

Håkan Lökvist

**Prognose tall for sosialklientdata
tilknyttet Styrings- og
informasjonshjulet for helse- og
sosialtjenesten i kommunene**

Prognosetall for sosialklientdata tilknyttet Styrings- og informasjonshjulet for helse- og sosialtjenesten i kommunene.

av

Håkan Lökvist

1. INNLEDNING.	3
2. PROBLEM.	4
3. KJENNEMERKER.	5
4. METODE	7
4.1 VALG AV METODE.....	7
4.2 ESTIMERING.	7
4.2.1 <i>Utvalg og populasjon.</i>	7
4.2.2 <i>Foreløpige tall og tallkvalitet.</i>	9
4.2.3 <i>Anvendelse av metodene på materialet.</i>	9
4.2.4 <i>Statistisk usikkerhet og skjevhet.</i>	10
4.3 KONKLUSJONER AV FORSØKENE MED SIMULERING AV FORHOLDENE I 1995.....	12
4.4 PROGNOSETALL: GJENNOMFØRING OG KONTROLL AV ESTIMERINGENE.	13
4.4.1 <i>Sammenligning med forhold i året før.</i>	13
4.4.2 <i>Gjennomføring av prognosetallsestimeringene i mars 1996.</i>	14
5. GENERELL OPPSUMMERING.	16
6. LITTERATUR	17
7. APPENDIKS.	18
7.1 MODELLER FOR ESTIMERING AV PROGNOSETALL.....	18
7.1.1 <i>Enkel rateestimering.</i>	18
7.1.2 <i>Olkin's modell for multivariat rateestimering.</i>	19
7.2 MODELLTILPASSING VED TRINNVIS REGRESJON OG «COMMON SENSE».....	23

1. Innledning.

Dette notatet handler om en «tenkt» situasjon: vi befinner oss i vinteren/våren 1995, og vi ønsker å finne tallene for antallet sosialstønadstilfeller i 1994. Problemet er at vi ikke har fått inn alle opplysningene som vi trenger på dette tidspunktet, og de tallene som vi allerede har er bare foreløpige og derved ikke helt pålitelige. Vi må derfor ta i bruk hjelpeinformasjon fra kjennemerker som vi kjenner på det aktuelle tidspunktet og putte inn disse i en estimeringsmodell. Denne estimeringsmodellen må vi finne og vurdere. Ettersom vi per idag kjenner det virkelige utfallet kan vi gjøre en kvalitetsvurdering av resultatet og finne frem til erfaringer som går å bruke for hensikten å lage tilsvarende estimater i årene fremover.

I første rekke er det landstall for antallet sosialstønadstilfeller og sosialstønad regnet i kronetall vi er interesserte i. Senere kan det bli aktuelt å gå ned på kommunenivå. Våre erfaringer sier imidlertid at kvaliteten på dataene foreløpig ikke er tilstrekkelig god for et slikt opplegg.

En slik fremgangsmåte forutsetter ikke bare at vi kjenner det virkelige utfallet. I tillegg må datagrunnlaget forventes å være stabilt over tid med hensyn på måleegenskaper, korrelasjonsstruktur o.l. hos kjennemerkene som vil bli brukt i estimeringsmodellene, slik at vi, med de erfaringer som forsøksåret gir oss, har en trygg grunn å stå på når vi bruker disse erfaringene til vårt prognoseformål under de kommende årene.

Når det gjelder selve datamaterialet har vi utført en del tilrettelegginger. Noen kommuner er blitt slått sammen i perioden 1994/95: Borge, Rolvsøy, Kråkerøy, Onsøy og tidligere kommunenummer 0103 Fredrikstad inngår etter sammenslåingen i 0106 Fredrikstad, og vi har valgt å justere kommunegrensene til situasjonen pr. idag.

I dette notatet vil et par forskjellige modeller for statistisk analyse av et datamateriale bli presentert. I presentasjonen har ambisjonen vært å holde seg til en språkbruk som ikke nødvendigvis forutsetter inngående kjennskap til statistisk og matematisk-statistisk terminologi. De litt mer tekniske beskrivelsene er det også henvist til i appendikset bak i notatet.

Det praktiske arbeid med prosjektet som ligger bak dette notatet er i hovedsak utført i samarbeid med den tidligere seksjonen for helse og sosialstatistikk. I denne sammenhengen vil forfatteren takke spesielt Amy Christiansen, Åne Osmunddalen og Harald Tønseth, som har vært de nærmeste samarbeidspartnere. Jon Folkedal og Truls Thirud, kontor for edb/avd. for personstatistikk, har utført en del korrektursetting av selve notatet og kommet med en del synspunkter som berører den språklige fremstillingen, og bør derved også bli omnevnt i denne sammenhengen. I tillegg vill forfatteren rette et takk til Arnfinn Schjalm, seksjon for statistiske metoder og standarder, som har sett på det metode-statistiske innholdet i notatet og fremført noen nyttige synspunkter og kommentarer.

