Interne notater

STATISTISK SENTRALBYRÅ

IN 80/17 4. juli 1980

SAMMENSATT ESTIMERING I FORBINDELSE MED BYRÅETS ARBEIDSKRAFTUNDERSØKELSER

Av

Roy Østensen*)

INNHOLD

*) Jeg vil takke John Helgeland og Johan Heldal for bistand under programmeringsarbeidet

1. Innledning

Dette notatet er skrevet i tilknytning til et program som er skrevet for å beregne endringstall og sammensatte nivåtall i Byråets arbeidskraftundersøkelser.

En sammensatt estimator er en veiet sum av estimatorer ved flere tidspunkter. Disse estimatorene er basert dels på hele utvalget ved de aktuelle tispunkter, dels på den del av utvalget som er felles ved flere av tidspunktene. Vektene er valgt slik at den sammensatte estimator blir forventningsrett.

Formålet med notatet er å gi en redegjørelse for det teoretiske grunnlaget for beregningene. Kapittel 2 gir derfor en redegjørelse for forskjellige typer sammensatte estimatorer når de baseres på bare de to siste tidspunktene undersøkelsene er foretatt. Det blir også foretatt sammenligninger av estimatorene på grunnlag av deres varians.

Siden Byrået bruker en rulleringsplan (beskrevet kap. 2) som har med enheter i utvalget som også var med for et år siden, blir det i kapittel 3 konstruert estimatorer som tar med observasjoner både fra et kvartal siden (som i kap. 2) og for et år siden. Da beregningene er relativt omfattende, har jeg ikke her gjort en generell sammenligning av disse estimatorene med de som er funnet tidligere. En sammenligning av variansen til estimatoren i 2.2 ($\hat{p}(t)$) og estimatoren i 3.2 ($\tilde{p}(t)$) viser at estimatene blir mere presise når vi tar med estimatorene fra tidligere tidspunkter, spesielt hvis korrelasjonskoeffisienten ρ er høy.

I Byråets arbeidskraftundersøkelse har vi estimert korrelasjonen ρ til 0,86 (se kap. 5 om autokorrelasjon). For denne verdien av ρ finner vi at $\hat{p}(t)$ reduserer variansen med 32% i forhold til gjennomsnittet $\overline{Y}$ av observasjonene ved tidspunkt t, mens $\tilde{p}(t)$ reduserer variansen med 44 prosent i forhold til $\overline{Y}$. Det kan derfor være nærliggende å se på estimatorer som tar med hele forhistorien, men da kommer vi over i tidsrekkeanalyse, noe som faller utenfor dette prosjektet.

I kapittel 4 går vi så over til å se på estimatorer for endringer i sysselsettingen fra et tidspunkt til et annet. Det blir vist at man får bedre estimatorer ved ikke å basere seg på de sammensatte nivåtallestimatorene. Resultatet er at de optimale estimatorene vil gi andre resultater for endringene enn det estimatorer som bruker estimatorene for nivåtallene, vil gi. I tabeller som publiseres kan dette være uheldig. De optimale estimatorene vil vel derfor først og fremst være et analyseverktøy.

Siden de fremkomne estimatorene baserer seg på korrelasjonene mellom observasjoner som er tatt til forskjellige tidspunkter, gjør kap. 5 kort rede for estimering av dem. Det er også tatt med to tabeller over korrelasjonene i de enkelte næringer, basert på tallene fra 1. kvartal 1975 og til 2. kvartal 1979. Disse tallene stemmer forøvrig godt overens med de tallene Dagsvik (1975) gir i tabell 1 og 2 i sitt notat.

Siste kapittel gjør så kort rede for bruken av selve programmet. Det er skrevet med tanke på at man skal kunne bruke det direkte fra terminal uten å måtte søke hjelp ved Systemkontoret i de enkelte tilfeller. Programmet finnes i to versjoner. Den ene versjonen ber om dataene etterhvert som maskinen trenger dem under kjøringen. Den andre versjonen leser dataene direkte fra file, men er forøvrig identisk med den første versjonen.

Den delen av programmet som beregner sammensatte nivåtall, beregner forøvrig både nivåtall basert på to tidspunkter (inneværende og forrige kvartal), og dessuten nivåtall hvor tallene for et år siden også er tatt med i beregningen.

I forbindelse med variansberegningene er det blitt antatt at hele landet er et enkelt utvalgsområde. Dette vil ha en del innflydelse på de resultatene som fremkommer, men det er grunn til å tro at denne innflydelsen er redusert, i og med at estimatorene baserer seg på oppblåste tall fra undersøkelsene. Som en rent kvalitativ sammenligning skulle dette i alle fall være godt egnet.

2. Sammensatt estimering på grunnlag av to tidspunkter

Anta det trekkes et utvalg på n enheter ved hvert tidspunkt trekkingen foregår. La observasjonene ved forrige tidspunkt være gitt ved $X_1, X_2, \ldots, X_n$, og ved inneværende tidspunkt er de $Y_1, Y_2, \ldots, Y_n$. Anta trekningen foregår med tilbakelegging, slik at vi for individene i og j har $\mathrm{Cov}(X_i, X_j) = 0$, $\mathrm{Cov}(X_i, Y_j) = 0$ og $\mathrm{Cov}(Y_i, Y_j) = 0$ når $i \neq j$.

La nå $\rho = \rho\,(X_i, Y_i)$ = korrelasjonskoeffisienten mellom observasjonene for samme person fra forrige og nåværende tidspunkt, mens $\sigma^2 = \mathrm{var}\,(Y_i)$. Vi antar videre det er realistisk at ρ og σ^2 regnes konstant over tid.

Trekningen av enhetene foregår på følgende måte:

$n\lambda$ enheter går igjen fra et tidspunkt til det neste, mens de resterende $n\mu$ enhetene blir byttet ut ($\mu = 1 - \lambda$). La $\overline{X}'$ og $\overline{Y}'$ være gjennomsnittet av de enhetene som beholdes for henholdsvis forrige og nåværende tidspunkt, mens $\overline{X}''$ er gjennomsnittet av de enhetene ved forrige tidspunkt som ble byttet ut. Tilsvarende er $\overline{Y}''$ gjennomsnittet av de enhetene som er nye ved nåværende tidspunkt. Hvis de totale gjennomsnitt skrives som henholdsvis $\overline{X}$ og $\overline{Y}$, vil dermed disse kunne uttrykkes som følger:

$$\overline{Y} = \lambda\,\overline{Y}' + \mu\,\overline{Y}'' \tag{2.1}$$

$$\overline{X} = \lambda\,\overline{X}' + \mu\,\overline{X}''$$

Variansen til de enkelte puljenes gjennomsnitt $(\overline{X}', \overline{X}'', \overline{Y}', \overline{Y}'')$ er gitt ved

$$\mathrm{Var}\,(\overline{X}') = \mathrm{Var}\,(\overline{Y}') = \sigma^2/n\lambda \tag{2.2}$$

$$\mathrm{Var}\,(\overline{X}'') = \mathrm{Var}\,(\overline{Y}'') = \sigma^2/n\mu$$

Vi ønsker å estimere $E(Y_i) = p$. Generelt kan p estimeres ved det totale gjennomsnitt $\overline{Y}$, men ved å benytte at hver enkelt observasjon er korrelert med tidligere observasjoner, kan vi få redusert variansen. Dette gir opphavet til sammensatte estimatorer. To typer estimatorer peker seg ut:

2.1 Minimum varians forventningsrett estimator (MVU-estimator) på grunnlag av de to siste tidspunktene

Den første estimatoren er en MVU-estimator på grunnlag av de to siste tidspunktene. Den kan generelt skrives som

$$\hat{p}_2 = a(\overline{X}'' - \overline{X}') + c\overline{Y}' + (1-c)\overline{Y}'' \tag{2.3}$$

som generelt vil være forventningsrett. Minimering av variansen gir konstantene lik:

$$a = \lambda\mu\rho/(1-\rho^2\mu^2) \tag{2.4}$$

$$c = \lambda/(1-\rho^2\mu^2)$$

slik at $\hat{p}_2$ kan skrives som

$$\hat{p}_2 = [\lambda\mu\rho(\overline{X}'' - \overline{X}') + \lambda\overline{Y}' + \mu(1-\rho^2\mu)\overline{Y}'']/(1-\rho^2\mu^2) \tag{2.5}$$

Variansen er gitt ved

$$\mathrm{Var}(\hat{p}_2) = (\sigma^2/n)(1-\rho^2\mu)/(1-\rho^2\mu^2) \tag{2.6}$$

2.2 Sammensatt MVU-estimator på grunnlag av flere tidspunkter

Når vi først har kommet på tanken å ta med tidligere tidspunkter ved estimeringen av nivået p, er det naturlig å ta med den sammensatte estimatoren fra forrige tidspunkt. Dermed vil vi indirekte dra nytte av resultatene fra tidligere tidspunkter.

