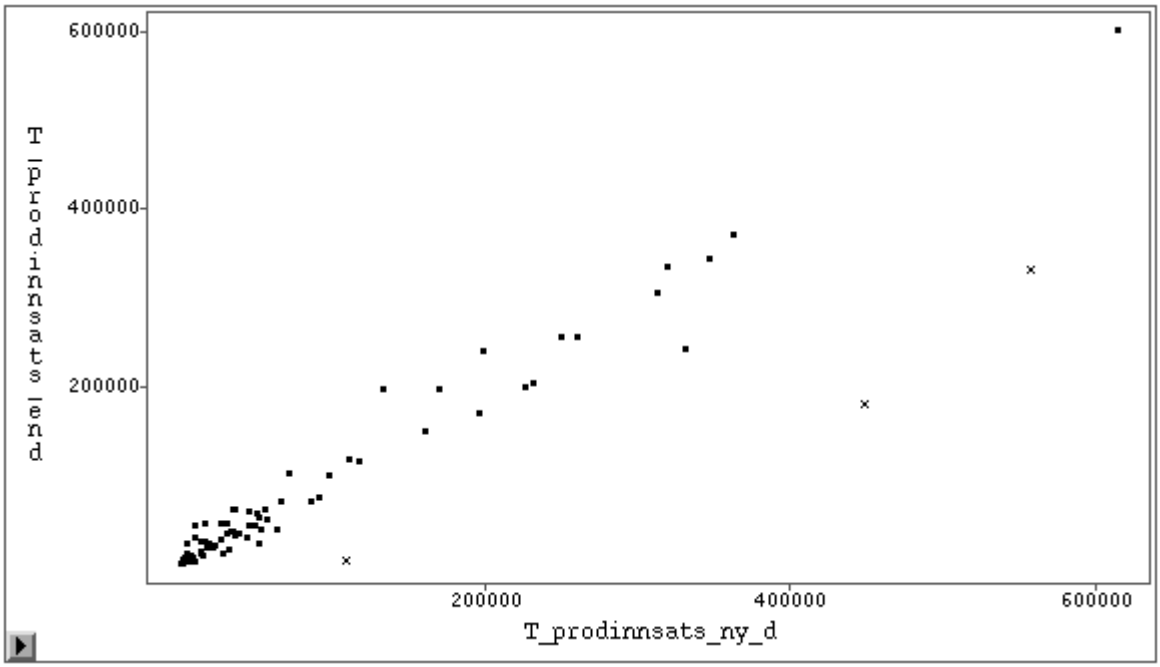
I dette avsnittet skal vi se på kvaliteten av de foreløpige tallene for år 2002 ved å sammenlikne de med de endelige tallene for samme år. Alle dataene er revidert før beregning av de endelige tallene og det er derfor mye bedre estimeringsgrunnlag for disse, men vi må huske på at også de endelige tallene er estimert og vi kan ikke betrakte dem som fasit.

Siden tallene på mikronivå ikke er de riktige verdiene skal vi sammenlikne akkumulerte tall på 3-siffer næringsnivå. Figurer 4, 5, 6, 7 og 8 viser spredningsplott over estimerte verdier for 4 hovedposter: produksjonsverdi, produktinnsats, bearbeidingsverdi og total lønn (estimert med hjelp av omsetning og estimert med hjelp av sysselsetting).

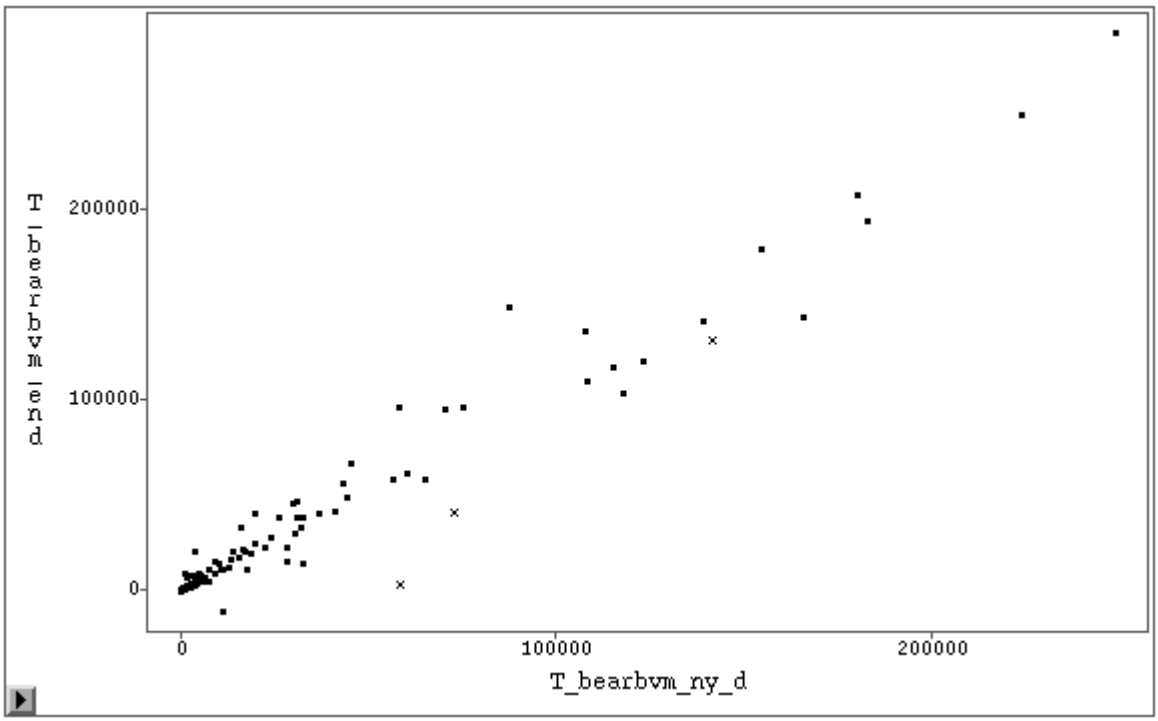
Figur 4. Estimert produksjonsverdi. Foreløpige tall mot de endelige



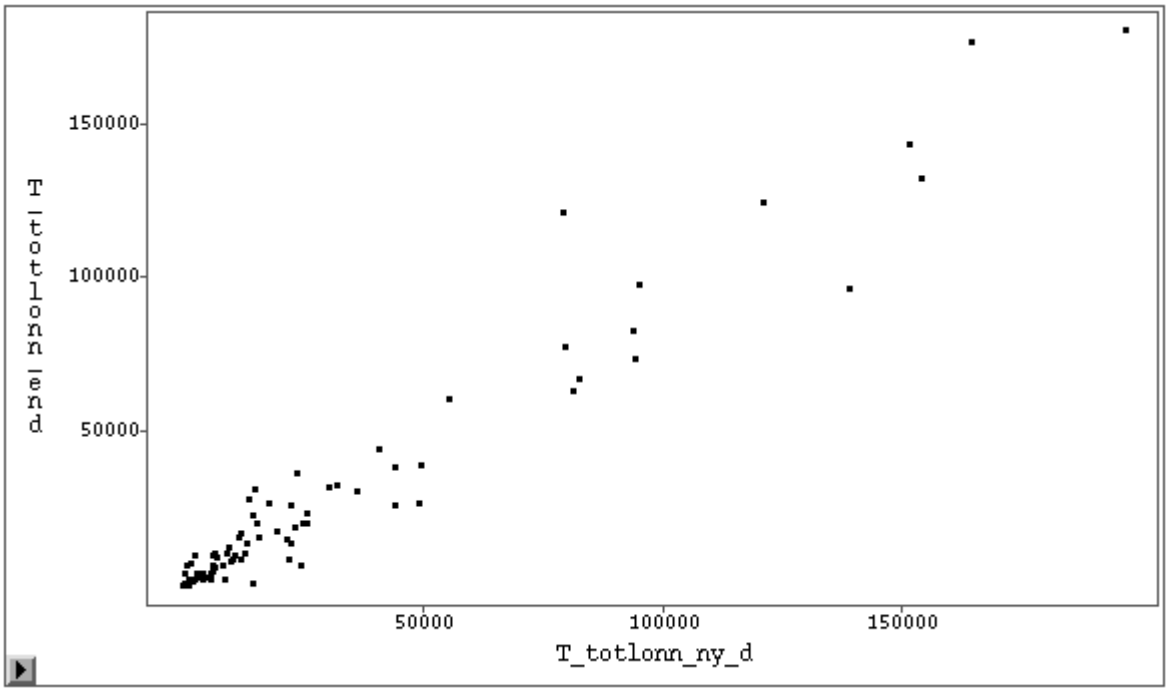
Figur 5. Estimert produktinnsats. Foreløpige tall mot de endelige



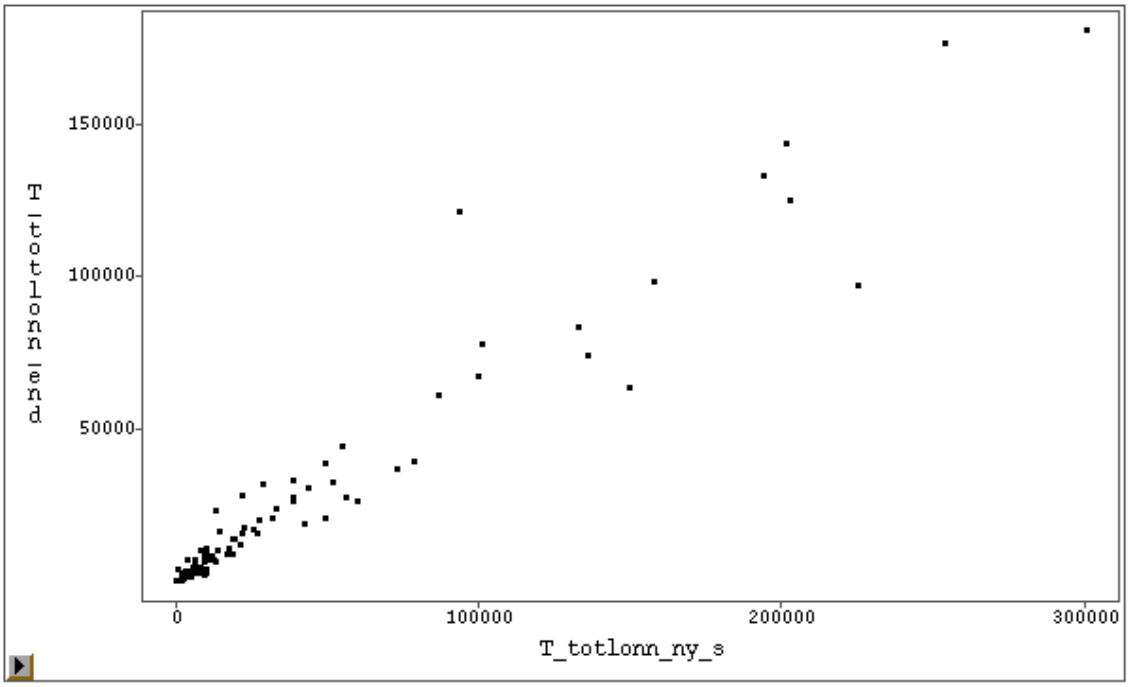
Figur 6. Estimert bearbeidingsverdi. Foreløpige tall mot de endelige