2. Problem.

Av rammepopulasjonen på 435 kommuner i modellforsøkene er vi bare kjent med et utvalg i størrelsesorden 80-120 kommuner. Av en rekke naturlige grunner, blant annet avhengighet mellom kommunetype og rutiner i datahåndteringen som er av avgjørende betydning for evt. forsinkelser når det gjelder innsending av skjema, representerer disse et sterkt skjevt utvalg. I tillegg er datakvaliteten noe usikker ettersom dataene ikke er endelig rettet opp når vi får dem inn i løpet av februar måned. Vi må altså bruke *foreløpige tall* for kommunene.

Vårt problem er hvordan vi skal prøve å finne et estimeringsopplegg som gir optimal tilpassing, sett ut fra de forutsetningene som gjelder. Dette arbeidet handler om å

1. finne gode forklaringsvariabler (kjennemerker)
2. velge en (eller flere alternative) estimeringsmodell(er)
3. utføre en del detaljarbeid
4. utføre en evaluering av kvaliteten på de estimatene som er funnet og trekke konklusjoner

Vi skal i de følgende kapitlene drøfte disse fire punktene. I kapittel 3 ser vi på en rekke kjennemerker som er interessante i sammenhengen. Kapittel 4 handler om valg av estimeringsmodell, sett som et simuleringsforsøk på 1994-data i februar/mars 1995. Detaljarbeidet, som blant annet omfatter hvordan vi kan forbedre datakvaliteten ved f.eks. å fjerne kommuner som gir lav validitet i dataene, blir også gjennomgått her. I tillegg blir det lagd et forsøk til evaluering, hvor simuleringsforsøk basert på tilsvarende forhold for året før blir benyttet. Et sammendrag av de viktigste punktene i dette notatet blir gitt i kapittel 5.

3. Kjennemerker.

Før vi går løs på å finne den eller de estimeringsmodeller som er best egnet for vårt prognoseformål må vi hente inn de opplysninger vi trenger. Fra forskjellige datamateriale har vi valgt å hente inn følgende kjennemerker:

FLSOS95	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -95, foreløpige tall våren -96	fra skjema
FLPSOS95	Andel <i>stønadstilf</i> , sos.hjelp ialt -95 pr. innb. 1.1.96, foreløp. tall v. -96	=FLSOS95/N_INNB96
SOS94	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -94 (endelige tall)	fra skjema
PSOS94	Andel <i>stønadstilfeller</i> , sosialhjelp ialt -94 pr. innbyggere 1.1.95	=SOS94/N_INNB95
FLSOS94	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -94, foreløpige tall våren -95	fra skjema
FLPSOS94	Andel <i>stønadstilf</i> , sos.hjelp ialt -94 pr. innb. 1.1.95, foreløp. tall v. -95	=FLSOS94/N_INNB95
SOS93	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -93	SSB Reg.stat. database
PSOS93	Andel <i>stønadstilfeller</i> , sosialhjelp ialt -93 pr. innbyggere 1.1.94	=SOS93/N_INNB94
FLSOS93	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -93, foreløpige tall våren -94	fra skjema
FLPSOS93	Andel <i>stønadstilf</i> , sos.hjelp ialt -93 pr. innb. 1.1.94, foreløp. tall v. -94	=FLSOS93/N_INNB94
SOS92	Antall <i>stønadstilfeller</i> , sosialhjelp ialt -92	SSB Reg.stat. database
PSOS92	Andel <i>stønadstilfeller</i> , sosialhjelp ialt -92 pr. innbyggere 1.1.94	=SOS92/N_INNB93
FLKRO95	Sosialstønad i <i>kronetall</i> ialt -95, foreløpige tall våren -96	fra skjema
FLPKRO95	Sosialstønad i <i>kronetall</i> ialt -95 pr. innb. 1.1.96, foreløp. tall v. -96	=FLKRO95/N_INNB96
KRO94	Sosialstønad i <i>kronetall</i> ialt -94 (endelige tall)	fra skjema
PKRO94	Sosialstønad i <i>kronetall</i> ialt -94 pr. innbyggere 1.1.95	=KRO94/N_INNB95
FLKRO94	Sosialstønad i <i>kronetall</i> ialt -94, foreløpige tall våren -95	fra skjema
FLPKRO94	Sosialstønad i <i>kronetall</i> ialt -94 pr. innb. 1.1.95, foreløp. tall v. -95	=FLKRO94/N_INNB95
KRO93	Sosialstønad i <i>kronetall</i> ialt -93	SSB Reg.stat. database
PKRO93	Sosialstønad i <i>kronetall</i> ialt -93 pr. innbyggere 1.1.94	=KRO93/N_INNB94
FLKRO93	Sosialstønad i <i>kronetall</i> ialt -93, foreløpige tall våren -94	fra skjema
FLPKRO93	Sosialstønad i <i>kronetall</i> ialt -93 pr. innb. 1.1.94, foreløp. tall v. -94	=FLKRO93/N_INNB94
KRO92	Sosialstønad i <i>kronetall</i> ialt -92	SSB Reg.stat. database
PKRO92	Sosialstønad i <i>kronetall</i> ialt -92 pr. innbyggere 1.1.94	=KRO92/N_INNB93
N_INNB96	Antall innebyggere ialt 1.1.96 (foreløp. tall fra 4. kvartalet 1995)	SSB Reg.stat. database
N_INNB95	Antall innebyggere ialt 1.1.95	SSB Reg.stat. database
N_INNB94	Antall innebyggere ialt 1.1.94	SSB Reg.stat. database
N_INNB93	Antall innebyggere ialt 1.1.93	SSB Reg.stat. database
KOMMKL	kommuneklasse, standard for kommuneklassifisering -95	fra NOS-publ. C 193
K1 - K7	kommuneklasse, «dummyvariabler» som svarer mot hhv. klasse 1 - 7	KN=(KOMMKL=N)
BARNFA93	Andel barnefamiljer pr. familie ialt -93	SSB
ENSMOR93	Andel enslige mødre pr. barnefamilje ialt -93	SSB