I dette tilfellet vil vi bruke en litt annen notasjon enn foran. $\overline{Y}'(t)$ og $\overline{Y}''(t)$ referer til gjennomsnittene ved tidspunkt t, altså ved inneværende tidspunkt. Tilsvarende refererer $\overline{X}'(t-s)$ og $\overline{X}''(t-s)$ til gjennomsnittene ved tidspunkt (t-s). Forøvrig har ' og "" samme betydning som tidligere. I tillegg vil vi nå basere oss på den sammensatte estimatoren fra forrige tidspunkt. La derfor $\hat{p}(t-s)$ betegne den sammensatte estimatoren fra tidspunkt (t-s). Des Raj (1968) viser (pkt. 1.11 s. 17-18) at en nødvendig og tilstrekkelig betingelse for at en estimator er MVU, er at den er ukorrelert med enhver null-funksjon. I pkt. 7.11 (s. 160-162) bruker han dette til å vise at den sammensatte estimatoren får formen

$$\hat{p}(t) = \phi(t)\overline{Y}''(t) + [1-\phi(t)][\overline{Y}'(t) + \rho(\hat{p}(t-1) - \overline{X}'(t-1))] \tag{2.7}$$

hvor $\phi(t)$ er en konstant som avhenger av tiden t.

Siden $\hat{p}(t)$ er en MVU-estimator ved tidspunkt t, er ifølge ovennevnte lemma i Des Raj

$$\mathrm{Cov}[\hat{p}(t) - \overline{Y}''(t),\ \hat{p}(t)] = 0 \tag{2.8}$$

slik at

$$\mathrm{Var}(\hat{p}(t)) = \mathrm{Cov}[\overline{Y}''(t),\ \hat{p}(t)] \tag{2.9}$$

Siden $\overline{Y}''(t)$ er gjennomsnittet av de nye enhetene ved tidspunkt t, er $\overline{Y}''(t)$ ukorrelert med alle ledd i uttrykket for $\hat{p}(t)$ unntatt seg selv. Følgelig er

$$\mathrm{Var}(\hat{p}(t)) = \phi(t)\mathrm{Var}(\overline{Y}''(t)) = \phi(t)\sigma^2/(n\mu) \tag{2.10}$$

For å finne variansen, og dermed konstanten ϕ, setter vi

$$\mathrm{Var}[\hat{p}(t)] = (\sigma^2/n)\ G(t) \tag{2.11}$$

hvor G(1) = 1.

Det er aktuelt å sammenligne hvor stor gevinst vi får ved å bruke $\hat{p}(t)$ istedenfor $\hat{p}_2$. John Dagsvik (IO 75/24, s.6) viser at vi har en viss gevinst, men at forskjellene mellom $\hat{p}(t)$ og $\hat{p}_2$ er små, unntatt når korrelasjonene er store.

I Byråets arbeidskraftundersøkelse (AKU) bruker man et slikt rotasjonssystem at utvalget er delt i fire puljer. Hver pulje er med i alt fire ganger, slik at halve utvalget var med for et år siden og halve var med for et kvartal siden, på følgende måte:

Kvartal	1	2	3	4	1	2
Pulje 1	x					
2	x	x				
3		x	x			
4			x	x		
5	x			x	x	
6	x	x			x	x
7		x	x			x
8			x	x		
9				x	x	
10					x	x
11						x

For Byrået blir derfor de aktuelle verdiene på λ og μ lik 0,5. Den eksakte verdi for G(t) finnes av uttrykket (se Des Raj (7.61)).

$$1/G(t) = 0{,}5 + \left[\rho^2\ G(t-1) + 2\ (1-\rho^2)\right]^{-1} \tag{2.12}$$

idet vi husker at G(1) = 1. Da får vi følgende tabell for G(t) over tid: (Merk at t=2 gir G(2), dvs. verdien til Var($\hat{p}_2$)).

Tabell 1. Reduksjon av Var($\hat{p}$(t)) over tid for forskjellige verdier av ρ

t =	1	2	3	5	10	∞
ρ = 0,5	1	0,9333	0,9286	0,9282	0,9282	0,9282
0,8	1	0,8095	0,7647	0,7509	0,7500	0,7500
0,86	1	0,7732	0,7066	0,6790	0,6758	0,6758
0,9	1	0,7461	0,6597	0,6151	0,6072	0,6071
0,95	1	0,7086	0,5888	0,5052	0,4770	0,4759

Vi merker oss at G(t) relativt raskt nærmer seg sin grenseverdi, som vi vil kalle G, iallfall så lenge ρ<0,9. Grenseverdien finner vi ved å sette G(t-1) = G(t) = G i (2.12). Da får vi

$$\lim_{t \to \infty} G(t) = G = 2[\sqrt{1-\rho^2} - (1-\rho^2)]/\rho^2 \tag{2.13}$$

Herav finner vi $\phi = \mu G$.

Siden konvergensen går relativt raskt, skulle vi kunne bruke verdien G for G(t) når vi søker ϕ(t). Den feilen vi gjør ved å benytte grenseverdien ϕ=μG som en tilnærmelse istedenfor den optimale verdien ϕ(t), skulle derfor raskt gå mot 0.

La oss nå sammenlikne $\hat{p}$(t) med $\hat{p}_2$ ved effisiensen e($\hat{p}(t)/\hat{p}_2$). I grensen er den gitt som G/G(2). Tabell 2 gir en oversikt over denne effisiensen for forskjellige verdier av ρ.

Tabell 2. Effisiensen e($\hat{p}(t)/\hat{p}_2$) for forskjellige verdier av ρ

ρ	0,5	0,6	0,7	0,8	0,86	0,9	0,95
G/G(2)	0,9945	0,9865	0,9684	0,9265	0,8740	0,8137	0,6716

I programmet for beregning av sammensatte nivåtall og endringstall har jeg regnet ρ lik 0,86, noe som gir $\hat{p}(t)$ lik:

$$\hat{p}(t) = 0{,}34\overline{Y}''(t) + 0{,}66\{\overline{Y}'(t) + 0{,}86[\hat{p}(t-1) - \overline{X}'(t-1)]\} \qquad (2.14)$$

3. Sammensatt estimering på grunnlag av tallene for forrige kvartal og et år siden

La $\hat{d}(t-s,t)$ være en estimator for endringen fra tidspunkt (t-s) til t. Vi ønsker da å se på estimatorer på formen

$$\tilde{p}(t) = a[\hat{p}(t-4) + \hat{d}(t-4,t)] + b[\hat{p}(t-1) + \hat{d}(t-1,t)] + (1-a-b)\hat{p}(t) \qquad (3.1)$$

hvor a og b er konstanter som ligger i intervallet [0,1], og hvor også summen a+b ligger i det samme intervallet.

La videre $\rho_s = \mathrm{Cov}(X_i(t-s), X_i(t))$,s=1,2,... Vi ønsker spesielt å se på to valg av $\hat{p}(t-s)$.

3.1 Sammensatt estimator på grunnlag av gjennomsnittene på tidspunktene (1-4), (I-1) og I

Generelt er andelen i hver gruppe som utvalget er delt opp i, lik $\lambda(I,t)$ for gruppe I ved tidspunkt t. Siden Byrået bruker $\lambda(I,t) = 1/4$, vil også vi bruke den verdien videre for å forenkle uttrykkene. Vi setter gjennomsnittet for gruppe I ved tidspunkt t lik $\overline{Y}(I,t)$ og gjennomsnittet for gruppe I ved tidspunkt (t-s) lik $\overline{X}(I,t-s)$, hvor I betegner for hvilken gang gruppen er med. ($\overline{X}(2,t-s)$ betegner f.eks. gjennomsnittet i gruppen som er med for 2. gang ved tidspunkt (t-s)). Dermed blir $\hat{p}(t) = \overline{Y}(t)$, $\hat{p}(t-1) = \overline{X}(t-1)$ og $\hat{p}(t-4) = \overline{X}(t-4)$ de respektive gjennomsnitt ved tidspunktene t, t-1 og t-4. La så

$$\hat{d}(t-1,t) = \{[\overline{Y}(2,t)-\overline{X}(1,t-1)] + [\overline{Y}(4,t)-\overline{X}(3,t-1)]\}/2$$

$$\hat{d}(t-4,t) = \{[\overline{Y}(3,t)-\overline{X}(1,t-4)] + [(\overline{Y}(4,t)-\overline{X}(2,t-4)]\}/2 \qquad (3.2)$$

Da blir estimatoren seende ut som følger:

$$\tilde{p}(t) = \{a[-\bar{X}(1,t-4) - \bar{X}(2,t-4) + \bar{X}(3,t-4) + \bar{X}(4,t-4)] \\ + b[-\bar{X}(1,t-1) + \bar{X}(2,t-1) - \bar{X}(3,t-1) + \bar{X}(4,t-1)] \\ + (1-a-b)\bar{Y}(1,t) + (1-a+b)\bar{Y}(2,t) + (1+a-b)\bar{Y}(3,t) \\ + (1+a+b)\bar{Y}(4,t)\}/4 \tag{3.3}$$

Vi ønsker å finne konstanter a og b slik at variansen blir minimert. Vi finner:

$$\mathrm{Var}(\tilde{p}(t)) = (\sigma^2/n)\ \{1-\rho_4 a-\rho_1 b+(2-\rho_4)a^2+(2-\rho_1)b^2 \\ +\rho_3 ab/2\} \tag{3.4}$$