Figur 7. Estimert lønn (med hjelp av omsetning). Foreløpige tall mot de endelige



Figur 8. Estimert lønn (med hjelp av sysselsetting). Foreløpige tall mot de endelige



Vi ser at det er godt samsvar mellom de endelige og de foreløpige tallene, bortsett fra enkelte næringene som skiller seg ut. For produksjonsverdi og produktinnsats er det 3 næringer som har overestimerte verdier for de foreløpige tallene. Mens det ser ut som foreløpige tall for lønn estimert med hjelp av sysselsetting er overestimert i forhold til de endelige tallene. Dette skyldes til dels store enkeltselskaper hvor det er vanskelig å få riktige grunnlagstall for omsetning og sysselsetting på bedriftsnivå. Slike selskaper bør derfor prioriteres i revisjonen.

7. Bruk av S-KJR applikasjonen til å estimere variasjonskoeffisienter

For en rask estimering i produksjonsprosessen av foreløpige tall på 3- eller 5- siffer nivå og beregning av variasjonskoeffisienter kan vi bruke "S-KJR" applikasjonen utviklet av Leiv Solheim og Matz Ivan Faldmo. Applikasjonen beregner punktestimater og forskjellige mål for usikkerhet på en rask og brukervennlig måte.

7.1. Brukerveiledning

7.1.1. Krav til filer

Før man kan starte kjøringen av applikasjonen må man ha to datasett (lagret på Unix): en fil med populasjonen og en med utvalget.
Populasjonen må inneholde følgende variabler:

- strata (næring på 3- eller 5-siff. nivå);
- forklaringsvariabel (omsetning og sysselsetting) for alle bedrifter.

Utvalget må inneholde:

- strata (næring på 3- eller 5-siff. nivå);
- forklaringsvariabel (omsetning og sysselsetting) for alle bedrifter;
- statistikk variabler som ønskes å estimere tall for (produksjonsverdi, produktinnsats, bearbeidingsverdi og total lønn).

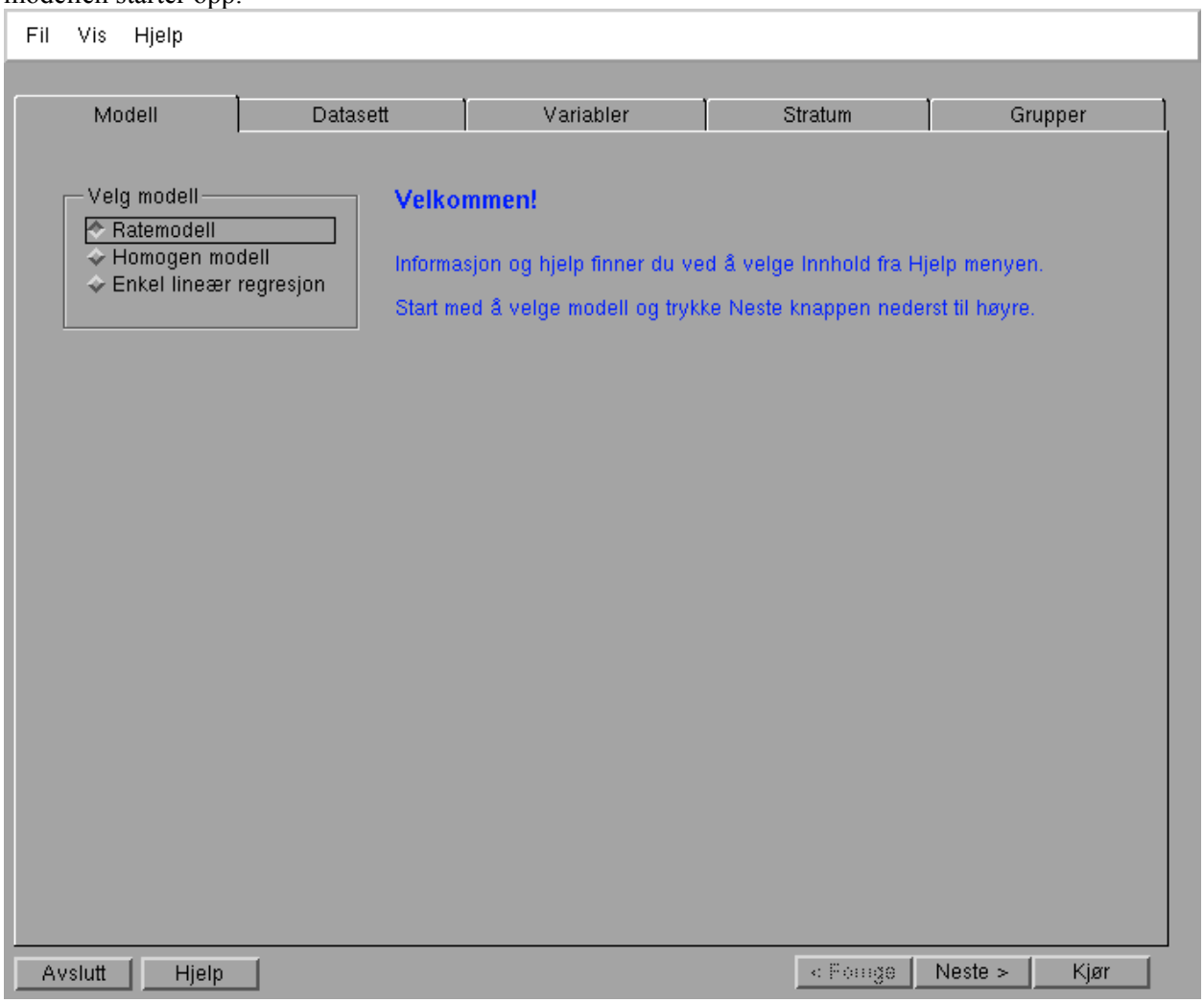
Strata må være karaktervariabel, mens forklarings- og statistikkvariabler må være numeriske.

I produksjonsprosessen anbefales det å kjøre Trinn 1, 2 og 3 fra estimeringsoppleget før bruk av S-KJR applikasjonen og ekskludere hjelpebedriftsforetakene og bedrifter med ekstreme og negative verdier fra både populasjons- og utvalgsfilen. Disse legges til etter estimeringen.

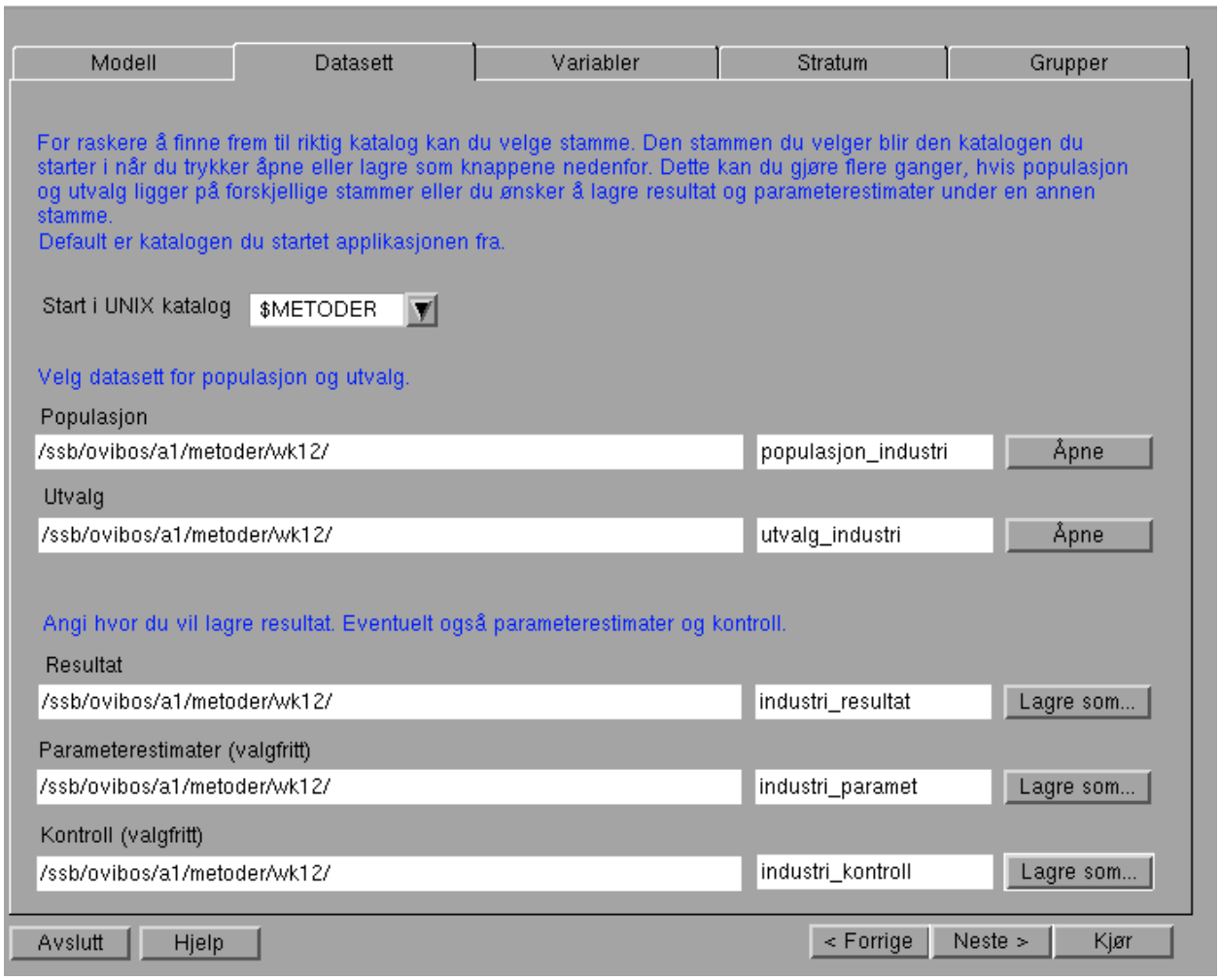
Hvis man vil studere data nærmere for enkelte næringer kan man kjøre trinn 1 og 3, dvs. ekskludere hjelpebedriftsforetakene og slå sammen næringer som ikke har tilstrekkelig antall bedrifter i utvalget, deretter bruke SAS/Insight til å vurdere enkelte bedrifter i forskjellige næringer (se nærmere forklaring i avsnitt 7.2), og så kjøre applikasjonen på nytt for å se hvordan eksklusjon av de enkelte bedrifter påvirker resultatene.