Kolonnen til venstre angir kjennemerkenes kortnavn, mens kolonnen i midten forklarer innholdet. Kildeangivelse og eventuell transformering for avledede variabler er angitt i høyrekolonnen. Betegnelsen «fra skjema» innebærer at kjennmerkene er blitt hentet inn direkte fra skjemaopplysninger, de er altså ikke offentliggjort i noen publikasjon.

De foreløpige tallene for sosialstønad i -94 (FLSOS94, FLPSOS94, FLKRO94 OG FLPKRO94) er i etterhånd gjenskapt for å angi hva vi hadde kjennskap til før den endelige opprettingen av opplysningene som skjedde i løpet av sommeren/høsten -95. Disse «upålitelige» (og i enkelte tilfeller manglende) opplysningene har avgjørende betydning for formålet å få simulert de virkelige forholdene og derved trekke realistiske konklusjoner. Tilsvarende gjelder for 93-tallene, mens de foreløpige 95-tallene gjengir den *faktiske* kjennskapet som vi hadde til dataene når prosjektet endelig ble gjennomført første gangen i mars 1996.

Definisjonen av de syv kommuneklassene bør kanskje forklares nærmere. Vi har brukt hovedklassifiseringen i 1995, men unnlatt å bruke undergrupperingene¹. Klasse 1 er primærnæringskommuner, mens klasse 2 er blandede landbruks- og industrikommuner og klasse 3 er en- eller flersidige rene industrikommuner. Klasse 4 er mindre sentrale, blandede tjenesteytings- og industrikommuner, klasse 5 er sentrale slike kommuner, klasse 6 er mindre sentrale tjenesteytingskommuner og, til sist, de sentrale tjenesteytingskommunene finnes i klasse 7.

Kjennmerkene har vi valgt ut med tanke på hva vi rent erfaringsmessig har vurdert som mest formålstjenlig når det gjelder oppbyggingen av modeller. Det finnes god grunn til å tro at f.eks. familietyperelaterte variabler som andelen barnefamilier eller andelen enslige mødre pr. barnefamilie i det minste partielt *kan* forklare størrelsen på SOS94 (eller noen avledet variabel til denne) og derved være brukbare som tilleggsopplysninger til SOS93 og SOS92 som forklaringsvariabler i en multivariat modell. Tilsvarende gjelder for kronetall, hvor KRO94 rimeligvis bør være avhengig av de samme opplysningene i tillegg til KRO93 og KRO92. Likeledes er det rimelig å tro at kommunetype *kan* ha relevans i denne sammenhengen, ikke minst når vi kommer inn på drøftingen om utvalgs-skjevhet. Avledningen av SOS94, SOS93 og SOS92 til andelen sosialhjelpstilfeller (PSOS94, PSOS93 etc.), og tilsvarende avledning av kjennmerkene for kronetall, henger sammen med at vi tror at disse transformasjonene vil gi dataene økt validitet.

¹ Klasse 1 kan inndeles i landbruks- og fiskekommuner, klasse 3 i hhv. ensidige og flersidige industrikommuner.

4. Metode.

4.1 Valg av metode.

Valget av estimeringsmodell avhenger av hvordan sammenhengen mellom kjennemerkene er strukturert, og det bør allerede nå bli nevnt at det generelt er umulig å gi en tolking som er hundre prosent entydig og objektiv. Vi har valgt å se på to ratestimeringsmodeller og sammenlignet disse. Den *enkle* modellen, hvor vi i utgangspunktet beregner raten mellom responsvariabelen og *en* forklaringsvariabel, har vært i bruk tidligere (bl.a. i en artikkel skrevet av Even Flaatten i 1984) og kan bli betraktet som vår referansemodell. Den *multivariate* modellen er et forsøk på å forbedre tallene som denne referansemodellen har gitt.