Vi finner a og b lik:

$$a = 2\{4[2-\rho_1]\rho_4-\rho_1\rho_3\}/\{16[2-\rho_1][2-\rho_4]-\rho_3^2\} \tag{3.5}$$

$$b = 2\{4[2-\rho_4]\rho_1-\rho_3\rho_4\}/\{16[2-\rho_1][2-\rho_4]-\rho_3^2\}$$

La nå ρ_1 = 0,86, ρ_3 = 0,79 og ρ_4 = 0,77. Disse verdiene synes å passe i arbeidskraftundersøkelsen (AKU) (se avsnitt 5). Da finner vi a = 0,26 og b = 0,332. Fordi korrelasjonskoeffisientene vil være tilnærmede verdier, skulle vi derfor kunne sette a = 1/4 og b = 1/3 uten større feil. I så tilfelle finner vi

$$\mathrm{Var}(\tilde{p}(t)) = 0{,}7573\ \sigma^2/n \tag{3.6}$$

I forhold til råestimatet $\bar{Y}(t)$ reduseres altså variansen med 24%. Til sammenligning er variansen til $\hat{p}_2$ ifølge (2.6) lik 0,7732 σ^2/n. Vi har derfor svært lite å vinne på å ta med de enkelte gruppegjennomsnittene $\bar{X}(1,t-4),\ldots,\bar{X}(4,t-4)$ fra et år tilbake.

De to estimatorene $\hat{p}_2$ fra (2.5) og $\tilde{p}(t)$ fra (3.4) baserer seg kun på gruppegjennomsnittene fra de forskjellige tidspunktene. Vi finner at den sammensatte estimatoren $\hat{p}(t)$ fra (2.14) som også inkluderer den sammensatte estimatoren fra tidspunkt (t-1) har en varians lik 0,6758 σ^2/n når $\rho = 0{,}86$. Dette er en variansreduksjon i forhold til $\hat{p}_2$ på 13%. Det er innlysende at vi vil få variansen redusert enda litt til hvis vi i tillegg inkluderer den sammensatte estimatoren fra tidspunkt (t-4). Men den sammenligningen vi foretok foran mellom variansene til $\hat{p}_2$ og $\tilde{p}(t)$ kan antyde at den ekstra variansereduksjonen vi får i forhold til $\hat{p}(t)$ blir liten. Vi vil se nærmere på dette i neste avsnitt.

3.2 Sammensatt estimator på grunnlag av de sammensatte estimatorene ved tidspunkt (t-4) og (t-1)

La $\tilde{p}(t-s)$ betegne den sammensatte estimator på tidspunkt (t-s). Nå vil vi betrakte estimatorer på formen

$$\tilde{p}(t) = a(t)[\tilde{p}(t-4) + \hat{d}(t-4)] + b(t)[\tilde{p}(t-1) + \hat{d}(t-1,t)] + [1-a(t)-b(t)]\overline{Y}(t) \tag{3.7}$$

hvor a(t) og b(t) er konstanter som avhenger av tiden t. De er gitt slik at deres verdier, samt summen a(t)+b(t) ligger i intervallet [0,1]. $\hat{d}(t-1,t)$ og $\hat{d}(t-4,t)$ er de samme som før.

Da beregningene av a(t) og b(t) er meget uoversiktlig i det generelle tilfellet, gjør vi visse forenklinger. Når man bare benytter den sammensatte estimatoren fra tidspunkt (t-1) i beregningene, viser det seg at vekten $\phi(t)$ (se avsnitt 2) hurtig nærmer seg sin asymptotiske verdi når t vokser. Det er naturlig å tro at dette er tilfelle også når vi tar med den sammensatte estimator fra tidspunkt (t-4). Vi setter derfor $\text{Var}[\tilde{p}(t)] = (\sigma^2/n)G$, $a(t) = a$ og $b(t) = b$.

Videre vil vi fortsatt sette puljestørrelsene like, dvs. lik n/4. Da får estimatoren utseende:

$$\begin{aligned}\tilde{p}(t) = {} & a\{\tilde{P}(t-4) - [\overline{X}(1,t-4) + \overline{X}(2,t-4)]/2\} + b\{\tilde{p}(t-1) - \\ & - [\overline{X}(1,t-1) + \overline{X}(3,t-1)]/2\} + \{(1-a-b)\overline{Y}(1,t) + \\ & + (1-a+b)\overline{Y}(2,t) + (1+a-b)\overline{Y}(3,t) + (1+a+b)\overline{Y}(4,t)\}/4\end{aligned} \tag{3.8}$$

For å forenkle beregningen av a og b, setter vi

$$\tilde{p}(t) = a\tilde{p}(t-4) + b\tilde{p}(t-1) + R(t) \tag{3.9}$$

hvor

$$R(t) = -\{a[\overline{X}(1,t-4) + \overline{X}(2,t-4)] + b[\overline{X}(1,t-1) + \overline{X}(3,t-1)]\}/2 + \{(1-a-b)\overline{Y}(1,t) + (1-a+b)\overline{Y}(2,t) + ((1+a-b)\overline{Y}(3,t) + (1+a+b)\overline{Y}(4,t)\}/4 \qquad (3.10)$$

La nå

$$\gamma(s) = \mathrm{Cov}[\tilde{p}(t-s), \tilde{p}(t)] \qquad (3.11)$$

P.g.a. symmetri er $\gamma(-s) = \gamma(s)$. La videre

$$\tilde{\gamma}(s) = \mathrm{Cov}[R(t), \tilde{p}(t-s)]$$

$$\kappa(s) = \mathrm{Cov}[R(t), R(t-s)] \qquad (3.12)$$

Da får vi

$$\gamma(s) = a\gamma(s-4) + b\gamma(s-1) + \tilde{\gamma}(s) \qquad (3.13)$$

$$\tilde{\gamma}(s) = a\tilde{\gamma}(s+4) \; b\tilde{\gamma}(s+1) + \kappa(s)$$

Merk forøvrig at $\tilde{\gamma}(s) = 0$ for $s \geq 6$.

Dette gir oss følgende relasjoner:

$$\gamma(0) = a\gamma(4) + b\gamma(1) + \tilde{\gamma}(0)$$

$$\gamma(1) = a\gamma(3) + b\gamma(0) + \tilde{\gamma}(1)$$

$$\gamma(2) = a\gamma(2) + b\gamma(1) + \tilde{\gamma}(2) \qquad (3.14)$$

$$\gamma(3) = a\gamma(1) + b\gamma(2) + \tilde{\gamma}(3)$$

$$\gamma(4) = a\gamma(0) + b\gamma(3) + \tilde{\gamma}(4)$$

Videre skal $\tilde{\gamma}$ (s) oppfylle følgende relasjoner for s lik eller mindre enn 6:

$$\tilde{\gamma}(6) = 0$$

$$\tilde{\gamma}(5) = \kappa(5)$$

$$\tilde{\gamma}(4) = b\gamma(5) + \kappa(4)$$

$$= b\kappa(5) + \kappa(4)$$

$$
\begin{aligned}
\tilde{\gamma}(3) &= b\tilde{\gamma}(4) + \kappa(3) \\
&= b^2\kappa(5) + b\kappa(4) + \kappa(3) \\
\tilde{\gamma}(2) &= b\tilde{\gamma}(3) + \kappa(2) \\
&= b^3\kappa(5) + b^2\kappa(4) + b\kappa(3) + \kappa(2) \\
\tilde{\gamma}(1) &= a\tilde{\gamma}(5) + b\tilde{\gamma}(2) + \kappa(1) \\
&= (a+b^4)\kappa(5) + b^3\kappa(4) + b^2\kappa(3) + b\kappa(2) + \kappa(1) \\
\tilde{\gamma}(0) &= a\tilde{\gamma}(4) + b\tilde{\gamma}(1) + \kappa(0) \\
&= b(2a+b^4)\kappa(5) + (a+b^4)\kappa(4) + b^3\kappa(3) + b^2\kappa(2) + b\kappa(1) + \kappa(0)
\end{aligned}
\tag{3.15}
$$

Vi trenger ligningene ovenfor for å kunne finne de verdiene av konstantene a og b som minimerer $\mathrm{Var}[\tilde{p}(t)] = \gamma(0)$. Men først må vi uttrykke $\kappa(s)$ ved $\rho_1, \ldots \rho_5$ og σ^2. Ut fra uttrykket for R(t) finner vi følgende uttrykk for $\kappa(s)$:

$$
\begin{aligned}
\kappa(0) &= (\sigma^2/n)\{1+3\,a^2 + 3\,b^2 - 2\,b\,(1+b)\,\rho_1 + 2\,a\,b\,\rho_3 - 2\,a\,(1+a)\rho_4\} \\
\kappa(1) &= (\sigma^2/n)\{-a(1-a)-b(1-b) + (1+a^2+b^2)\rho_1 - 4\,a\,(1+a-b)\rho_3 + \\
&\quad + 2\,a\,b\,\rho_4 - a(1+a+b)\rho_5\}/8 \\
\kappa(2) &= 0 \\
\kappa(3) &= (\sigma^2/n)\{4\,a\,b - 2\,a\,(1-a+b)\rho_1 - 2\,b\,(1+a-b)\rho_4 + (1-(a-b)^2)\rho_3\}/16 \\
\kappa(4) &= (\sigma^2/n)\{-4\,a\,(1-a) + 4\,a\,b\,\rho_1 - 2\,b(1-a+b)\rho_3 + 2(1+a^2+9\,b^2)\rho_4 \\
&\quad - 2\,b(1+a+b)\rho_5\}/16 \\
\kappa(5) &= (\sigma^2/n)(1-a-b)\{-2\,a\,\rho_1 - 2\,b\,\rho_4 + (1+a+b)\rho_5\}/16 \\
\kappa(6) &= 0
\end{aligned}
\tag{3.15}
$$