7.1.2. Programkjøring

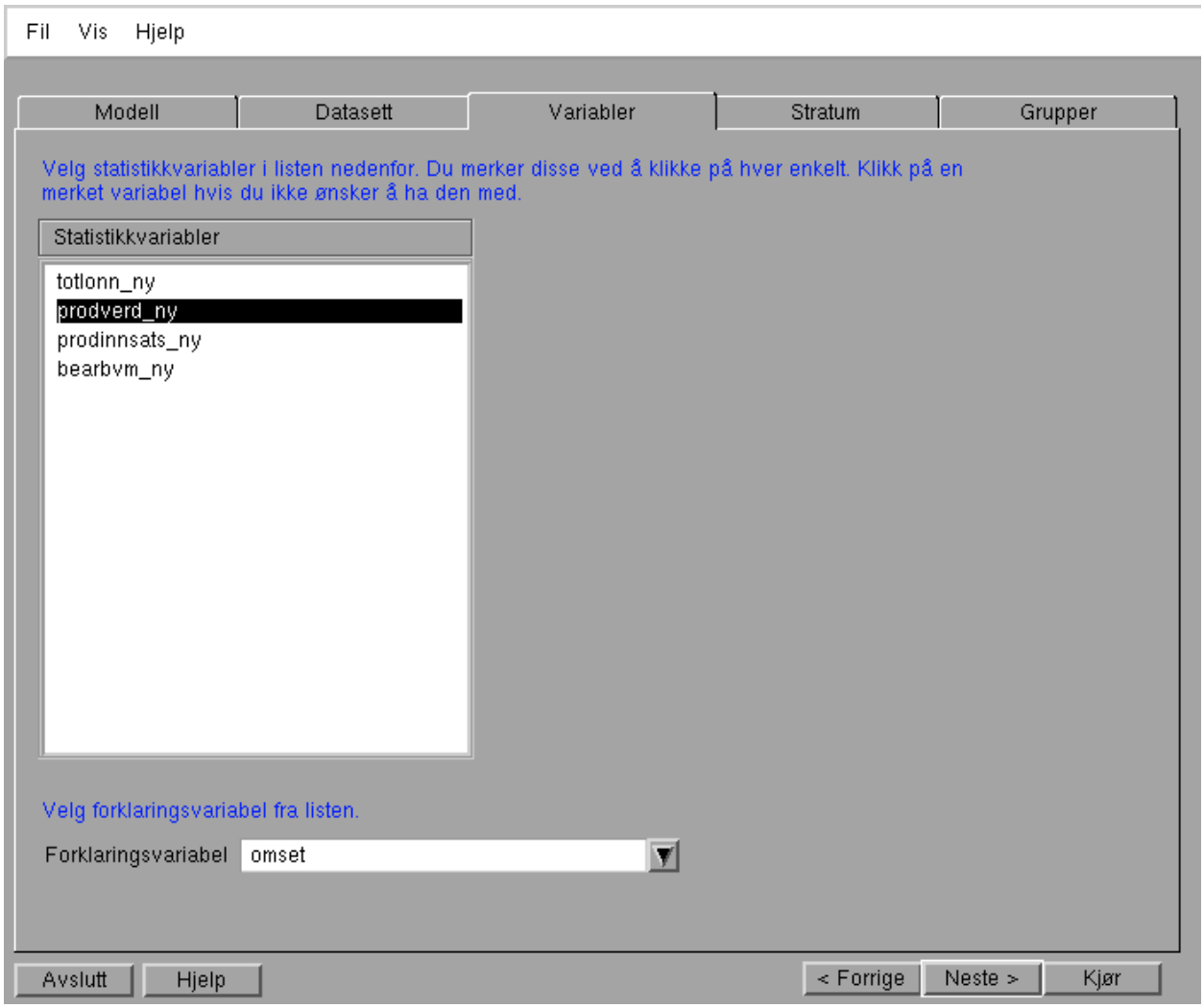
Logg deg inn på Unix og skriv kommandoen "S-KJR" for å starte applikasjonen. Trykk "enter" og modellen starter opp.



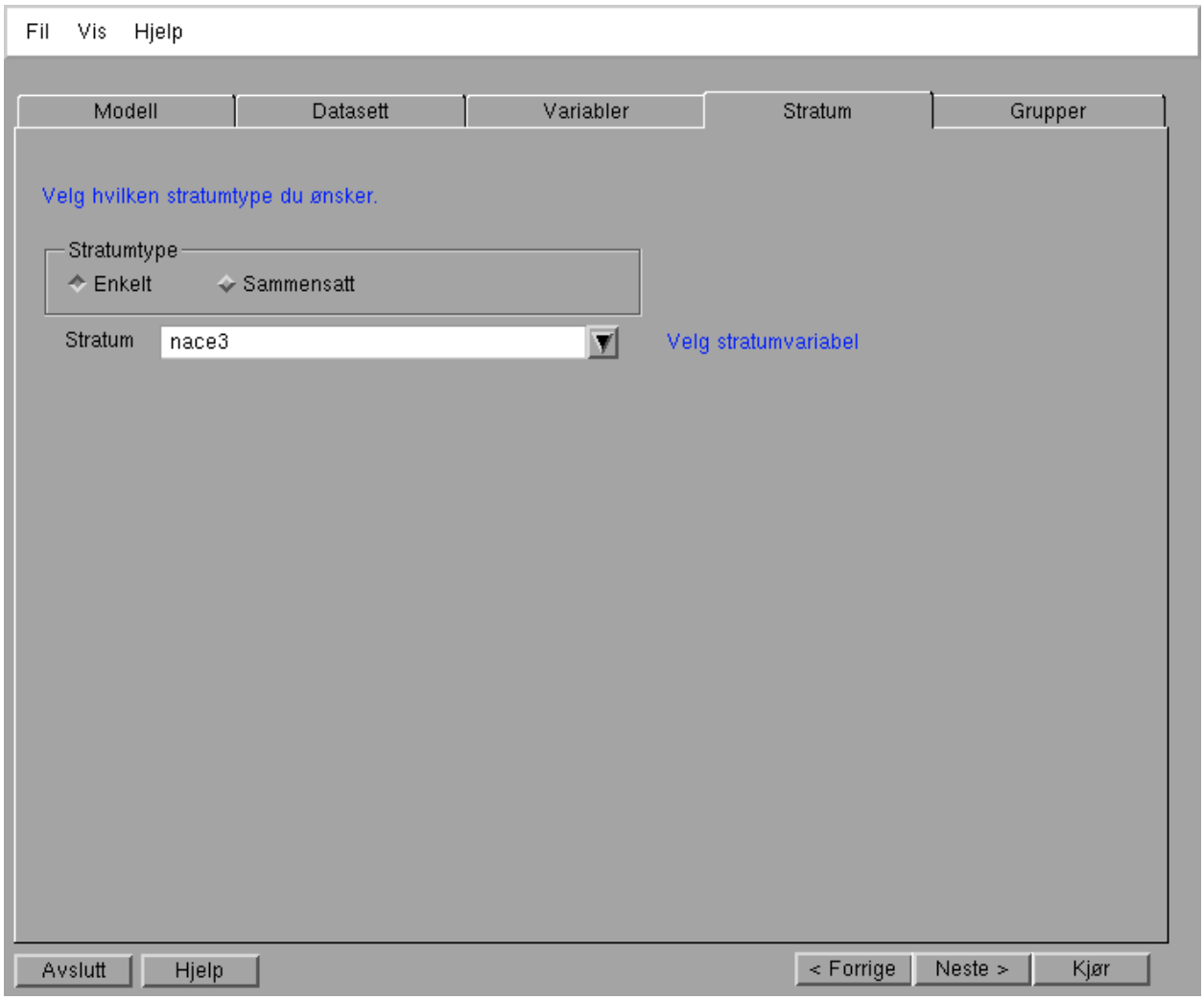
- Trykk "Ratemodell" under "Velg modell".
- Gå videre med å trykke på fanen "Datasett" eller knappen "Neste" nede til høyre.



- Definer området hvor populasjons- og utvalgs- filene er lagret.
- Angi navn på filene for resultatene, parameterestimatene og kontrollene og området hvor de skal legges.
- Gå videre til fanen "Variabler" eller "Neste" nede til høyre.



- Angi hvilke statistikkvariable som skal estimeres. Det er mulig å estimere inntil 50 variabler, men det anbefales å estimere en av gangen for å ha lettere oversikt på utskriften.
- Angi hvilken variabel som skal være forklaringsvariabel (sysselsetting for å estimere lønn, eller omsetning for å estimere alle 4 statistikkvariablene).
- Trykk på fanen "Stratum" eller "Neste" nede til høyre.



- Angi hva modellen skal benytte som stratumvariabel (næring på 3- eller 5-siff. nivå)
- Velg stratumtype "Enkelt"

Nå er alt klart til å starte estimeringen. Velg "Kjør" nederst til høyre. Det kommer opp et Outputvindu med kjøringsresultatene.

7.1.3. Forklaring av utskriften

```
Ratemodell - predikerte totaler, variasjonskoeffisienter og konfidensintervall

                 Sum_N_                    T_prodverd_   CV_prodverd_   LB_prodverd_   UB_prodverd_
Obs    land        pop      Sum_X_pop         ny            ny             ny             ny          modell
```

Første linje angir verdier for hele populasjonen, utenom utliggere og bedrifter som er ekskludert fra estimeringsgrunnlaget.
Sum_N_pop - antall bedrifter i populasjonen;
Sum_X_pop - verdi av hjelpevariabelen for hele populasjonen (her omsetning);
T_prodverd_ny - punktestimatet for variabelen vi ønsket å estimere (her produksjonsverdi);
CV_prodverd_ny - CV står for "Variation of coefficient" eller variasjonskoeffisienten. Den tolkes som hvor stor prosent standardavviket er av selve punktestimatet og regnes ut som 100 multiplisert med standardavviket dividert med punktestimatet (resultatene i "output"-vinduet gir ikke standardavviket direkte).