Valget av forklaringsvariabler bygger på to ting. Vi har valgt dels å se på hva som kommer ut dersom vi kjører trinnvis multipel regresjonsanalyse på datamaterialet, og dels brukt en mer erfaringsmessig og intuitiv vurderingsevne. Dette vil få en nærmere presentasjon i kapittel 7.2. Resultatet er to ratestimeringsmodeller, den konvensjonale «enkle» modellen som får formen

$$EST_{\text{«enkel»}}(\text{SOS94}) = \sum_{i=1}^N \text{SOS93}_i \left(\frac{\sum_{i=1}^n \text{FLSOS94}_i}{\sum_{i=1}^n \text{SOS93}_i} \right)$$

og den *multivariate*, også benevnt «Olkin's metode», som bygger på en kombinasjon av flere enkle modeller og derved får formen

$$EST_{\text{multipel}}(\text{PSOS94}_i) = W_{\text{PSCS93}} \text{TPSOS93} \left(\frac{\sum_{i=1}^n \text{FLPSOS94}_i}{\sum_{i=1}^n \text{PSOS93}_i} \right) + W_{\text{PSCS92}} \text{TPSOS92} \left(\frac{\sum_{i=1}^n \text{FLPSOS94}_i}{\sum_{i=1}^n \text{PSOS92}_i} \right) + W_{\text{ENSMOR93}} T_{\text{ENSMOR93}} \left(\frac{\sum_{i=1}^n \text{FLPSOS94}_i}{\sum_{i=1}^n \text{ENSMOR93}_i} \right)$$

hvor *TPSOS93*, *TPSOS92* og *T_ENSMOR93* er proporsjonstall som er blitt beregnet på grunnlag av hele landet (se nærmere beskrivelse i kommentaren til vår anvendelse av Olkin's metode i kapittel 7.1.2).

De to estimeringsmodellene vil bli mer nøyaktig beskrevet i kapittel 7.1, hvor vi også går inn på begrepene utvalgsvarians og variasjonskoeffisient.

For *kronetall* vil estimeringsmodellene være syntaktisk identiske, mens variabelnavnen selvfølgelig vil bli skiftet ut. Dette vil si at SOS94 erstattes med KRO94, SOS93 med KRO93, og SOS92 erstattes med KRO92.

4.2 Estimering.

4.2.1 Utvalg og populasjon.

Vinteren (februar/mars) 1995 kjente vi bare sosialstønadskjennemerket SOS94 for et mindretall av de 435 kommunene i vår populasjon. I betraktning av dette, og at vi i tillegg bare var kjent med de *foreløpige* tallene fra skjemaopplysningene, vil vårt *nettutvalg* som danner grunnlaget for estimeringen under «realistiske» forhold være begrenset av følgende betingelser:

1. Skjemaene måtte ha vært innlevert før en bestemt dato. Det er ønskelig at vi lager en første prognose allerede i tidsrommet 15. februar til 1. mars (som er den egentlige utgangen av tidsfristen for skjemainnlevering fra kommunene til SSB).
2. Skjema må bare være innlevert en gang, det må ikke være returnert til kommunene for retting og komplettering.
3. Kommunene må ha et EDB-basert opplegg for innsamling av data.
4. Det må i det minste være registrert foreløpige tall for sosialstønad.
5. Storbykommunene Oslo, Bergen og Trondheim *må* være med uansett, dersom punkt 1-4 ikke er oppfylt må opplysningene - hvis dette er mulig - bli førbearbeidet manuelt.
6. Noen kommuner kan det bli tale om å fjerne fordi vi på forhånd vet at datakvaliteten er dårlig.

Punkt 1 er viktig fordi hensikten er å finne et tidspunkt for laging av imputasjoner som tilfredsstillende ønsker om å være «tidlig ute». Punktene 2 og 3 har relevans for kvaliteten på de foreløpige tallene. Disse, som er meget varierende, avhenger av rutiner og prosedyrer for innlevering av opplysninger fra kommunene til SSB. Når det gjelder punkt 4 bør det rimeligvis i tillegg bli stilt et kvalitetsmessig minimumskrav på de foreløpige tallene, her må vi gå etter skjønn. Som kommentar til punkt 5 kan nevnes at vi har spesialbehandlet Oslo kommune etter tre forskjellige s.k. *utvalgsmodeller*. Men dette vender vi tilbake til (vi vil da forklare at vi har tatt utgangspunkt i at vi kjenner sosialstønadstallene for noen enkelte bydeler i kommunen på det aktuelle tidspunktet).

Når det gjelder punkt 6 har vi av forskjellige grunner besluttet å fjerne seks kommuner fra utvalget før estimeringene i forsøksåret 1995 (og «kontrollåret» 1994). De seks kommunene er Lørenskog, Karmøy, Utsira, Sortland, Tromsø og Alta.

Tabell 4.1 viser hvordan antallet kommuner fordeler seg etter de forskjellige kriteriene som ligger bak definisjonen av et utvalg. Fire forskjellige frister for skjemainnlevering i 1995 blir sammenlignet, dessuten oppgis det totale antallet kommuner som oppfyller de forskjellige kriteriene.

Vi har vurdert at det er mest hensiktsmessig å bruke et utvalg som bygger på 1. mars som «prognosedato».