Selve minimeringen ble gjort numerisk på datamaskinen. Følgende verdier ble brukt på korrelasjonskoeffisienten:

$$\rho_1 = 0{,}86,\ \rho_3 = 0{,}79,\ \rho_4 = 0{,}77,\ \rho_5 = 0{,}76.$$

Det gav

$$a = 0{,}2348,\ b = 0{,}3905$$

Dette gav en minimumsverdi på variansen lik

$$\mathrm{Var}[\tilde{p}(t)] = 0{,}5643\ (\sigma^2/n) \tag{3.17}$$

Som nevnt i siste avsnitt av 3.1 hadde den tilsvarende estimatoren fra 2.2 en varians på 0,68 σ^2/n. Ved å ta med den sammensatte estimatoren fra tidspunkt (t-4), får vi dermed en ekstra variansreduksjon på 16%. Ut fra det vi kom frem til i 3.1 er dette en overraskende høy gevinst.

4. Estimering av endringstall

Vi ønsker nå å se nærmere på noen estimatorer for endringstallene. La parameteren av interesse, endringstallet fra tidspunkt s til t bli betegnet med $\hat{d}(s,t)$. Så lenge publiserte tabeller er basert på gjennomsnittet $\overline{Y}(t)$ som estimator for endringstallene, bør man vel helst bruke $d(s,t) = \overline{Y}(t) - \overline{X}(s)$ som estimator for endringstallene hvis man skal publisere tall for endringer. Denne estimatoren har en varians lik $2(1-\lambda\rho)\,\sigma^2/n$ når $s=t-1$. For $\lambda=\frac{1}{2}$ blir dette $(2-\rho)\,\sigma^2/n$. Vi kan merke oss at vi får minimum varians om vi bruker paneldata, altså at vi benytter samme utvalg fra gang til gang.

Siden man er interessert i både nivåtall og endringstall, er ikke dette praktisk. Vi kan likevel oppnå en betydelig variansreduksjon for estimatoren til endringstallene. De to neste avsnittene gjør nærmere rede for dette.

4.1 Estimatet basert på estimatorene for nivåtallene

La $\hat{p}(t)$ betegne den sammensatte estimatoren fra avsnitt 2.2 og $\tilde{p}(t)$ den sammensatte estimatoren fra avsnitt 3.2. Hovedforskjellen på dem er at $\hat{p}(t)$ tar med bare den sammensatte estimatoren $\hat{p}(t-1)$ fra forrige tidspunkt, mens $\tilde{p}(t)$ tar med de sammensatte estimatorene $\tilde{p}(t-1)$ og $\tilde{p}(t-4)$. Se forøvrig avsnitt 2.2 og 3.2 for en nærmere redegjørelse av disse estimatorene.

Slik det ble vist i kapittel 2 og 3, gir $\hat{p}(t)$ og $\tilde{p}(t)$ en ganske betydelig variansreduksjon, spesielt når korrelasjonen ρ er høy. Hvis man derfor publiserer tabeller over nivåtall på grunnlag av de sammensatte estimatorer, vil man samtidig oppnå publiserbare tall over endringer som har en betydelig reduksjon av variansen.

Siden $\tilde{p}(t)$ har den minste variansen av de estimatorene for nivåtall som vi har sett på, bør man i tilfelle basere estimatoren for endringer på denne. Men dette forutsetter som nevnt foran at man samtidig bruker $\tilde{p}(t)$ som estimator for de nivåtallene man publiserer. Nå er denne estimatoren vanskelig å studere rent teoretisk. Vi kan riktignok studere estimatorens egenskaper ved simuleringsforsøk, men det har jeg ikke gjort. Isteden vil vi derfor se på estimatoren for endringstall når vi bruker $\hat{p}(t)$, siden den har enklere egenskaper. Den vil da få følgende utseende:

$$\hat{d}(s,t) = \hat{p}(t) - \hat{p}(s) \tag{4.1}$$

Sett $\mathrm{Var}\ d(s,t) = (\sigma^2/n)\ H(t,t-s)$

For $s = t-1$ får vi spesielt

$$\begin{aligned}\hat{d}(t-1,t) &= \hat{p}(t) - \hat{p}(t-1)\\ &= \phi(t)\overline{Y}''(t) + [1-\phi(t)][\overline{Y}'(t)-\rho_1\overline{X}'(t-1)]\\ &= \{1-\rho_1[1-\phi(t)]\}\hat{p}(t-1)\end{aligned}$$

La som tidligere $\mathrm{Var}[\hat{p}(t)] \doteq G(t)\,\sigma^2/n$. Da blir

$$\begin{aligned}\mathrm{Var}[\hat{d}(t-1,t)] &\doteq (\sigma^2/n)\{G(t) + G(t-1)[1-2\rho_1(1-\phi(t))]\} \qquad (4.3)\\ &= (\sigma^2/n)\,H(t,1)\end{aligned}$$

hvor $G(t)$ finnes av (2.13) når $\mu = \lambda = 1/2$. Grenseverdien for $H(t,1)$ finner vi ved å sette $G(t) = G(t-1) = G$. Da finner vi:

$$\lim_{t\to\infty} H(t,1) = H(1) = 2G[1-\rho_1(1-\phi)] \tag{4.4}$$

hvor vi finner G og ϕ fra (2.13). For forskjellige verdier av t og ρ_1 finner vi $H(t,1)$ lik

Tabell 3. Den normerte varians for $\hat{d}(t-1)$ over tid for forskjellige verdier av ρ

t =		3	5	H(1)
ρ =	0,25	1,72	1,72	1,718
	0,50	1,36	1,36	1,359
	0,70	1,00	0,99	0,986
	0,80	0,77	0,75	0,750
	0,86	0,62	0,59	0,582
	0,90	0,51	0,46	0,453
	0,95	0,35	0,28	0,263

Legg til sammenligning merke til at når $\lambda = \mu = 1/2$, har estimatoren $\hat{d}(t-1,t) = \overline{Y}(t) - \overline{X}(t-1)$ en varians lik $(2-\rho)\sigma^2/n$, som er større enn σ^2/n for alle verdier av ρ. For store ρ-verdier blir følgelig gevinsten betydelig hvis vi isteden bruker (4.2).

La oss se på tilfellet $s = t-4$. Da får estimatoren utseende

$$\hat{d}(t-4,t) = \hat{p}(t) - \hat{p}(t-4) \tag{4.5}$$

For å finne variansen til $\hat{d}(t-4,t)$, trenger vi $\mathrm{Cov}(\hat{p}(t), \hat{p}(t-4))$. For å lette beregningen av denne kovariansen, vil vi går frem som i avsnitt 3.2. Vi antar vektene har nådd sine asymptotiske verdier, slik at vi kan skrive

$$\hat{p}(t) = \rho(1-\phi)\hat{p}(t-1) + R(t) \tag{4.6}$$

hvor

$$R(t) = \phi\overline{Y}(1,t) + (1-\phi)\overline{Y}(2,t) + \phi\overline{Y}(3,t) + (1-\phi)\overline{Y}(4,t) - \rho(1-\phi)\,[\overline{X}(1,t-1) + \overline{X}(3,t-1)] \tag{4.7}$$

La så

$$\begin{aligned} \gamma(s) &= \mathrm{Cov}(\hat{p}(t-s), \hat{p}(t)) \\ \tilde{\gamma}(s) &= \mathrm{Cov}(R(t), \hat{p}(t-s)) \\ \kappa(s) &= \mathrm{Cov}(R(t), R(t-s)) \end{aligned} \tag{4.8}$$

Ved å kombinere (4.6) og (4.8) ser vi at

$$\gamma(s) = \gamma_1(1-\phi)\rho(s-1) + \tilde{\gamma}(s) \tag{4.9}$$

P.g.a. symmetri er $\gamma(-s) = \gamma(s)$. La oss så sette $(\sigma^2/n) = 1$ for å gjøre uttrykkene noe penere. Idet vi husker at

$$\gamma(0) = \text{Var } \hat{p}(t) = G \tag{4.10}$$

(se (2.13)),

finner vi dermed

$$\begin{aligned}
\gamma(1) &= \rho_1(1-\phi)G + \tilde{\gamma}(1) \\
\gamma(2) &= \rho_1^2(1-\phi)^2G + \rho_1(1-\phi)\tilde{\gamma}(1) + \tilde{\gamma}(2) \\
\gamma(3) &= \rho_1^3(1-\phi)^3G + \rho_1^2(1-\phi)^2\tilde{\gamma}(1) + \rho_1(1-\phi)\tilde{\gamma}(2) + \tilde{\gamma}(3) \\
\gamma(4) &= \rho_1^4(1-\phi)^4G + \rho_1^3(1-\phi)^3\tilde{\gamma}(1) + \rho_1^2(1-\phi)^2\tilde{\gamma}(2) + \\
&\quad + \rho_1(1-\phi)\tilde{\gamma}(3) + \tilde{\gamma}(4)
\end{aligned} \tag{4.12}$$