LB_prodverd_ny - angir nedre verdi for et 95% konfidensintervall. Nedre verdi regnes ut på følgende måte: punktestimatet - 1,96*standardavviket;
UB_prodverd_ny - angir øvre verdi for et 95% konfidensintervall: punktestimatet + 1,96*standardavviket.

De samme variablene gjentas for de enkelte næringene.

Ratemodell - predikerte totaler, variasjonskoeffisienter og konfidensintervall

Obs	nace3	Sum_N_ pop	Sum_X_pop	T_prodverd_ ny	CV_prodverd_ ny	LB_prodverd_ ny	UB_prodverd_ ny

T_prodverd_ny gir altså et estimat for den variabelen vi er opptatt av, produksjonsverdien. Hvor godt dette estimatet er kan vi bedømme ut fra å se på variasjonskoeffisienten (CV_prodverd_ny) og konfidensintervallet. Hvis vi har en variasjonskoeffisient på 10% vil dette innebære at den nedre og øvre grense for konfidensintervallet vil være 19,6% fra punktestimatet (1,96*variasjonskoeffisienten ved et 95% konfidensintervall). Nedre og øvre grense for konfidensintervallet angir for hvilket området vi med 95% grad av sikkerhet kan si at den sanne verdien befinner seg. Jo mindre dette intervallet er jo sikrere vil også estimatet for produksjonsverdien være.

Vi ser av tabellen D.1 i vedlegg D at variasjonskoeffisienten er forholdsvis stor for enkelte næringer, noe som innebærer at man enten må revidere flere innen disse næringene, eller bør se nærmere på beregningsgrunnlaget for disse. Her har vi brukt omsetningstall fra seksjon 240 sin korttidsstatistikk som hjelpevariabelen. For noen flerbedriftsforetak fordeler de omsetning på enkelte bedrifter etter en nøkkel. Hvis den nøkkelen er feil får vi feil beregningsgrunnlag for vårt formål, noe som resulterer i store variasjonskoeffisienter. Hvis vi bruker omsetningstall fra BoF som hjelpevariabelen minskes variasjonskoeffisienten for enkelte næringer betraktelig (se tabell D.1 i vedlegg D).

Videre i Output-vinduet angis verdier for parameterestimatene beta og sigma.

Ratemodell - parameterestimatene for beta og sigma

Obs	nace3	N_pop	X_POP	N_utv	X_UTV	SIGMA_ prodverd_ ny	BETA_ prodverd_ ny

Beta angir den estimerte sammenhengen mellom den valgte statistikkvariabel (her produksjonsverdi) og forklaringsvariabelen (her omsetning). Sigma er en proporsjonalitetsfaktor i standardavviket.

Til slutt i Output-vinduet kommer kontroller av datagrunnlaget.

Ratemodell - kontroll av datagrunnlaget

Obs	nace3	N_pop	X_POP	N_utv	X_UTV

Her angis det hvor mange bedrifter det er i hver enkelt strata (næring), verdi på forklaringsvariabelen i denne næringen, antall bedrifter i utvalget (utvidet) og verdi på forklaringsvariabelen i utvalget.

7.2. Eksempel på resultater og bruk av SAS/Insight for analyse av enkelte næringer

Tabell 3 viser et eksempel på total tall beregnet uten å ta hensyn til observasjoner med ekstremverdier.

Tabell 3. Estimerte tall for år 2002 (uten å ta hensyn til utliggere)

Populasjon Antall=21741 Omset=415886848,5				
	Estimerte tall	CV	Nedre grense	Øvre grense
Total lønn	86835459,56	10,37	69188275,73	104482643,38
Produksjonsverdi	438246730,36	6,34	383780077,67	492713383,05
Produktinnsats	298885244,77	5,41	267158376,73	330612112,8
Bearbeidingsverdi	139361485,59	10,77	109918717,86	168804253,32

Vedlegg E presenterer resultater av kjøringer av applikasjonen for produksjonsverdi for alle bedrifter uten å ta hensyn til utliggere (hjelpebedriftforetak er ekskludert av beregningene). Vi ser at enkelte næringer har en veldig høy variasjonskoeffisient og høy "sigma" verdi.

For eksempel, næring 282 har en variasjonskoeffisient lik CV=708. Vi kan studere den næringen nærmere ved hjelp av SAS/Insight på følgende måte:

- Vi beregner vekten som er inverse til x (B_omsetn)
  Edit ⇒ New variable ⇒ 1/Y

Y her er forklaringsvariabelen som er omsetning i dette tilfellet.

- Nå kan vi bruke ratemodellen:
  Analyze ⇒Fit (Y X)

Y - statistikkvariable (prodverd_ny)
X - forklaringsvariabel (omsetning)
Group - nace3
Weight - vekten som er laget i forrige trinn (B_omset)

Vi må også markere at det ikke skal være noe konstantledd ved å fjerne markeringen fra boksen **Intercept**. I tillegg kan vi be om å få det studentiserte residualet og dfbetas (se kap. 4.4) ved å velge

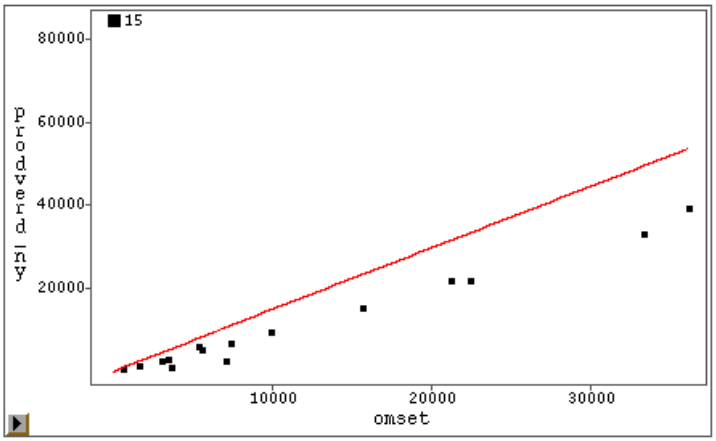
Output ⇒ Output variables ⇒ Studentized Residual og Dfbetas

Vi får følgende skjermbilde:

nace3 = 282

prodverd_ny = omset	
Response Distribution:	Normal
Link Function:	Identity

Model Equation	
prodverd_ny =	1.4858 omset



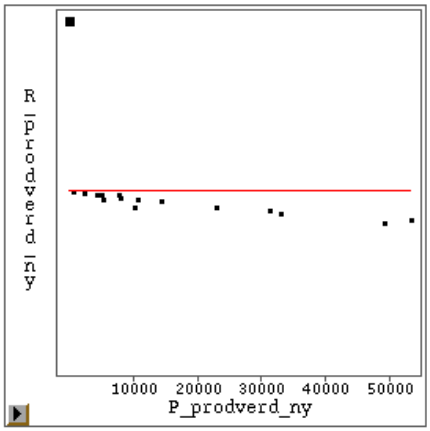
Parametric No-Intercept Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
	1	1	391504.936	15	242710161	0.0001	1.6E-03	0.9685

Summary of Fit			
Mean of Response	84678.6178	R-Square	0.0001
Root MSE	15579.1579	Adj R-Sq	0

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	391504.936	391504.936	1.6E-03	0.9685
Error	15	3.641E+09	242710161		
U Total	16	3.641E+09			

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
omset	1	391504.936	391504.936	1.613E-03	0.9685

Parameter Estimates							
Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
omset	1	1.4858	36.9936	0.04	0.9685	1.0000	1.0000



Ved å klikke på enkelte observasjoner kan vi se verdi for det studentiserte residualet (RT_prodverd_ny) og estimaten av raten dfbetas (Bpr_omset). For observasjon nummer 15 har de indikatorene svært høye verdier (RT_prodverd_ny=3,37E+03 og Bpr=1,13E+01).
Vi kan ta den observasjonen ut slik at vi kan se hva som skjer når den er fjernet fra analysen. Det gjøres ved å merke den observasjonen i plottet og deretter gjøre valget

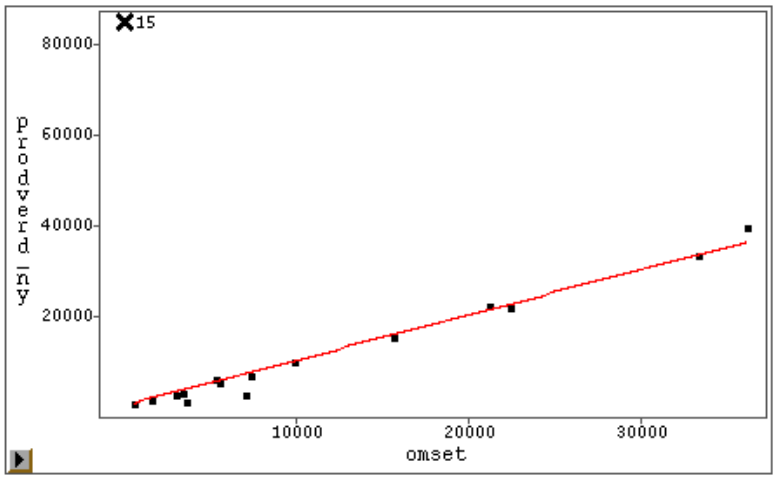
- **Edit ⇒ Observations ⇒ Exclude in Calculations**

Skjermbilde nedenfor viser at vi får en bedre tilpasning av modellen ved å holde denne observasjonen utenfor analysen.

nace3 = 282

prodverd_ny = omset	
Response Distribution:	Normal
Link Function:	Identity

Model Equation	
prodverd_ny =	1.0046 omset



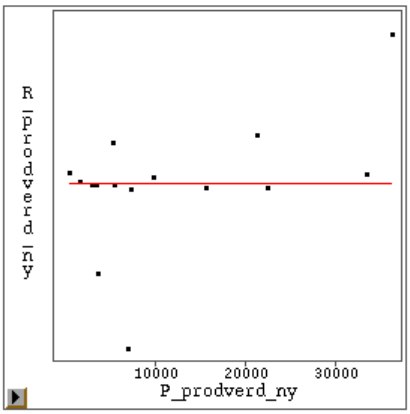
Parametric No-Intercept Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
	1	1	178994.700	14	319.8021	0.9756	559.70	<.0001

Summary of Fit			
Mean of Response	3642.7785	R-Square	0.9756
Root MSE	17.8830	Adj R-Sq	0.9739

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	178994.700	178994.700	559.70	<.0001
Error	14	4477.2295	319.8021		
U Total	15	183471.929			

Parameter Estimates

Variable	DF	Estimate	Std Error	t Stat	Pr >|t|	Tolerance	Var Inflation
omset	1	1.0046	0.0425	23.66	<.0001	1.0000	1.0000



Målet med denne øvelsen er å kartlegge utliggere i de forskjellige næringene. Men den tar lang tid og man bør selvfølgelig ikke 'rote seg bort' i alle detaljene. Programmet som er fremstilt i avsnitt 5 beregner indikatorene og skiller ut utliggere automatisk. Hvis man vil få en bedre følelse av datagrunnlaget kan man bruke SAS/Insight og metoden beskrevet ovenfor.