Tabell 4.1. Antallet kommuner i populasjonen som oppfyller kriteriene for akseptabel datakvalitet. Dette ligger til grunn for utvalgsselekteringen. Det valgte nettutvalget pr. den 1. mars, som vises med uthevet skrift, omfatter 121 kommuner.

«Prognose-dato», d.v.s. dato da alle opplysninger må være levert inn	Alle kommuner som har levert skjema (inkl. Oslo)	Komm. som har levert skjema bare en gang	Komm. med EDB-basert reg. av skjema	Komm. som har levert foreløpige tall	Bruttoutvalg: EDB-kommuner som har lev. skjema en gang + Oslo, Bergen og Trondheim	Nettoutvalg: Brutto-utvalget minus kommuner som mangler foreløpige tall samt Lørenskog, Karmøy m.fl. ekskluderte kommuner
15. febr. 1995	290	287	104	249	104	99
1. mars 1995	342	338	130	280	129	121
15. mars 1995	382	377	156	306	154	144
1. april 1995	405	396	170	321	166	155
Totalt	435	416	186	335	175	158

4.2.2 Foreløpige tall og tallkvalitet.

Tallene som vi får inn i perioden februar/mars, og som ligger til grunn for prognosene våre, er ofte beheftet med forskjellige typer av feil og kan derfor bli karakterisert som foreløpige. Det finnes flere forklaringer til den manglende kvaliteten, de viktigste er feilpunching under innsamlingen av data eller at sosialhjelpsytteren har flyttet under året og derfor er blitt registrert i to kommuner. For Oslo er situasjonen i tillegg noe spesiell: datainnhenting er i denne kommunen organisert slik at hver bydel er ansvarlig for innsamling og videresending av opplysningene, slik at vi får disse opplysningene inn på forskjellige tidspunkter for hver bydel. Dette vil få konsekvenser som vi kommer tilbake til i neste kapittel.

4.2.3 Anvendelse av metodene på materialet.

Som vi allerede nevnte i innledningen er det landstall for antall sosialstønadstilfeller og sosialstønad angitt i kronetall vi er interesserte i å gi prognosetall for. Vi har valgt to estimeringsmodeller, enkel og multivariat ratestimering.

Omlag 11 prosent av Norges befolkning bor i Oslo, som har en særstilling når det gjelder datainnhenting, ettersom alle bydelene ikke har levert alle opplysningene når den endelige fristen (som vi har satt til 1. mars) har løpt ut. Derfor har vi brukt tre alternativer for særbehandling av Oslo, som vi henholdsvis har gitt betegnelsene «*utvalgsmodell I, II og III*». I *utvalgsmodell I* utfører vi en «før-estimering» av det foreløpige tallet ved å bruke enkel ratestimering (uansett hva vi velger for estimeringsmetode senere, til «hovedestimeringen»). I denne sammenhengen bør det nevnes at Oslo pr. 1. mars 1995 hadde et utvalg på 16 besvarte bydeler (av totalt 27). Deretter betrakter vi Oslo som hvilken som helst kommune som har levert opplysninger i rett tid. I *utvalgsmodell II* utelukker vi Oslo helt når vi lager estimatorerene (Oslo er altså ikke med i utvalget). *Utvalgsmodell III* tar sikte på å dele Oslo bydeler inn i to strata, eller delkommuner, hvor den ene har levert opplysninger i tid og der ved er med i utvalget, mens den andre delkommunen består av bydeler som mangler slike opplysninger og derfor blir kategorisert som en «kommune» utenfor utvalget. En ulempe med den sistnevnte modellen er at opplegget for multivariat estimering ikke kan brukes ettersom vi ikke kjenner tallene for variabelen «enslige mødre» på bydelsnivå.

De estimerte prognosetallene på landsbasis fremgår av **tabellene 4.2.a og 4.2.b**. I **tabell 4.2.a** handler det om antall tilfeller. Her kan vi se at den enkle modellen generelt sett ser ut til å gi best «treff»: det relative avviket, som blir beregnet som differensen mellom det estimerte utfallet og det virkelige utfallet («fasitverdien») dividert på det sistnevnte, ligger her på et gjennomgående betydelig lavere nivå, både for utvalgsmodell I og II (med hhv 1,17 og 1,60 % mot den multivariate modellens 2,49 og 2,51). Hvis vi ser på utvalgsmodell III, enkel estimering, ser vi at det relative avviket her er lavest med 1,02 %.

Sosialstønad regnet i kronetall, slik det fremgår av **tabell 4.2.b**, oppviser resultater som er litt annerledes. Her gir det multivariate opplegget bedre «treff» hvis vi ser på utvalgsmodell II. Det er imidlertid fortsatt den enkle rateestimatoren i utvalgsmodell III som er den optimale, her med et «treff» som underestimerer «virkeligheten» med omtrent 0,05 prosent eller to millioner kroner.