Fra (4.6) og (4.8) ser vi så at

$$\tilde{\gamma}(s) = \rho_1(1-\phi)\tilde{\gamma}(s+1) + \kappa(s) \tag{4.13}$$

Vi ser fra (4.7) at $\tilde{\gamma}(s) = \kappa(s) = 0$ når $s \geq 6$. Fra (4.13) får vi derfor

$$\begin{aligned}
\tilde{\gamma}(5) &= \kappa(5) \\
\tilde{\gamma}(4) &= \rho_1(1-\phi)\kappa(5) + \kappa(4) \\
\tilde{\gamma}(3) &= \rho_1^2(1-\phi)^2\kappa(5) + \rho_1(1-\phi)\kappa(4) + \kappa(3) \\
\tilde{\gamma}(2) &= \rho_1^3(1-\phi)^3\kappa(5) + \rho_1^2(1-\phi)^2\kappa(4) + \rho_1(1-\phi)\kappa(3) + \kappa(2) \\
\tilde{\gamma}(1) &= \rho_1^4(1-\phi)^4\kappa(5) + \rho_1^3(1-\phi)^3\kappa(4) + \rho_1^2(1-\phi)^2\kappa(3) + \\
&\quad + \rho_1(1-\phi)\kappa(2) + \kappa(1)
\end{aligned} \tag{4.14}$$

Vi kunne tatt med $\tilde{\gamma}(0)$ også, men vi er primært interessert i $\gamma(4) = \text{Cov}(\hat{p}(t-4), \hat{p}(t))$, og der inngår ikke $\tilde{\gamma}(0)$ annet enn indirekte via G.

Av (4.7) finner vi nå uttrykkene for de $\kappa(s)$ vi trenger (jeg utelater $\kappa(0)$, siden vi ikke har behov for å kjenne den):

$$
\begin{aligned}
\kappa(1) &= 0 \\
\kappa(2) &= 0 \\
\kappa(3) &= (\rho_3-\rho_1\rho_4)\phi(1-\phi) \\
\kappa(4) &= \rho_4\phi^2 + [\rho_4(1+\rho_1^2) - \rho_1(\rho_3+\rho_5)](1-\phi)^2 \\
\kappa(5) &= (\rho_5-\rho_1\rho_4)\phi(1-\phi)
\end{aligned}
\tag{4.15}
$$

Setter vi uttrykkene fra (4.14) inn i uttrykket for $\gamma(4)$ fra (4.12) og kombinerer det med (4.15) får vi:

$$
\begin{aligned}
\gamma(4) &= \rho_1^4(1-\phi)^4G + \{[1-\rho_1^7(1-\phi)^7][\rho_1(1-\phi)\kappa(5) + \kappa(4)] + \\
&\quad [1-\rho_1^5(1-\phi)^5]\kappa(3)\}/[1-\rho_1(1-\phi)] \\
&= \rho_1^4(1-\phi)^4G + \{[1-\rho_1^7(1-\phi)^7]\Big(\rho_4\phi^2 + (1-\phi)^2 \\
&\quad [(\rho_4-\rho_1\rho_3) - (\rho_5-\rho_1\rho_4)\rho_1(1-\phi)]\Big) + [1-\rho_1^5(1-\phi)^5] \\
&\quad (\rho_3-\rho_1\rho_4)\phi(1-\phi)\}/[1-\rho_1(1-\phi)]
\end{aligned}
\tag{4.16}
$$

Setter vi $\mu = \frac{1}{2}$, finner vi fra (2.13) at

$$
\phi = [\sqrt{1-\rho_1^2} - (1-\rho_1^2)]/\rho_1^2 \tag{4.17}
$$

$$
G = 2\phi
$$

Var $\hat{d}(t-4,t)$ finner vi dermed lik

$$
\begin{aligned}
\text{Var } \hat{d}(t-4,t) &\doteq 2(G-\gamma(4)) \\
&= H(4)
\end{aligned}
\tag{4.18}
$$

Tabell 4 viser hvordan H(4) forandrer seg med forskjellige verdier av korrelasjonskoeffisientene. Jeg har brukt modellen

$$\rho_r = \rho^{2,5\sqrt{r}} \tag{4.19}$$

hvor r = t-s. Dette er den modellen jeg fant var mest realistisk i AKU (se kapittel 6).

Tabell 4. Den normerte varians H(4) for $\hat{d}(t-4,t)$ under modellen $\rho_r = \rho^{2,5\sqrt{r}}$ for noen verdier av ρ

ρ	$\rho(3)$	$\rho(4)$	$\rho(5)$	H(4)
0,25	0,116	0,090	0,071	1,832
0,50	0,341	0,299	0,267	1,465
0,70	0,575	0,537	0,507	1,050
0,80	0,707	0,678	0,654	0,791
0,86	0,791	0,769	0,750	0,607
0,90	0,849	0,832	0,818	0,467
0,95	0,924	0,915	0,907	0,264

4.2 Optimale estimatorer

Hvis vi gir avkall på at estimatet skal stemme overens med differansen mellom de respektive nivåtallestimater, har vi mulighet til å forbedre estimatoren enda en del. Dagsvik (IO 75/24) foreslår:

La $\hat{p}(t,s)$ være en estimator for p(t) basert på den delen av utvalget som er felles ved tidspunkt t og s. For s = t-1 vil det si:

$$\hat{p}(t,t-1) = \overline{Y}'(t) \tag{4.20}$$

$$\hat{p}(t-1,t) = \overline{X}'(t)$$

Da er

$$\hat{d}(s,t) = \hat{p}(t,s) - \hat{p}(s,t) \tag{4.21}$$

en estimator for endringen p(t) - p(s). For s = t-1 har vi spesielt

$$\hat{d}(t-1,t) = \overline{Y}'(t) - \overline{X}'(t-1) \tag{4.22}$$

$$\mathrm{Var}[\hat{d}(t-1,t)] = (\sigma^2/n)\, 2(1-\rho)/\lambda$$

hvor $\rho = \rho_1$. La nå $p^*(t)$ være en estimator for $p(t)$ basert på utvalget ved tidspunkt t. Dvs. $p^*(t) = \overline{Y} = \mu\overline{Y}''(t)+\lambda\overline{Y}'(t)$. En bedre estimator for endringene fremkommer da ved å sette:

$$\tilde{d}(s,t) = (1-k)[p^*(t) - p^*(s)] + k\hat{d}(s,t) \tag{4.23}$$

hvor k er en konstant mellom 0 og 1. For s = t-1 gir dette:

$$\begin{aligned}\tilde{d}(t-1,t) &= (1-k)[\overline{Y}(t) - \overline{X}(t-1)] + k[\overline{Y}'(t) - \overline{X}'(t-1)] \qquad (4.24)\\ &= (1-k)\mu[\overline{Y}''(t) - \overline{X}''(t-1)] + (\lambda+\mu k)[\overline{Y}'(t) - \overline{X}'(t-1)]\end{aligned}$$

Setter vi nå $\mathrm{Var}[\tilde{d}(t-1,t)] = (\sigma^2/n)H''$, får vi

$$H'' = 2\{(1-k)^2\mu + (\lambda+\mu k)^2(1-\rho)/\lambda\} \tag{4.25}$$

Minimerer vi dette, får vi

$$k = \lambda\rho/(1-\mu\rho)$$

som gir endringsestimatoren lik

$$\tilde{d}(t-1,t) = \{\mu(1-\rho)[\overline{Y}''(t) - \overline{X}'(t-1)] + \lambda[\overline{Y}'(t) - \overline{X}'(t-1)]\}/(1-\mu\rho) \tag{4.27}$$

og den normerte variansen H'' blir

$$H'' = 2(1-\rho)/(1-\mu\rho) \tag{4.28}$$

Setter vi $\lambda = \mu = 0{,}5$, får vi spesielt

$$\tilde{d}(t-1,t) = \{(1-\rho)[\overline{Y}''(t-1) - \overline{X}''(t-1)] + [\overline{Y}'(t) - \overline{X}'(t-1)]\}/(2-\rho) \tag{4.29}$$

og

$$H'(1) = 4(1-\rho)/(2-\rho) \tag{4.30}$$

Følgende tabell sammenligner H'' med H(1) for noen verdier av ρ:

Tabell 5. De asymptotiske verdiene for de normerte varianser til $\hat{d}(t-1,t)$ og $\tilde{d}(t-1,t)$ for noen verdier av ρ

ρ	H'(1)	H(1)	H'(1)/H(1)
0,25	1,71	1,72	1,00
0,50	1,33	1,36	0,98
0,70	0,92	0,99	0,94
0,80	0,67	0,75	0,89
0,86	0,49	0,58	0,84
0,90	0,36	0,45	0,80
0,95	0,19	0,26	0,72

Vi finner enkelt at lim(H"/H) = 0,5.