8. Oppsummering

I dette notatet ble opplegget for beregning av foreløpige tall presentert. Modell, estimeringsopplegg og program, samt S-KJR applikasjon for estimering av variasjonskoeffisienter og konfidensintervaller ble beskrevet.

Programmet beregner foreløpige tall på mikronivå, dvs. for enkelte bedrifter, mens applikasjonen for estimering av variasjonskoeffisienter bruker næring som strata.

Vi har sett på hvordan ekskludering av enkelte observasjoner med ekstreme verdier påvirker resultatene. Vi har også visst hvordan kvaliteten på hjelpevariabelen påvirker variasjonskoeffisient og konfidensintervall, noe som tyder på stor betydning av prioritet revisjonsarbeid.

Referanser

Solheim, Leiv, Faldmo, Matz Ivan og Jan Sander (2004): Prediksjon og usikkerhet i S-KJR modeller - I. Kursnotater, SM-03.

Tabell A.1. Antall bedrifter og dekning (%) fordelt på 3-siffer næring

nace3	antall	dekkning
101	3	0.15
103	14	0.01
131	4	0.02
132	4	0.11
141	187	0.38
142	482	0.66
143	18	0.11
144	1	0.01
145	28	0.12
151	339	4.67
152	653	5.50
153	85	0.76
154	30	0.73
155	115	2.52
156	80	0.66
157	137	2.13
158	832	3.40
159	98	2.93
160	10	1.48
171	14	0.06
172	64	0.06
173	43	0.06
174	345	0.23
175	237	0.37
176	28	0.03
177	96	0.07
181	17	0.01
182	549	0.24
183	39	0.02
191	9	0.04
192	41	0.01
193	30	0.04
201	822	1.43
202	38	0.59
203	901	1.83
204	105	0.09
205	427	0.11
211	40	2.49
212	89	0.78
221	1695	4.57
222	2025	2.27
223	111	0.05
231	1	0.00
232	12	2.49
241	100	5.09
243	40	0.48
244	35	0.95
245	70	0.59
246	54	0.33
251	62	0.11
252	401	1.44
261	131	0.52
262	230	0.17
263	1	0.00
264	5	0.01
265	21	0.21
266	369	1.43
267	187	0.15
268	91	0.53
271	46	1.49
272	53	0.13
273	6	0.01

nace3	antall	dekkning
274	44	4.93
275	51	0.34
281	544	1.72
282	26	0.06
283	4	0.02
284	28	0.04
285	1151	0.92
286	168	0.47
287	486	1.55
291	246	2.69
292	1091	2.14
293	671	0.54
294	92	0.14
295	487	1.22
296	38	0.68
297	53	0.42
300	52	0.20
311	77	0.54
312	96	0.55
313	24	0.63
314	6	0.01
315	77	0.23
316	274	0.36
321	69	0.68
322	34	1.32
323	48	0.51
331	452	0.67
332	126	1.08
333	33	0.49
334	14	0.04
335	3	0.00
341	10	0.23
342	74	0.62
343	69	0.97
351	1110	11.97
352	11	0.09
353	24	0.29
354	22	0.05
355	10	0.01
361	1307	1.76
362	245	0.13
363	37	0.01
364	54	0.19
365	52	0.01
366	327	0.14
371	78	0.32
372	78	0.11

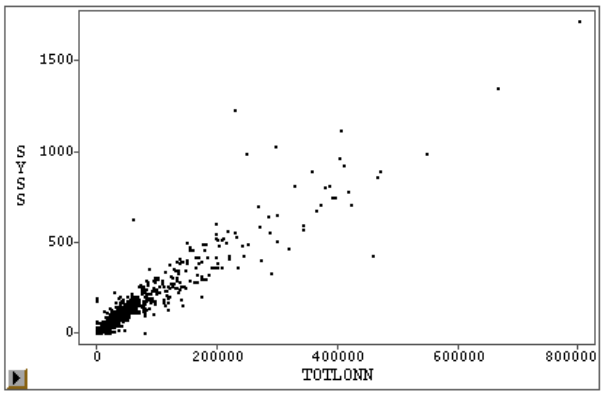
Tabell A2. Dekning av omsetning foredelt på utvalgsbedrifter, bedrifter med NO tall, bedrifter med BKF tall og resterende bedrifter i hver næring

		bk	no	rs	ut
101	Sum	.	.	.	100.00
103	Sum	27.70	4.70	67.60	.
131	Sum	.	.	4.90	95.10
132	Sum	.	0.20	.	99.80
141	Sum	10.60	12.90	6.90	69.70
142	Sum	17.00	14.30	14.20	54.50
143	Sum	0.30	3.20	3.20	93.30
144	Sum	.	.	100.00	.
145	Sum	9.90	0.40	2.70	87.00
151	Sum	2.70	1.90	1.50	93.90
152	Sum	7.40	6.90	3.10	82.60
153	Sum	7.50	2.80	2.10	87.60
154	Sum	6.90	0.40	0.00	92.60
155	Sum	0.00	3.40	0.40	96.20
156	Sum	7.80	12.30	11.20	68.70
157	Sum	3.10	3.00	5.80	88.10
158	Sum	3.40	3.30	3.80	89.60
159	Sum	0.80	0.40	0.50	98.30
160	Sum	.	.	.	100.00
171	Sum	1.20	0.70	0.20	97.90
172	Sum	0.40	1.70	3.70	94.20
173	Sum	9.10	13.10	4.50	73.30
174	Sum	8.40	16.80	11.70	63.10
175	Sum	12.40	6.90	5.70	75.10
176	Sum	22.10	0.70	1.90	75.20
177	Sum	6.70	6.20	10.00	77.10
181	Sum	86.70	7.70	5.60	.
182	Sum	3.40	15.50	9.10	71.90
183	Sum	11.70	12.80	23.50	52.10
191	Sum	0.30	1.20	0.20	98.30
192	Sum	43.10	22.60	34.30	.
193	Sum	8.60	5.30	6.60	79.50
201	Sum	13.10	5.80	7.30	73.70
202	Sum	2.30	6.80	0.40	90.50
203	Sum	9.40	10.90	4.40	75.30
204	Sum	16.30	17.70	9.20	56.80
205	Sum	17.00	25.90	27.90	29.30
211	Sum	0.10	0.30	0.90	98.60
212	Sum	6.10	2.30	3.90	87.70
221	Sum	6.60	5.80	2.60	85.00
222	Sum	14.70	15.30	6.60	63.40
223	Sum	31.60	13.00	14.40	40.90
231	Sum	.	100.00	.	.
232	Sum	0.00	0.00	0.10	99.80
241	Sum	0.40	0.30	0.00	99.30
243	Sum	3.80	0.70	5.60	89.90
244	Sum	0.30	0.40	0.50	98.80
245	Sum	1.20	1.50	0.80	96.50
246	Sum	6.90	6.10	0.90	86.10
251	Sum	17.80	8.30	15.80	58.10
252	Sum	9.30	8.30	2.60	79.80
261	Sum	2.30	5.20	1.70	90.80
262	Sum	1.60	2.70	6.00	89.70
263	Sum	.	.	100.00	.
264	Sum	.	5.10	1.00	93.90
265	Sum	4.20	.	0.30	95.60
266	Sum	8.90	11.30	4.30	75.60
267	Sum	15.40	29.40	10.80	44.40

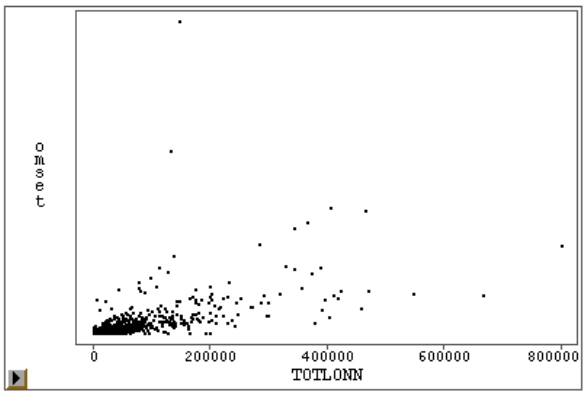
		bk	no	rs	ut
268	Sum	0.20	1.60	0.70	97.50
271	Sum	0.90	0.40	0.30	98.40
272	Sum	13.00	16.30	13.00	57.70
273	Sum	.	1.10	0.80	98.00
274	Sum	0.00	0.00	0.20	99.80
275	Sum	5.70	3.20	0.60	90.40
281	Sum	10.90	10.60	3.00	75.60
282	Sum	13.70	7.20	5.50	73.50
283	Sum	2.60	1.90	.	95.50
284	Sum	1.70	19.40	5.40	73.50
285	Sum	12.90	16.10	8.80	62.10
286	Sum	7.50	3.80	2.00	86.60
287	Sum	5.00	7.20	2.40	85.40
291	Sum	2.10	2.40	0.80	94.60
292	Sum	10.60	10.20	5.30	73.90
293	Sum	7.20	11.50	16.40	65.00
294	Sum	10.40	10.80	5.60	73.20
295	Sum	10.70	8.20	3.30	77.70
296	Sum	0.30	0.50	0.30	98.90
297	Sum	0.40	1.30	2.90	95.40
300	Sum	4.20	6.60	4.10	85.10
311	Sum	2.70	3.30	1.70	92.20
312	Sum	6.30	5.90	0.60	87.30
313	Sum	0.10	1.00	0.20	98.60
314	Sum	.	38.30	1.00	60.80
315	Sum	3.50	4.80	1.90	89.90
316	Sum	11.40	20.70	5.70	62.20
321	Sum	1.80	2.90	0.40	94.80
322	Sum	0.20	0.80	0.00	99.00
323	Sum	0.50	0.90	0.20	98.30
331	Sum	6.80	10.00	3.10	80.10
332	Sum	6.50	3.50	0.50	89.40
333	Sum	5.50	2.70	0.10	91.70
334	Sum	7.00	4.30	1.80	86.90
335	Sum	89.30	.	10.70	.
341	Sum	0.80	1.00	0.30	98.00
342	Sum	1.20	6.30	1.70	90.80
343	Sum	0.60	1.20	1.50	96.70
351	Sum	2.30	2.00	0.70	95.00
352	Sum	0.30	5.10	0.10	94.40
353	Sum	1.50	1.20	2.60	94.70
354	Sum	0.00	12.40	3.80	83.80
355	Sum	1.10	3.70	1.40	93.80
361	Sum	5.80	9.50	5.60	79.10
362	Sum	8.80	8.90	18.70	63.60
363	Sum	8.10	28.00	26.50	37.40
364	Sum	6.00	2.90	1.60	89.50
365	Sum	51.80	32.10	16.10	.
366	Sum	23.40	21.20	10.20	45.10
371	Sum	15.50	13.80	11.20	59.50
372	Sum	36.80	30.30	5.80	27.10