Vi bør i denne sammenhengen være oppmerksom på at ordet «treff» *ikke* er synonymt med «treffsikkerhet». Det relative avviket gjengir resultatet av *et* forsøk og må derfor ikke bli tolket som et generelt brukbart statistisk usikkerhetsmål. Situasjonen kan sammenlignes med et eksempel fra idrettsverden: en profesjonell golfspiller vil sannsynligvis prestere et mye bedre resultat i en konkurranse enn en nybegynner, men det finnes ingenting som sier at ikke nybegynneren rent tilfeldig skulle kunne få inn en «hole-in-one» i et vellykket slag (selv om sannsynligheten for dette er svært, svært liten), samtidig som den profesjonelle spilleren slår ballen i sjøen ved siden av banen.

**Tabell
4.2.a.**Sosialstønad, *antall tilfeller*, hele landet, for 1994. Prognose pr. 1. mars 1995.N_innb95:
4348410
"Fasit":
177657

<u>Utvalgsmode</u> ll:	<u>Enkel est.</u>	<u>pr. 100 innb.</u>	<u>rel.avvik</u>	<u>Utvalgs- størrelse</u>
I	179744	4,13	1,17 %	121
II	180498	4,15	1,60 %	120
III	179464	4,13	1,02 %	121
<u>Utvalgsmode</u> ll:	<u>Multiv. est.</u>	<u>pr. 100 innb.</u>	<u>rel.avvik</u>	<u>Utvalgs- størrelse</u>
I	182088	4,19	2,49 %	121
II	182118	4,19	2,51 %	120
	-	-	-	-

**Tabell
4.2.b.**Sosialstønad i *kronetall* (mkr), hele landet, for 1994. Prognose pr. 1. mars 1995.N_innb95:
4348410
"Fasit":
3964

<u>Utvalgsmode</u> ll:	<u>Enkel est.</u>	<u>kr pr. innb</u>	<u>rel.avvik</u>	<u>Utvalgs- størrelse</u>
I	3951	908,6	-0,33 %	121
II	4011	922,4	1,19 %	120
III	3962	911,1	-0,05 %	121
<u>Utvalgsmode</u> ll:	<u>Multiv. est.</u>	<u>kr pr. innb</u>	<u>rel.avvik</u>	<u>Utvalgs- størrelse</u>
I	3995	918,7	0,78 %	121
II	3994	918,5	0,76 %	120
	-	-	-	-

4.2.4 Statistisk usikkerhet og skjevhet.

Begrepet *presisjon*, eller «treffsikkerheten» som vi foretrekker å kalle den statistiske usikkerheten i det forrige kapitlet, kan måles kvantitativt. Detsamme gjelder for begrepet *nøyaktighet*, eller omvendt, den systematiske *skjevheten* i et estimat.

De estimeringsmetodene vi har brukt er tilnærmet forventingsrette, og med en utvalgsstørrelse på 121 kommuner i 1995 (eller 79 kommuner i «kontrollåret» 1994, som vi kommer tilbake til i kapittel 4.4) så har vi valgt å se bort fra skjevhet som kan forklares av estimatorene i seg selv ettersom denne er av en ubetydelig størrelsesorden.

En annen kilde til systematisk skjevhet, som har høy relevans for *vårt* materiale, er det faktum at vi bruker foreløpige tall til responsvariabelen. De foreløpige tallenes avvik fra de endelige på landsbasis blir presentert i **tabell 4.3**. De relative avvikene ligger på en forholdsvis rimelig nivå, rundt hhv. 0,8 og 0,4 prosent, avhengig av om det handler om antall tilfeller eller utbetalt stønad i kronetall. Vi ser også at skjevheten er svakt positiv. Dette vil si, at rateestimatorne (både den enkle og den multivariate) vil gi en svak overestimering på grunn av at vi setter inn de foreløpige tallene i tellerne til disse. En interessant iakttagelse er dessuten at de utvalgte bydelene i Oslo, som skiller seg ut fra resten av landet når det gjelder rutiner for datainnhenting, ikke ser ut til å gi noen problemer i denne sammenhengen.

Tabell 4.3. De foreløpige tallenes avvik fra de endelige. Sosialhjelp i antallet tilfeller og kronetall i 1994. Fra nettutvalget pr. den 1.3.95.

<u>Område</u>	<u>Antallet tilfeller i 1994</u>			<u>Kronetall (mkr) i 1994</u>		
	Endelige tall	Foreløpige tall	Avvik i %	Endelige tall	Foreløpige tall	Avvik i %
Utvalgte bydeler i Oslo	21385	21537	0,71 %	663	665	0,30 %
Øvrige komm. i utv.	65556	66105	0,84 %	1428	1434	0,42 %
Hele utvalget	86941	87642	0,81 %	2091	2099	0,38 %

Presisjonen i estimatene ser ut til å være sterkt avhengig av metodevalget, mens valget av utvalgsmodell gir et litt svakere mønster. I **tabell 4.4** ser vi at den estimerte variasjonskoeffisientenⁱⁱ kommer best ut dersom vi bruker det multivariate estimeringsopplegget. Ser vi på den enkle estimeringsmodellen finnes det et svakt tegn som peker i retning av at utvalgsmodell III er mest optimal. Dette gjelder både for antallet sosialstønadstilfeller og utbetalt sosialstønad i kronetall.