På tilsvarende vis finner vi estimatoren for endringer fra (t-4) til t lik

$$\tilde{d}(t-4,t) = \{\mu[1-\rho_4][\bar{Y}''(t)-\bar{X}''(t-4)] + \\ + \lambda[\bar{Y}'(t) - \bar{X}'(t-4)]\}/[1-\mu\rho_4] \tag{4.31}$$

hvor de enkelte ledd i uttrykket har samme betydning som i uttrykk (4.12), men nå bare med referanse til tidspunktene t og t-4 istedenfor t og t-1. Også nå er $\lambda = \mu = 0,5$ i AKU, slik at estimatoren får formen

$$\tilde{d}(t-4,t) = \{(1-\rho_4)[\bar{Y}''(t) - \bar{X}''(t-4)] \\ +[\bar{Y}'(t) - \bar{X}'(t-4)]\}/(2-\rho_4) \tag{4.32}$$

med varians

$$\begin{aligned} \text{Var}\ \tilde{d}(t-4,t) &= (\sigma^2/n)2(1-\rho_4)/(1-\mu\rho_4) \qquad (4.33) \\ &= (\sigma^2/n)4(1-\rho_4)/(2-\rho_4) \\ &= (\sigma^2/n)H'(4) \end{aligned}$$

Tabell 6 gir en sammenligning av H'(4) med H(4) for noen verdier av ρ_4 under modellen (4.19). For å lette oversikten er ρ tatt med i forspalten.

Tabell 6. De asymptotiske verdiene for de normerte varianser til $\hat{d}(t-4,t)$ og $\tilde{d}(t-4,t)$ for noen verdier av ρ_4

ρ	ρ_4	H'(4)	H(4)	H'(4)/H(4)
0,25	0,09	1,91	1,83	1,04
0,50	0,30	1,65	1,47	1,13
0,70	0,54	1,27	1,05	1,20
0,80	0,68	0,97	0,79	1,23
0,86	0,77	0,75	0,61	1,24
0,90	0,83	0,57	0,47	1,23
0,95	0,91	0,31	0,26	1,19

Dette antyder at vi kan få en viss gevinst ved å ta med observasjonene fra flere tidspunkter når estimatoren for endringstallene, $\tilde{d}(t-4,t)$, skal konstrueres. Jeg har ikke gjort noe forsøk på dette.

Vi vet at

$$\tilde{d}'(s,t) = \tilde{p}(t) - \tilde{p}(s) \tag{4.34}$$

har lavere varians enn $\hat{d}(s,t)$ fra (4.1). Vi kan derfor av tabell 6 slutte at $\tilde{d}'(s,t)$ er den beste estimatoren for endringstallene fra t-4 til t blant de estimatorene vi har sett på.

Når det gjelder endringen d(t-1,t) fra tidspunkt t-1 til t, må vi studere $\tilde{d}'(t-1,t)$ spesielt før vi kan se hvilken estimator som er best. Å gjøre dette rent analytisk vil ta mye tid, så jeg har ikke gjort dette.

5. Autokorrelasjon

Det er på sin plass å nevne litt om autokorrelasjon. For en mere fyldig drøftelse henviser vi til Dagsviks notat, IO 75/24, og Steinar Bjerves notat, IO 74/36.

La $\rho(s,t)$ betegne autokorrelasjonen mellom tidspunkt s og t. Dagsvik viser at en estimator for denne er gitt ved

$$\hat{\rho}(s,t) = \{\hat{p}(s,t) - \hat{p}(s)\hat{p}(t)\}/\{\hat{p}(t)[1-\hat{p}(t)][1-\hat{p}(s)]\hat{p}(s)\}^{\frac{1}{2}} \tag{5.1}$$

hvor

$\hat{p}(t)$ = relativ andel personer i utvalget som tilhører en gruppe G ved tidspunkt t,

$\hat{p}(s,t)$ = relativ andel personer i utvalget som tilhører gruppen G både ved tidspunkt s og t (t>s). (5.2)

Hvis vi baserer oss på tallene i tidsrommet første kvartal 1975 til annet kvartal 1979 for hver hovednæring (ensifret næring) og tar gjennomsnittet for hver næring, fremkommer følgende tabell:

Tabell 7. Gjennomsnittlige verdier for korrelasjonskoeffisienten $\rho(t-r)$ for r = 1, 3, 4, 5 i tidsrommet 1. kvartal 1975 til 2. kvartal 1979

r=:	1	3	4	5
Næring:				
Jordbruk, skogbruk, fiske og fangst	0,82	0,78	0,78	0,79
Bergverk, industri, vann og kraftforsyning	0,88	0,81	0,80	0,79
Bygg og anlegg	0,85	0,79	0,76	0,73
Varehandel, hotell og restaurant	0,85	0,75	0,74	0,70
Transport	0,88	0,80	0,78	0,75
Bank, forsikring, finans og eiendomsdrift	0,91	0,81	0,80	0,76
Offentlig administrasjon, politi og rettsvesen	0,87	0,85	0,76	0,80
Totalt gjennomsnitt	0,86	0,80	0,77	0,77

De totale gjennomsnittene er gjennomsnitt av korrelasjonen for hver enkelt næring. Årsaken til at korrelasjonskoeffisientene for r = 2 mangler ovenfor, er Byråets rulleringsplan, se figuren s.5. Den gjør at tidspunktene (t-2) og t ikke har felles observasjoner.

Det er en ting ved tallene vi kan merke oss, og som kan virke forbausende, nemlig at korrelasjonen jevnt over blir mindre over tid. Med andre ord at $\rho(t-4,t)$ faktisk er mindre enn $\rho(t-3,t)$. Siden det er mange yrker som er svært sesongpreget, skulle man kunne vente motsatt resultat. Men også svenske resultater viser det samme. Vi vil ikke her forsøke å gi noen forklaring på hvorfor det er slik.

Vi ønsker nå å se litt nærmere på forskjellige modeller for korrelasjonen. Vi vil anta at korrelasjonen $\rho(t,t+h)$ er uavhengig av t, slik at vi i fortsettelsen vil sette

$$\rho(t,t+h) = \rho(h) \tag{5.3}$$

Bjerve (IO 74/36) nevner litt om forskjellige modeller for korrelasjonen. Den mest vanlige, og som bl.a. Des Raj (1968) bruker, er modellen:

$$\rho(h) = \rho^{h} \tag{5.4}$$

(ρ betegner som vanlig $\rho(1)$).

Av tabell 5 ser vi at en slik modell, selv om den er enkel å bruke, gir for lave verdier for $\rho(h)$, selv for små verdier av h. Bjerve nevner i sitt notat (2(e) s. 8) at Rao og Graham foreslår et rotasjonsmønster som tar hensyn til at det kan være sesongvariasjon slik som i AKU, nemlig modellen:

$$\rho(4i+j) = \rho_1^{i}\,\rho_2^{j} \tag{5.5}$$

for j=0,1,2,3, og i=0,1...

Ideen med dette korrelasjonsmønsteret er at ρ_1 representerer den årvisse korrelasjonen, mens ρ_2 representerer den kvartalsvise. Sammenligner vi denne modellen med tabell 5, finner vi at den er noe bedre, men at den fortsatt gir for små verdier til $\rho(3)$ og $\rho(4)$, iallfall brukt i AKU.

Det er derfor av interesse å finne en modell som stemmer bedre overens med dataene i tabellen. Siden tabellen viser at $\rho(5)$ ikke faller særlig i forhold til $\rho(4)$, har jeg tatt utgangspunkt i følgende modell:

$$\rho(h) = \rho^{f(h)} \tag{5.6}$$

hvor jeg forsøksvis har prøvd funksjonen

$$f(h) = \sqrt[a]{h} \tag{5.7}$$

Ved å ta utgangspunkt i siste linje i tabell 7, altså velge $\rho = 0,86$ og $\rho(4) = 0,77$, får vi a = 2,5. M.a.o. skal vi bruke funksjonen

$$\rho(h) = \rho^{\sqrt[2,5]{h}} \tag{5.8}$$

Det gir følgende beregnede verdier for korrelasjonene til de enkelte næringer:

Tabell 8. Beregnede korrelasjonskoeffisienter $\rho(h)$ på grunnlag av funksjonen $\rho(h) = \rho^{2,5\sqrt{h}}$

h = :	1	2	3	4	5
Jordbruk, skogbruk, fiske og fangst....	0,82	0,77	0,73	0,71	0,68
Bergverk, industri, vann og kraftforsyning	0,88	0,85	0,82	0,80	0,79
Bygg og anlegg	0,85	0,80	0,77	0,75	0,73
Varehandel, hotell og restaurant	0,85	0,80	0,77	0,75	0,73
Transport	0,88	0,84	0,81	0,79	0,78
Bank, forsikring, finans og eiendomsdrift	0,91	0,88	0,86	0,85	0,83
Offentlig administrasjon, politi og tettsvesen	0,87	0,83	0,80	0,78	0,76
Totalt gjennomsnitt	0,86	0,82	0,80	0,77	0,76

Sammenligner vi med verdiene i tabell 5, ser vi at (5.8) gir en overraskende god tilnærmelse for de fleste næringer.