Kontroll av hovedvariabel mot hjelpevariabel

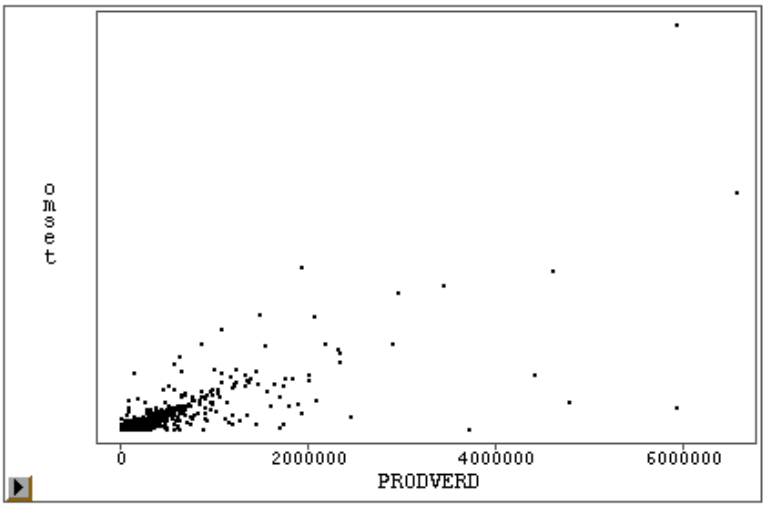
Figur 1. Totallønn mot sysselsetting



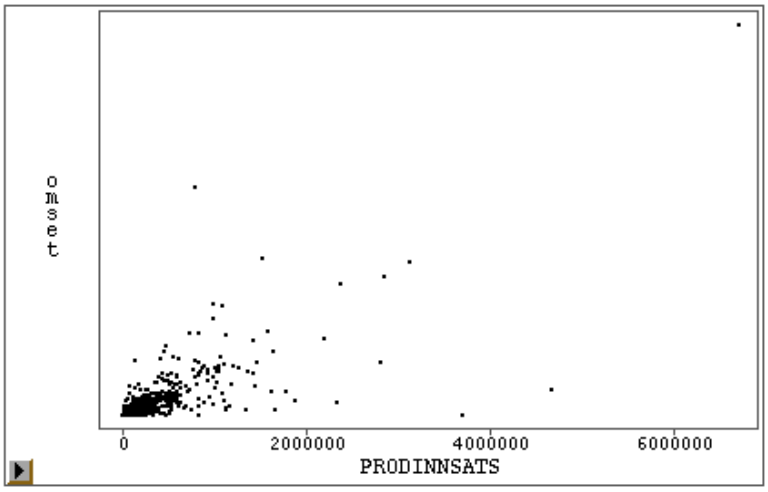
Figur 2. Totallønn mot omsetning



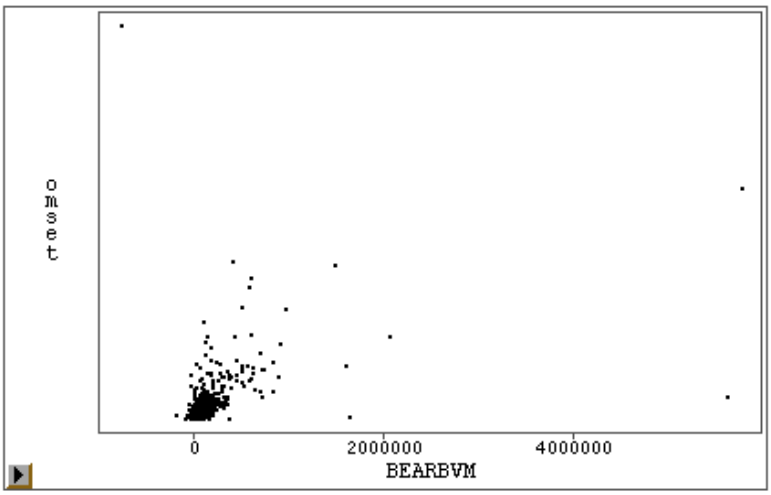
Figur 3. Produksjonsverdi mot omsetning



Figur 4. Produktinnsats mot omsetning



Figur 5. Bearbeidingsverdi mot omsetning



Program for beregning av foreløpige tall.

Trinn 1:

```
* Spesifiserer filen som skal leses;
libname ind"$INDUSTR/struktur/wk24/is02";
run;

data industri;
set ind.forelopig_20040705;
run;

data industri;
set industri;
length nace2 $2;
length nace3 $3;
length nac3 $3;
length nacex $5;
nace2=substr(nace1,1,2);
nac3=substr(nace1,4,3);
nacex=nace2||nac3;
nace3=substr(nacex,1,3);
drop nac3;
run;

* Skiller ut bedrifter med registerenhetstype 4-hjelpeforetak som ikke skal brukes ved oppblåsing;
data type4 spes estfil;
set industri;
if lonn_bkf<0 or drinnt_bkf<0 or varef_bkf <0 or drkost_bkf <0 then do;
drinnt_bkf = . ;
lonn_bkf = . ;
varef_bkf = . ;
drkost_bkf = . ;
end;
if mrk_utvalg='1' and reg_type='04' then output type4;
else if mrk_utvalg='1' and prodverd = . then output spes;
else output estfil;
run;

data type4;
set type4;
prodinnsats_ny=prodinnsats;
prodverd_ny=prodverd;
bearbvm_ny=bearbvm;
totlonn_ny=totlonn;
totlonn_ny_s=totlonn;
run;

* Setter merke på bedrifter med negative tall fra NO (skal ikke brukes i estimeringsgrunnlaget);
data estfil;
```

```
set estfil;
if mrk_utvalg=' ' and mrk_no='1' then do;
if totlonn_no<0 then mrk_no_neg='1';
else if prodinnsats_no<0 then mrk_no_neg='1';
else if prodverd_no<0 then mrk_no_neg='1';
else mrk_no_neg=' ';
end;
run;
proc sort data=estfil;
by nace1;
run;

data estfil1;
set estfil;
if reg_type > '01' then utbkf='1';
else if mrk_utvalg='1' and reg_type ne '01' then utbkf='2';
else if mrk_utvalg='1' and reg_type='01' and skjema='I2' then utbkf='9';
else if drinnt_bkf= . then utbkf='3';
else if drinnt_bkf<0 then utbkf='4';
else if drinnt_bkf=0 and prodverd>0 then utbkf='5';
else if lonn_bkf=0 and totlonn>0 then utbkf='6';
else utbkf ='0';
run;

data estfil2;
set estfil1;
if oms_valgt= .  or oms_valgt<0 then omset=oms_bof;
else omset=oms_valgt;
run;

* Skiller ut bedrifter med negative omsetning eller sysselsetting;
data estfil3 spes1;
set estfil2;
if omset<0 or syss<0 then output spes1;
else output estfil3;
run;

* Rettes på formateringen for å unngå vekter som er mer enn 1;
data estfil4;
set estfil3;
syss_ny=syss*1;
omset_ny=omset*1;
if 0<syss_ny<0.5 then syss=0;
else if 0.5<=syss_ny<1 then syss=1;
if 0<omset_ny<0.5 then omset=0;
else if 0.5<=omset_ny<1 then omset=1;
run;

data estfil5;
set estfil4;
if mrk_utvalg='1' then do;
mrk='ut';
prodverd_ny=prodverd;
prodinnsats_ny=prodinnsats;
```

```
bearbvm_ny=bearbvm;
totlonn_ny=totlonn;
totlonn_ny_s=totlonn;
end;
else if reg_type='01' and mrk_utvalg=' ' and mrk_no='1' then do;
if mrk_no_neg=' ' then do;
mrk='no';
prodverd_ny=prodverd_no;
prodinnsats_ny=prodinnsats_no;
bearbvm_ny=bearbvm_no;
totlonn_ny=totlonn_no;
totlonn_ny_s=totlonn_no;
end;
else if mrk_no_neg='1' and mrk_bkf='1' and utbkf='0' and drinnt_bkf>0 then do;
mrk='bk';
prodverd_ny=drinnt_bkf;
prodinnsats_ny=varef_bkf+adrkost_bkf;
bearbvm_ny=prodverd_ny-prodinnsats_ny;
totlonn_ny=lonn_bkf;
totlonn_ny_s=lonn_bkf;
end;
else mrk='rs';
end;
else if reg_type='01' and mrk_utvalg=' ' and mrk_no=' ' and mrk_bkf='1' and utbkf='0' and drinnt_bkf>0
then do;
mrk='bk';
prodverd_ny=drinnt_bkf;
prodinnsats_ny=varef_bkf+adrkost_bkf;
bearbvm_ny=prodverd_ny-prodinnsats_ny;
totlonn_ny=lonn_bkf;
totlonn_ny_s=lonn_bkf;
end;
else mrk='rs';
run;
```