ⁱⁱ Begrepet *variasjonskoeffisient*, som vi gir en mer teknisk innføring til i avsnittene 7.1.1 og 7.1.2, uttrykker utvalgsstandardavviket i forhold til det estimerte forventete verdi. Vi bør allerede her gå ut med en advarsel: utvalgsvarians, utvalgsstandardavvik og variasjonskoeffisient slik de er beregnet i dette notatet bygger på at vi har et *enkelt tilfeldig utvalg*. Da definisjonen av det vi har valgt å benevne «et utvalg av kommuner» egentlig ikke kan sies å være et resultat av en tilfeldig trekning, men isteden et resultat av en selektering som bygger på en rekke kriterier som må være oppfylt (edb-kommuner som har levert inn skjema i tid blant annet, se forøvrig de seks punktene i avsnitt 4.2.1), bør vi være meget forsiktige i tolkingen av disse statistiske usikkerhetsmålene. Vi har likvel vurdert at variasjonskoeffisienttallene gir informasjon som er brukbar for vårt formål, å sammenligne kvaliteten på de forskjellige estimeringsoppleggene og gjøre et forsøk å trekke noen konklusjoner.

Tabell 4.4. De estimerte variasjonskoeffisientene, i prosent, for antallet sosialstønadstilfeller og utbetalt sosialstønad i kronetall for 1994, pr. den 1.3.1995.

Sosialstønad, antall tilfeller :

<u>Utvalgsmode</u> ll:	<u>Enkel estimering</u>	<u>Multivariat est.</u>
I	1,37 %	0,97 %
II	1,63 %	0,98 %
III	1,44 %	-

Sosialstønad i kronetall :

<u>Utvalgsmode</u> ll:	<u>Enkel estimering</u>	<u>Multivariat est.</u>
I	2,52 %	1,47 %
II	2,68 %	1,50 %
III	2,33 %	-

Fotnote: en mer korrekt måte å beregne variasjonskoeffisientene til utvalgsmode

ll I på, er å lage en inndeling i to strata, Oslo inndelt i bydeler og resten av landet inndelt i kommuner som enheter, beregne utvalgsvariansen til «førestimatene» til Oslo separat og deretter veie sammen denne med variasjonen til stratumet som inneholder de andre kommunene. Vi har imidlertid valgt et enklere opplegg hvor Oslo, som de andre kommunene i utvalget, er en enhet i en ustratifisert populasjon (kommuner i Norge). Til vårt formål, å gi et vurderingsgrunnlag for en beslutning om hvilket estimeringsopplegg vi bør velge, tror vi at dette bør holde.

4.3 Konklusjoner av forsøkene med simulering av forholdene i 1995.

Hvis vi tar i betraktning at antallet tilfeller og sosialstønad i kronetall regnet rimeligvis bør forklares av de samme bakenforliggende faktorene, så kan det være klokt å velge *et* opplegg for hver av disse responsvariabler som grunner seg på en felles konklusjon. Derved ser det ut som om to slike «felles» opplegg kommer best ut i forsøkene. Det er utvalgsmode

ll II, multivariat estimering og den enkle estimeringen i utvalgsmodell III. De lave variasjonskoeffisientene og de «treffsikre» estimatene i forsøket med 1994 års tall er argumenter som taler for dette, hvis vi ser på både antallet tilfeller og kronetall og lager en totalvurdering av resultatene (hvor også faglig erfaring og kompetanse på området blir tatt i betraktning).

Dersom vi bare tar variasjonskoeffisientene, og ingenting annet, i betraktning er det ikke noen tvil om at de multivariate modellene gir en overlegent høy presisjon. Denne påstanden, som er vår hovedkonklusjon hvis vi ser på datamaterialet fra en rent kvantitativ betraktning, er imidlertid ikke entydig. Alle de fem oppleggene gir overlag estimerer med relativt god presisjon, og vi har altså derfor hatt anledning til å «ressonere oss frem» til en løsning som i høy grad bygger på en mer «pragmatisk» tilnærming til problemene. Vi har også valgt å se bort fra skjevhetene i estimatene ettersom disse generelt ligger på et relativt lavt nivå, uansett modellvalg.

4.4 Prognose tall: Gjennomføring og kontroll av estimeringene.

4.4.1 Sammenligning med forhold i året før.

La oss gjøre en liten kontroll ved å se på forholdene i det foregående år. Vi antar at vi befinner oss i tiden rundt 1. mars 1994. På det tidspunktet har vi samlet inn foreløpige sosialstønadstall for 1993 fra 79 kommuner. Av flere grunner, som blant annet har å gjøre med rutine for datainnhenting i kommunene det året, er datakvaliteten dessuten generelt sett dårligere enn for tilsvarende 1994-tall samme tidspunkt året etter.