6. Brukerbeskrivelse av program for nivå- og endringstall

6.1 Innledning

Det som følger er en kort brukerbeskrivelse av programmet som beregner endringstall og sammensatte estimater for nivåtallene på grunnlag av de estimatene som er funnet i tidligere kapitler. Det er skrevet med tanke på at man skal kunne kjøre programmet direkte fra terminal uten å måtte søke hjelp for hver gang.

Ved beregning av nivåtallene bruker programmet følgende tre estimatorer:

$$\hat{p}(t) = \overline{Y}(t)$$

$$\hat{p}_2(t) = 0,34\overline{Y}''(t) + 0,66\{\overline{Y}'(t) + 0,86[\hat{p}(t-1) - \overline{X}'(t-1)]\} \qquad (6.1)$$

$$\hat{p}_3(t) = 0,2348[\hat{p}(t-4) - \overline{X}'(t-4)] + 0,3905[\hat{p}(t-1) - \overline{X}'(t-1)] + \{0,3746\overline{Y}(1,t) + 1,1557\overline{Y}(2,t) + 0,8443\overline{Y}(3,t) + 1,6254\overline{Y}(4,t)\}/4$$

$\hat{p}_2(t)$ er identisk med estimatoren $\hat{p}(t)$ i (2.14), og $\hat{p}_3(t)$ er identisk med $\tilde{p}(t)$ fra (3.8) når verdiene a = 0,2348, b = 0,3905 innsettes i uttrykket for $\tilde{p}(t)$. Jeg har i kap. 2 og 3 sammenlignet disse estimatorene med andre mulige estimatorer, mere spesifikt med følgende estimatorer:

$$\hat{p}_2 = a(\overline{X}''-\overline{X}') + c\overline{Y}' + (1-c)\overline{Y}'' \tag{6.2}$$

$$\tilde{p}(t) = a[\overline{X}(t-4) + \hat{d}(t-4,t)] + b[\overline{X}(t-1) + \hat{d}(t-1,t)] + (1-a-b)\overline{Y}(t)$$

hvor $\hat{d}(t-s,t) = \overline{Y}'(t) - \overline{X}'(t-s)$ er en estimator for endringen fra t-s til t. Jeg viste at disse to estimatorene gir en variansgevinst sammenlignet med $\overline{Y}(t)$, men at $\hat{p}_2(t)$ og $\hat{p}_3(t)$ gir enda større variansgevinster enn disse igjen. Den beste estimatoren blant de jeg har sett på, er $\hat{p}_3(t)$. Innen utvalgsområdene gir den en variansreduksjon sammenlignet med $\overline{Y}(t)$ på vel 40%, mens de andre gir en reduksjon på 30% eller mindre. Jeg har latt programmet tabellere alle de tre estimatorene i (6.1) for nivåtallene for å lette sammenligningen av dem i forbindelse med en eventuell analyse av tallene.

Programmet beregner også tall over endringene fra tidspunkt t-1 og t-4 til t. Jeg har valgt å bruke de optimale estimatorene fra avsnitt 4.2. Setter vi inn ρ = 0,86 i (4.4) og (4.15), vil man finne at den optimale estimatoren

$$\tilde{d}(t-1,t) = \{0,14[\overline{Y}''(t) - \overline{X}''(t-1)] + [\overline{Y}'(t) - \overline{X}'(t-1)]\}/1,14 \tag{6.3}$$

gir en variansgevinst på 14% i forhold til estimatoren

$$\hat{d}(t-1,t) = \hat{p}_2(t) - \hat{p}_2(t-1) \tag{6.4}$$

For endringen fra t-4 til t har jeg tilsvarende basert programmet på

$$\tilde{d}(t-4,t) = \{0,23[\overline{Y}''(t) - \overline{X}''(t-4)] + [\overline{Y}'(t) - \overline{X}'(t-4)]\}/1,23 \tag{6.5}$$

Et nærmere studium av

$$\hat{d}(t-4,t) = \hat{p}_2(t) - \hat{p}_2(t-4) \tag{6.6}$$

viser at estimatoren i (6.6) har lavest varians av disse to estimatorene (se tabell 6). Det vil derfor være mest fordelaktig å basere estimeringen av endringstallene fra t-4 til t på estimatoren

$$\tilde{d}'(t-4,t) = \hat{p}_3(t) - \hat{p}_3(t-4) \tag{6.7}$$

Dette fordi $\hat{p}_3(t)$ har mindre varians enn $\hat{p}_2(t)$, noe som ble vist spesielt for tilfellet $\rho = 0{,}86$. Det vil følgelig lønne seg å basere en analyse av endringstallene på både $\tilde{d}(s,t)$ og $\tilde{d}'(s,t)$. $\tilde{d}(s,t)$ blir tabellert av maskinen, og man finner $\tilde{d}'(s,t)$ uten store vanskeligheter ved å ta differensene mellom de nye nivåtallsestimatene $\hat{p}_3(t)$ og de man har fra tidspunkt s. Se forøvrig kapittel 4.

Ved beregning av vektene som inngår i uttrykkene for de enkelte estimatorene har jeg tatt utgangspunkt i følgende verdier på korrelasjonskoeffisientene: $\rho(1) = 0{,}86$, $\rho(3) = 0{,}79$, $\rho(4) = 0{,}77$, $\rho(5) = 0{,}76$.

Selve programmet finnes i to versjoner, nemlig

SSB/SSB15/ROS/NIVAA

og

SSB/SSB15/ROS/NIVFIL

Den eneste forskjellen på de to versjonene er at mens NIVAA forutsetter at man taster inn dataene fra terminalen under kjøringen av programmet, leser NIVFIL dataene fra en file. De to neste avsnittene gjør nærmere rede for hver av de to versjonene.

6.2 Inntasting av data under kjøring

Ønsker man å taste dataene inn direkte under kjøring av programmet, gir man følgende to kommandoer:

```
FORT OLD /SSB15/ROS/NIVAA
RUN
```

Kommandoen "FORT" forteller maskinen at den skal bruke FORTRAN-kompilatoren. Kommandoen "OLD" forteller at det er en gammel ("old") file. Den siste stringen på linjen er så selve filenavnet. "RUN"-kommandoen på neste linje forteller maskinen at programmet skal kjøres.

Når disse to linjene er tastet inn, blir det et opphold på i underkant av et minutt før noe skjer. Det kommer av at maskinen søker efter feil, samt kompilerer programmet før det kjøres. Når den så har avsluttet kompileringen, vil det bli skrevet følgende på terminalen:

PROGRAMMET BEREGNER ABSOLUTTE TALL FOR DE ENKELTE NAERINGER,
SAMT ENDRINGSTALL FRA FORRIGE KVARTAL OG FORRIGE AAR.
HVER GANG TALLENE ER INNTASTET, BLIR DE GJENTATT PAA NESTE LINJE
FOR KONTROLLENS SKYLD.
OPPGI TOTALT ANTALL PERSONER FOR ET AAR SIDEN I GRUPPE 1,2,3,4.
GI TALLENE I HELE HUNDRER UTEN BLANK MELLOM HVERT TALL,
SOM SEKSSIFREDE TALL.
=

Når likhetstegnet kommer på terminalen, er det et varsel fra maskinen at den venter data. Man taster da inn de tallene maskinen ber om. Istedenfor å bruke blanke mellomrom mellom tallene, kan man bruke komma isteden. Det vil si at man kan inntaste tallene på en av de følgende måter: (Notasjonen (cr) betyr at linjene avsluttes med trykk på RETURN-knappen.)

7001,7002,7003,7004 (cr)

eller

7001 7002 7003 7004 (cr)

Terminalen gjentar så tallene på neste linje. Deretter fremkommer følgende tekst på terminalen.

TOTALT ANTALL I GRUPPE 1,2,3,4 FOR ET KVART. SIDEN,
6-SIFREDE TALL.

Man svarer da på samme måte som på det forrige spørsmålet når likhetstegnet fremkommer. Terminalen gjentar så de nye dataene. Deretter kommer neste spørsmål frem:

TOTALT ANTALL PESONER I HVER AV GRUPPENE 1,2,3,4
I INNEVAERENDE KVARTAL, 6-SIFREDE TALL.

som besvares på samme måte. Tilsvarende besvares de enkelte spørsmålene som kommer efter tur:

ANTALL NAERINGER, 3-SIFRET TALL.
= (antall næringer inntastes) (cr)

OPPGI NAERINGSNUMRENE, 3-SIFRET TALL,
MED EN BLANK MELLOM, 10 NUMRE PR. LINJE
=

Her inntaster man 10 og 10 numre om gangen. For hver gang en linje er ferdig og man har trykket på RETURN-knappen, gjentaes tallene på linjen av maskinen. Deretter vil et nytt likhetstegn komme, slik at man kan fortsette å inntaste data.