Trinn 2:

```
* Lager vekter;
data estfil_ny;
set estfil6;
if omset=0 then vekt=0;
else vekt=1/omset;
run;

data estfil_ny1;
set estfil_ny;
if syss=0 then vekt_syss=0;
else vekt_syss=1/syss;
run;

* Beregner det studentisete residualet og DFBETAS basert på omsetning mot produksjonsverdi;
proc means data=estfil_ny1 noprint nway;
class nacex;
```

```
var omset syss;
output out=test (rename=(_freq_=antx1) drop=_type_)
sum(omset)=sum_omset_nace
sum(syss)=sum_syss_nace
;
run;

proc sort data=estfil_ny1;
by nacex;
run;

data estfil_ny2 tull;
merge estfil_ny1(in=a) test(in=b);
by nacex;
if a and b then output estfil_ny2;
else output tull;
run;

proc reg data=estfil_ny2 noprint;
by nacex;
model prodverd=omset/ influence noint;
weight vekt;
output out=predxa
  p=pprodv
  rstudent=rprodv
  ;
run;

data predxa;
set predxa;
dfbetas_prodv=rprodv*sqrt(omset/(sum_omset_nace-omset));
run;

* Beregner det studentisete residualet og DFBETAS basert på sysselsetting mot totall lønn;
proc reg data=estfil_ny2 noprint;
by nacex;
model totlonn=syss/ influence noint;
weight vekt_syss;
output out=predxa_syss
  p=ptotlonn
  rstudent=rtotlonn
  ;
run;

data predxa_syss;
set predxa_syss;
dfbetas_totlonn=rtotlonn*sqrt(syss/(sum_syss_nace-syss));
run;

* Identifiserer utliggere basert på omsetning;
data estfil_x utligg_oms;
set predxa;
if mrk='ut' or mrk='no' or mrk='bk' then
do;
```

```
   if abs(rprodv)>2 and abs(dfbetas_prodv)>2 then output utligg_oms;
else output estfil_x;
end;
else output estfil_x;
run;

* Identifiserer utliggere basert på sysselsetting;
data est_ny utligg_syss;
set predxa_syss;
if mrk='ut' or mrk='no' or mrk='bk' then do;
   if abs(rtotlonn)>2 and abs(dfbetas_totlonn)>2 then output utligg_syss;
else output est_ny;
end;
else output est_ny;
run;

* Skiller ut utvalgsbedrifter med negative produksjonsverdi eller produktinnsats og bedrifter med
sysselsetting mer enn 200;
data est_ny1 utligg1;
set estfil_x;
if mrk_utvalg='1' then do;
if prodinnsats<0 then output utligg1;
if prodverd<0 then output utligg1;
if syss>200 then output utligg1;
end;
else output est_ny1;
run;

* Kobler sammen filene med utliggere;
proc sort data=utligg_oms; by bnr;
run;

proc sort data=utligg_syss; by bnr;
run;

proc sort data=utligg1; by bnr;
run;

data utliggere;
merge utligg1(in=a) utligg_syss(in=b) utligg_oms(in=c);
by bnr;
if a then utligg='1';
if b then utligg_syss='1';
if c then utligg_oms='1';
run;

data utliggere;
set utliggere;
prodinnsats_ny=prodinnsats;
prodverd_ny=prodverd;
bearbvm_ny=bearbvm;
totlonn_ny=totlonn;
totlonn_ny_s=totlonn;
run;
```

```
proc sort data=utliggere; by bnr;
proc sort data=estfil_ny2; by bnr;
run;

data est_ny;
merge estfil_ny2(in=a) utliggere(in=b);
by bnr;
if a and not b;
run;
```

Trinn 3:

```
* Kontroll av utvalg mot populasjon etter næring;
proc freq data=estfil5 ;
tables nace1*mrk/ nocol nocum nopercent norow;
run;

* OBS!!! Denne delen må kontrolleres mot tabellen;
data estfil6;
set estfil5;
if nace1='14.400' then do;
  nacex='14300';
  nace3='143';
  nace2='14';
end;
if nace1='15.930'  then do;
  nacex='15940';
  nace3='159';
  nace2='15';
end;
if nace1='17.140' or nace1='17.170' then do;
  nacex='17130';
  nace3='171';
  nace2='17';
end;
if nace1='17.510' then do;
  nacex='17520';
  nace3='175';
  nace2='17';
end;
if nace1='20.520' then do;
  nacex='20510';
  nace3='205';
  nace2='20';
end;
if nace1='24.200' or nace1='24.170' then do;
  nacex='24160';
  nace3='241';
  nace2='24';
end;
if nace1='24.640' then do;
  nacex='24630';
  nace3='246';
```

```
  nace2='24';
end;
if nace1='26.300' then do;
  nacex='26260';
  nace3='262';
  nace2='16';
end;
if nace1='35.410' then do;
  nacex='35420';
  nace3='354';
  nace2='35';
end;
if nace1='33.500' then do;
  nacex='33400';
  nace3='334';
  nace2='33';
end;
run;
```

Trinn 4:

```
proc sort data=est_ny;
by nacex;
run;

* Beregner antall bedrifter i populasjonen og i utvidet utvalg på 5-siffer nivå;
data telloppx;
set est_ny (keep= nacex nace3 nace2 mrk_utvalg mrk_no mrk);
retain  antpopx antmrkx;
by nacex;
antpopx+1;
if mrk='ut' or mrk='no' or mrk='bk' then antmrkx+1;
if last.nacex then do;
output;
antpopx=0;
antmrkx=0;
end;
run;

* Beregner antall bedrifter i populasjonen og i utvidet utvalg på 3-siffer nivå;
data tellopp3;
set est_ny (keep=nace3 mrk_utvalg mrk_no mrk);
retain  antpop3 antmrk3;
by nace3;
antpop3+1;
if mrk='ut' or mrk='no' or mrk='bk' then antmrk3+1;
if last.nace3 then do;
output;
antpop3=0;
antmrk3=0;
end;
run;
```

```
* Kobler filene sammen;
data tellopp2;
merge telloppx (drop= mrk_utvalg utbkf drinnt_bkf)
      tellopp3 (drop= mrk_utvalg utbkf drinnt_bkf);
by nace3;
run;

* Lager filer med bedrifter som skal estimeres på 5-siffer nivå og på 3-siffer nivå;
data tresif;
set tellopp2 (keep= nace3 antpopx antmrkx);
if antmrkx=0 and antpopx>0;
run;

proc sort data=tresif (keep=nace3) nodupkey;
by nace3;
run;

data tresif;
set tresif;
sif='3';
run;

data est;
merge est_ny tresif;
by nace3;
run;

data est5 est3;
set est;
if sif='3' then output est3;
else output est5;
run;

data est5;
merge est5 (drop=sif in=uu) telloppx (keep= nacex  antpopx antmrkx);
by nacex;
if uu;
run;

data est3;
merge est3 (drop=sif in=jj) tellopp3 (keep= nace3 antpop3 antmrk3);
by nace3;
if jj;
run;
```

Trinn 5:

```
* Beregner hovedpostene for bedrifter på 5-siffer nivå;
proc means data=est5 sum noprint;
by nacex;
var omset syss totlonn_ny prodinnsats_ny prodverd_ny bearbvm_ny;
output out=sumomsx
sum= oomset osyss ototlonn oprodinnsats oprodverd obearbvm;
run;
```

```
data andomsx;
set sumomsx;
atotlonn=ototlonn/oomset;
aprodinnsats=oprodinnsats/oomset;
satotlonn=ototlonn/osyss;
aprodverd=oprodverd/oomset;
abearbvm=obearbvm/oomset;
run;

data andomsx;
set andomsx (keep=nacex atotlonn aprodverd aprodinnsats abearbvm satotlonn);
run;

data estimerx;
merge est5 andomsx;
by nacex;
run;

data estimerex;
set estimerx;
if mrk='rs' then do;
totlonn_ny=round(atotlonn*omset, 0.1);
totlonn_ny_s=round(satotlonn*syss, 0.1);
prodverd_ny=round(aprodverd*omset, 0.1);
prodinnsats_ny=round(aprodinnsats*omset, 0.1);
bearbvm_ny=round(abearbvm*omset, 0.1);
end;
run;

* Beregner hovedpostene for bedrifter på 3-siffer nivå;
proc means data=est3 sum noprint;
by nace3;
var omset syss totlonn_ny prodinnsats_ny prodverd_ny bearbvm_ny;
output out=sumoms3
sum=oomset osyss ototlonn oprodinnsats oprodverd obearbvm;
run;

data andoms3;
set sumoms3;
atotlonn=ototlonn/oomset;
aprodinnsats=oprodinnsats/oomset;
satotlonn=ototlonn/osyss;
aprodverd=oprodverd/oomset;
abearbvm=obearbvm/oomset;
run;

data andoms3;
set andoms3 (keep=nace3 atotlonn aprodverd aprodinnsats abearbvm satotlonn);
run;

data estimer3;
merge est3 andoms3;
by nace3;
run;
```

```
data estimere3;
set estimer3;
if mrk='rs' then do;
totlonn_ny=round(atotlonn*omset, 0.1);
totlonn_ny_s=round(satotlonn*syss, 0.1);
prodverd_ny=round(aprodverd*omset, 0.1);
prodinnsats_ny=round(aprodinnsats*omset,0.1);
bearbvm_ny=round(abearbvm*omset, 0.1);
end;
run;

* Setter sammen filene med predikerte verdier, ekstreme verdier og hjelpebedriftsforetak;
data estferdig;
set estimerex estimere3 utliggere type4;
run;
```

Tabell D.1. Variasjonskoeffisienter beregnet med hjelp av omsetningstall fra seksjon 240 og omsetningstall fra BoF for 3-siffer næring

Obs	nace3	CV_prodverd_ny	CV_prodverd_bof
1		7.820	0.1251
2	101	0.000	0.0000
3	103	6.821	6.7300
4	131	0.000	0.0000
5	132	0.000	0.0000
6	141	1.740	0.8667
7	142	10.748	0.4358
8	143	0.726	0.2886
9	145	0.740	0.2465
10	151	23.569	0.4173
11	152	13.175	0.4581
12	153	11.219	10.1893
13	154	0.354	0.1399
14	155	0.435	0.1486
15	156	2.924	1.4960
16	157	2.691	0.3257
17	158	0.435	0.1848
18	159	1.188	0.4029
19	160	0.000	0.0000
20	171	3.265	0.1312
21	172	0.912	0.6349
22	173	2.777	2.4940
23	174	0.924	0.7131
24	175	32.663	0.5618
25	176	1.667	1.6542
26	177	0.909	0.6392
27	181	0.588	0.7828
28	182	1.235	0.8159
29	183	4.574	4.1002
30	191	3.432	3.4490
31	192	4.235	4.5622
32	193	3.821	5.1687
33	201	37.667	0.3073
34	202	0.246	0.0999
35	203	0.640	0.2002
36	204	1.344	0.6790
37	205	55.743	1.1015
38	211	2.170	0.3084
39	212	1.732	0.4646
40	221	56.609	0.7446
41	222	14.196	0.3128
42	223	0.961	0.5518
43	231	0.000	0.0000
44	232	26.341	1.2155
45	241	3.910	0.0974
46	243	2.399	1.5447
47	244	236.497	0.7186
48	245	0.458	0.3625
49	246	0.757	0.6941
50	251	2.829	2.3440
51	252	10.292	0.1145
52	261	62.964	0.6571
53	262	2.773	1.8886
54	264	9.704	9.2121
55	265	0.143	0.0236
56	266	11.019	0.2009
57	267	18.691	0.7183
58	268	1.041	0.1719
59	271	0.703	0.3556
60	272	2.482	0.9523
61	273	16.138	16.8224