Når vi er kommet hit, er maskinen klar til å motta dataene for den enkelte næring. Dermed fremkommer det på terminalen:

OPPGI DE ABSOLUTTE TALL FOR HVER NAERING I HELE HUNDRE.
BRUK TO LINJER PR. NAERING.
BRUK SEKSSIFREDE TALL UTEN BLANK MELLOM I FOELGENDE REKKEFOELGE:
LINJE 1: FJORAARETS TALL: PULJE 1,2,3,4;
FORRIGE KVARTAL: 1,2,3,4;
LINJE 2: DAGENS TALL: PULJE 1,2,3,4

NB! Her er det viktig at man passer på å taste dataene slik programmet ber om det. Dataene for fjoråret og forrige kvartal skal altså komme på samme linje. Man vil forøvrig fort oppdage at man har gjort feil, siden man vil få nuller når dataene gjentaes hvis man skifter linje før alle dataene er kommed med.

Det er mulig å avbryte programmet og begynne på nytt hvis man skulle ønske det. Da trykker man bare på BREAK-knappen øverst til høyre på tastaturet (gjelder SILENT-terminalen). Når det så kommer en stjerne (*) frem på terminalen, gir man en ny "RUN"-ordre.

Har man allerede tastet inn dataene for noen næringer når uhellet er ute, vil forannevnte prosedyre være svært tidkrevende. Det er derfor bedre å fortsette inntastingen av alle dataene uten å rette opp feilen og kjøre programmet som om dataene var korrekte. Når så de ønskede tabellene er ferdige og det fremkommer en stjerne på terminalen (hvilket varsler om at programmet er ferdig, og maskinen venter på nye ordrer), kjører man programmet på nytt for de næringene som man tastet inn feil første gang.

Når man endelig er ferdig med å taste inn disse dataene, vil terminalen skrive

OPPGI DE SAMMENSATTE NIVAATALL FOR ET AAR SIDEN
I HELE HUNDRER SOM 6-SIFREDE TALL UTEN BLANK MELLOM HVERT TALL.
BRUK 10 TALL PR. LINJE.

=

(data inntastes)

OPPGI DE SAMMENSATTE NIVAATALL FOR ET KVART. SIDEN I HELE HUNDRER SOM 6-SIFREDE TALL UTEN BLANK MELLOM HVERT TALL. BRUK 10 TALL PR. LINJE.

Når man så har tastet inn disse dataene, er man ferdig. Da vil maskinen spørre etter hva slags tabell man ønsker, ved følgende spørsmål:

HVILKE TABELLER SKAL SKRIVES UT?
0=BEGGE TABELLER; 1=BARE NIVAATALL; 2=BARE ENDRINGSTALL

Ettersom man svarer 0,1 eller 2, vil man få ut henholdsvis begge tabellene, bare tabellen over nivåtallene, eller bare tabellen over endringstallene. Når de er kommet, er kjøringen over, og man kan logge seg av ved kommandoen "BYE" som vanlig.

NB! En liten anmerkning til slutt:

Det hender man vil være plaget av forstyrrelser på linjen. Vanligvis vil terminalen be en gjenta den siste linjen man tastet inn med kommandoen

RETRANSMIT LAST LINE

Da må man gjenta linjen på vanlig måte.

Forstyrrelsene kan også resultere i at maskinen oppfatter tegn vi ikke har tastet inn. Anta for eksempel vi har tastet inn

391,382,378,385

men at maskinen p.g.a. forstyrrelser på linjen har oppfattet det som om vi begynte linjen med en Z. I det tilfelle vil maskinen skrive ut følgende:

```
Z391,382,378,385
^
FILE CODE 05 ILLEGAL CHAR;   CORRECTION =
```

Da gjentar man ikke hele linjen, men bare det tegnet som skulle være på den markerte plassen (markert ved en ^ under det tegnet som er galt). Hvis det ikke skulle være noe tegn der, slik som nå, gir man blank og trykker på RETURN.

6.3 Innlesning av data fra file

En ulempe med å taste inn tallene under kjøringen av programmet, er at man risikerer å gjøre feil og derfor må foreta kjøringen på ny. Dette vil kunne bli svært tidkrevende. Den andre versjonen av programmet er derfor laget med det formål for øye at man først legger dataene inn på en egen file, og så lar programmet lese fra filen. Det har også den fordel at de dataene som er konstant fra gang til gang, f.eks. næringsnumrene, ikke må inntastes på nytt hver gang, siden de allerede vil ligge på filen etter første gangs kjøring.

Programmet vil lese dataene inn fra filekode 1. Man må derfor omdøpe datafilen til 01. Dette gjøres enkelt på følgende måte:

Anta datafilen (dvs. den filen dataene ligger på fra tidligere) heter

```
SSB/SSB15/ROS/NIVDATA
```

Når man skal taste inn de nye dataene på filen, gir man følgende kommando:

```
EDIT OLD /SSB15/ROS/NIVDATA"01"
```

Dermed blir filen registrert i AFT (Accessible File Table) under betegnelsen 01. Når man så retter på filen, bruker man bare vanlige EDIT-kommandoer. Skulle man ikke kjenne bruken av systemet, finnes det beskrevet i manualen for TEXT EDITOR.

Når dataene inntastes på filen, må man passe på å bruke rett format. Denne gang kan man ikke adskille tallene med komma, men må bruke det korrekt antall blanke anslag som må til for at tallene skal bli rett plassert. Denne versjonen av programmet ønsker dataene i samme format og rekkefølge som versjon nr. 1. Man finner derfor i avsnitt 6.2 hvordan tallene skal ligge på filen. Forøvrig kan det lønne seg å ha en tidligere kjøring ved siden av seg når man skal taste inn de nye dataene på filen.

NB! Husk å fjerne de uaktuelle dataene på filen enten ved en D (delete)-kommando, eller R(replace)-kommando.

Når filen er rettet opp, gir man ordren

RESAVE 01

Når maskinen svarer

DATA SAVED —— 01

er den klar til å kjøre programmet. Man gir da kommandoen

FORT OLD /SSB15/ROS/NIVFIL
RUN

Dermed får man utlistet tabellene automatisk uten å måtte gjøre mere.

Når maskinen er ferdig med kjøringen og det kommer en stjerne på terminalen, bør man sjekke utlistingen av dataene, som maskinen foretar under kjøringen. Det vil ofte kunne hende man har fått enkelte av dataene feil plassert, enten fordi det er kommet fremmede tegn på linjen (forstyrrelser på telefonlinjen), eller fordi man har brukt for mange eller for få blanke anslag på en linje. Da må man kalle tilbake EDIT-systemet og få rettet disse feilene. Dette gjøres på følgende måte

DONE
OLD 01

Da er man tilbake i EDIT-systemet, og har fått tilbake datafilen.

Anta man på utlistingen av dataene oppdager at en linje ser ut som følger:

1015 2113 !1788 Z 1464

Da er ! og Z fremmede tegn som må fjernes. Nå kan det være vanskelig å se nøyaktig hva slags tegn som skal fjernes. Hvis man er usikker og ikke ønsker å bruke kommandoene

```
DS:/!/ DS:/Z/
```

kan man isteden gjøre som følger: Etter å ha fått pointeren til den angjeldende linjen (ved FS-kommandoen eller PS-kommandoen), skriver man

```
RS:/3/,/1/:/3 1/
RS:/88/,/1/:/88 1/
```

Dermed blir linjen rettet til

```
1015 2113 1788 1464
```

og de fremmede tegnene er fjernet.

Tilsvarende går man frem om det er antall blanke anslag som skal forandres. Når filen så er ferdig rettet, skriver man

```
RESAVE 01
FORT OLD NIVFIL
RUN
```

og programmet blir kjørt med de korrigerte dataene.

Når man er fornøyd med utlistingen, er man ferdig, og gir kommandoen BYE.

Referanser.

Bjerve, Steinar (1974): "Sammensatt estimering ved roterende utvalg. Matematisk - statistiske problemer knyttet til arbeidskraftundersøkelsene". Statistisk Sentralbyrå, Arbeidsnotat (IO 74/36).

Dagsvik, John (1975): "Presisjonsgevinst ved bruk av sammensatt estimering i Byråets arbeidskraftundersøkelser". Statistisk Sentralbyrå, Arbeidsnotat (IO 75/24).

Des Raj (1968): "Sampling Theory". McGraw-Hill, New York.

Malmberg, S. (1971): "Korrelasjonsberäkningar i arbetskraftsundersökningarna (A.K.U.)." Statistisk Centralbyrå, Resultatredovisning, 6.11.1970.

Malmberg, S. (1971): "Korrelasjonsberäkningar i arbetskraftsundersökningarna (A.K.U.) 1970-1971". Statistisk Centralbyrå, Utredningsinstitutet, Rapport, 26.3.1971.

Vannebo, Olav (1975): "Bruttostrømmer i arbeidsmarkedet. 1. kvartal 1972 - 2. kvartal 1974". Statistisk Sentralbyrå, Arbeidsnotat (IO 75/2).