Obs	nace3	CV_prodverd_ny	CV_prodverd_bof
62	274	0.630	0.0774
63	275	2.197	0.5881
64	281	6.415	0.4935
65	282	1.203	1.1572
66	283	0.000	0.0000
67	284	0.676	0.6259
68	285	72.106	0.1852
69	286	0.456	0.3312
70	287	0.478	0.2030
71	291	7.194	0.3022
72	292	11.323	0.2013
73	293	23.101	2.4747
74	294	0.646	0.5245
75	295	1.045	0.3091
76	296	1.438	0.3908
77	297	0.443	0.4735
78	300	3.982	2.5753
79	311	0.537	0.3731
80	312	0.244	0.1281
81	313	1.178	0.3842
82	314	22.863	22.8629
83	315	1.718	0.7754
84	316	6.242	0.4558
85	321	0.492	0.2469
86	322	0.071	0.0712
87	323	1.225	0.5284
88	331	20.476	3.9295
89	332	23.470	0.1372
90	333	89.152	0.1252
91	334	1.406	1.2356
92	341	3.705	3.0806
93	342	0.964	0.8336
94	343	571.176	0.6901
95	351	69.016	0.1028
96	352	0.117	0.0569
97	353	1.674	1.2945
98	354	2.252	1.8879
99	355	8.266	6.5812
100	361	0.664	0.2093
101	362	0.865	0.7481
102	363	8.080	3.3864
103	364	0.805	0.7674
104	365	9.329	2.1961
105	366	14.454	0.9990
106	371	91.268	1.2347
107	372	1.160	0.6170

Tabell E.1. Variasjonskoeffisient, "sigma" og "beta" verdi for produksjonsverdi for hele populasjonen (med utliggere) for 3-siffer næring

nace3	CV_prodverd_ny	SIGMA_prodverd_ny	BETA_prodverd_ny
	6.3410	.	.
101	0.0000	2.62	1.03964
103	6.8210	7.66	0.95995
131	0.0000	0.00	0.54620
132	0.0000	252.54	1.11799
141	1.7400	73.66	0.88508
142	10.7484	441.65	1.06853
143	0.7259	14.42	1.01385
145	0.3822	16.36	0.98826
151	12.1043	9172.25	2.22984
152	13.0746	2868.01	0.92699
153	8.8822	953.60	0.89962
154	0.1780	133.14	0.78432
155	1.1553	847.94	1.36788
156	2.4202	102.09	0.88847
157	2.6914	378.99	1.33536
158	0.2733	46.14	0.93349
159	0.4041	196.48	1.02549
160	0.0000	16.11	0.99328
171	2.3640	142.89	1.02687
172	0.6742	13.10	0.95402
173	1.3664	25.47	1.01080
174	0.9244	21.40	0.96082
175	31.7469	1264.72	0.93212
176	3.3598	53.05	0.64494
177	1.4645	18.42	0.84037
181	0.5024	3.71	1.00144
182	1.0543	27.73	0.92350
183	7.5520	67.34	2.18269
191	0.0617	6.77	1.02297
192	4.2350	12.70	0.90263
193	3.8209	13.70	0.88112
201	36.3230	3009.15	1.00815
202	0.4514	86.06	0.73384
203	0.5686	66.51	0.99191
204	1.3435	24.04	0.97692
205	55.7434	532.41	0.93016
211	1.3028	485.66	1.17049
212	1.1060	92.60	0.96089
221	33.3073	8750.21	1.10464
222	13.7013	1297.94	0.95787
223	1.0336	10.19	0.98489
231	0.0000	0.00	0.37349
232	0.0798	44.72	0.52507
241	1.9159	4322.12	0.78583
243	0.8625	49.13	0.97893
244	22.6139	10876.60	1.73154
245	0.4358	61.79	1.05195
246	0.5669	47.70	0.78784
251	1.7089	22.08	0.89088
252	9.9664	1355.51	0.95770
261	45.7306	4602.97	1.01112
262	1.7850	44.84	0.85279
264	9.7043	5.07	0.44134
265	0.1427	25.85	1.03010
266	10.1795	1111.04	0.95568
267	18.6907	438.21	1.06637
268	1.0414	173.09	1.09268
271	0.4902	215.91	0.96160
272	2.6302	46.11	0.96060

nace3	CV_prodverd_ny	SIGMA_prodverd_ny	BETA_prodverd_ny
273	16.14	4.75	0.90179
274	0.92	1256.23	1.24402
275	0.78	78.43	0.91885
281	6.29	821.05	1.03248
282	708.82	15579.16	1.48577
283	0.00	42.23	0.81456
284	2.14	17.24	0.86083
285	72.11	4257.30	1.01445
286	0.41	31.96	0.80393
287	0.43	51.49	0.84482
291	2.66	816.39	1.10082
292	8.90	1017.96	1.03961
293	15.76	351.81	0.69829
294	0.65	14.57	0.99794
295	3.58	410.98	1.01389
296	0.32	106.53	1.02167
297	0.94	40.94	1.07946
300	0.79	26.26	0.88466
311	0.54	43.27	0.82155
312	0.36	68.81	0.98213
313	0.58	259.45	1.22319
314	3.96	39.90	0.51171
315	1.37	60.96	0.87310
316	15.89	1051.36	1.36683
321	0.30	73.63	1.00360
322	0.05	78.79	0.88345
323	0.53	53.49	0.35752
331	9.12	673.20	0.78545
332	13.23	3680.14	1.00512
333	37.15	22314.69	1.08430
334	0.50	12.95	0.91298
341	0.64	86.89	0.71844
342	1.94	101.19	0.42113
343	172.33	29647.22	1.12343
351	27.30	20171.20	0.96010
352	0.72	91.08	0.71193
353	1059.88	108436.41	1.45440
354	9.14	178.17	1.27770
355	0.93	14.09	0.85663
361	0.66	62.16	1.02221
362	0.87	10.78	0.91066
363	8.08	20.72	0.88430
364	0.44	23.09	0.90690
365	9.33	40.35	0.92241
366	8.47	199.10	1.08656
371	91.27	3137.54	1.12589
372	1.16	33.35	1.01700

De sist utgitte publikasjonene i serien Notater

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

2005/11 A-K. Mevik: Usikkerhet i ordrestatistikken. 22s.

2005/12 A. Akselsen, S. Lien, Ø. Sivertstøl: FD - Trygd. Variabelliste. 56.

2005/13 T. Seland Forgaard: Monitor for sekundærflytting. En deskriptiv analyse om sekundærflyttinger blant flyktninger som ble bosatt i Norge i perioden 1994-2003. 48s.

2005/14 O. Villund: Kvalitet på yrke i registertbasert statistikk. Resultater og utfordringer. 48s.

2005/15 E. Engelien, M. Steinnes og V.V. Holst Bloch: Tilgang til friluftsområder. Metode og resultater 2004. 38s.

2005/16 G. Dahl: Uførepensjonisters bakgrunn. 56s.

2005/17 W. Drzwi: Økonomisk-politisk kalender 1964-1999

2005/18 A. Rolland: KOSTRA, tjenestekvalitetog kompetansefordeling i supermarkedstaten. 45s.

2005/19 H. Tønnseth. Årsrapport 2004. Kontaktutvalget for helse- og sosialstatistikk 10s.

2005/20 N.K. Buskoven: Vertskommunekompensasjon - kartlegging av kommunenes utgifter til asylmottak. 49s.

2005/21 H.C. Hougen: Omnibusundersøkelsen oktober/november 2004. Dokumentasjonsrapport. 52s.

2005/22 D. Sve, L. Solheim og G. Haraldsen: Eldres kvalitet. Dokumentasjon av datafangsten. 64s.

2005/23 E. Rauan: Undersøking om foreldrebetaling i barnehagar, januar 2005. 45s.

2005/24 L. Østby: Bruk av velferdsordninger blant nyankomne innvandrere fra de nye EØS-medlemslandene. 36s.

2005/25 A. Fagereng: Reestimering av faktoretterspørselen i KVARTS. 72s.

2005/26 O. Haugen: Utrekning av vekter til inntekts og formuesundersøkingane 2000, 2001 og 2002. 56s.

2005/27 M. Bråthen, J.I. Hamre og T. Pedersen: Evaluering av ordinære arbeidsmarkedstiltak. Beskrivende analyse av deltakerne i 2002 og forslag til ny evalueringsmetode. 33s.

2005/28 M. Høstmark: Forundersøkelse om kommunale helseutgifter knyttet til bosetting av flyktninger. 48s.

2005/29 A. Vedø: Analyse av revisjon. Lønn i bygge- og anleggsvirksomhet. 43

2005/30 H.C. Hougen: Samordnet levekårsundersøkelse 2004 - tverrsnittsundersøkelsen. Dokumentasjonsrapport. 139s.

2005/31 T. Hægeland, L.J. Kirkebøen og O. Raaum: Skoleresultater 2004. En kartlegging av karakterer fra grunn- og videregående skoler i Norge. 89s.

2005/32 A.Rolland: Brukertilfredshetsmålinger i offentlig sektor. Utredning for Moderniseringsdepartementet og regjeringens handlingsplan for modernisering. 96s.

2005/33 K. Aasestad, A. Finstad o g K. Loe Hansen: Bruk av helsefarlige produkter i grafisk industri. 27s.

2005/34 S.W. Bogen, K. Digre, A. Hedum, T. Hægeland, T.K. Schjerven og B. Vold: Et system for statistikk omstatlig virksomhet. Forprosjektnotat. 44s.

2005/35 Kostra. Arbeidsgrupperapporter 2005. 230s.