Tabell 4.5.a og **4.5.b** er direkte sammenlignbare med hhv. **tabellene 4.2.a, 4.2.b** og **4.4**. Også her ser vi tydelig at variasjonskoeffisientene peker i retning av at de multivariate modelleneⁱⁱⁱ kommer «best» ut, mens de relative avvikene fra «fasit» viser en noe svakere struktur. En noe «motsigelsesfull» iakttagelse som bør nevnes, er at variasjonskoeffisienten for utvalgsmodell II, enkel estimering, kommer dårligst ut for kronetallene med verdien 3,34 %, mens tilsvarende verdi for antallet tilfeller ligger på et meget godt nivå med verdien 1,74 %. For 94-tall i 1995 (se **tabell 4.4**) har vi gjort samme iakttagelse hvis vi begrenser oss til å sammenligne kronetall med antallet tilfeller. Men dersom vi isteden velger å sammenligne de tre utvalgsmodellene så peker **tabell 4.4** i retning av at utvalgsmodell II kommer dårligst ut i begge tilfeller, mens utvalgsmodell III bør være den mest optimale beslutning dersom vi begrenser oss til enkel estimering.^{iv}

Hvis vi betrakter den manglende datakvaliteten er det ikke urimelig at vår konklusjon i kapittel 4.3 om at det multivariate opplegget gir bedre estimater er riktig. Noen særlig avgjørende informasjon om det mest formålstjenlige valget av utvalgsmodell får vi derimot ikke ved å sammenligne med forholdene pr. 1. mars 1994.

ⁱⁱⁱ Til den multivariate modellen for andelen sosialstønadstilfeller i 1993 har vi brukt andelen sosialstønadstilfeller i 1991 og 1992 samt andelen enslige mødre ialt i 1991 som forklaringsvariabler. På tilsvarende måte har vi tilpasset den multivariate modellen for utbetalinger av sosialstønad i kronetall.

^{iv} Man må være klar over at utvalgsfraksjon- og størrelse i utvalgsmodell II skiller seg noe fra de andre modellene. For utvalgsmodell I og III har vi $N=435$ kommuner i populasjonen og $n=121$ kommuner i nettoutvalget (gjelder 94-tall i 1995), for modell II er tallene $N=434$ og $n=120$ (ettersom vi har ekskludert Oslo helt fra beregningene). Ettersom utvalgsstandardavviket beregnes ved at vi

braker korrigeringsfaktoren $\sqrt{\frac{1}{n}(1 - \frac{n}{N})}$, så vil forholdet mellom utvalgsmodell II og tilsvarende

modeller I og III være 1,0053:1. For tilsvarende forhold i året før vil n være hhv. 78 og 79, og dette vil gi forholdet 1,0075:1 (under gjennomføringen av prognostiseringene i mars 1996 brukte vi et utvalg på 88/89 kommuner, noe som ga raten 1,0068:1. En forskjell i korreksjonsfaktorene som ligger under godt og vel en prosent kan vi se bort fra i sammenligningene av utvalgsmodellenes variasjonskoeffisienter i teksten over.

Tabell 4.5.a.

Sosialstønad, *antall tilfeller*, hele landet, for 1993.
Prognose pr. 1.mars 1994.

N_innb94:
4324815
"Fasit":
177732

<u>Utvalgsmode</u> ll:	<u>Enkel est.</u>	<u>pr. 100 innb.</u>	<u>rel.avvik</u>	<u>cv</u>
I	184395	4,26	3,75 %	3,39 %
II	180055	4,16	1,31 %	1,74 %
III	183551	4,24	3,27 %	2,86 %
<u>Utvalgsmode</u> ll:	<u>Multiv.est</u>	<u>pr. 100 innb.</u>	<u>rel.avvik</u>	<u>cv</u>
I	179146	4,14	0,80 %	1,13 %
II	178849	4,14	0,63 %	1,15 %
	-	-	-	-

Tabell 4.5.b.

Sosialstønad i *kronetall (mkr)*, hele landet, for 1993.
Prognose pr. 1. mars 1994.

N_innb94:
4324815
"Fasit", mkr:
3706

<u>Utvalgsmode</u> ll:	<u>Enkel est.</u>	<u>kr. pr. innb.</u>	<u>rel.avvik</u>	<u>cv</u>
I	3618	836,6	-2,37 %	2,66 %
II	3643	842,3	-1,70 %	3,34 %
III	3672	849,1	-0,92 %	2,71 %
<u>Utvalgsmode</u> ll:	<u>Multiv.est</u>	<u>kr. pr. innb.</u>	<u>rel.avvik</u>	<u>cv</u>
I	3600	832,4	-2,86 %	2,07 %
II	3595	831,2	-3,00 %	2,11 %
	-	-	-	-

*Fotnote 1: Utvalgsmode*ll I bygger på førestimering av Oslo kommune på grunnlag av et utvalg på 10 bydeler (jfr. beskrivelsen i kap. 4.2.3).

Fotnote 2: kolonnen lengst til høyre, som er merket cv, angir den estimerte variasjonskoeffisienten, uttrykket i prosent.

4.4.2 Gjennomføring av prognosetallsestimeringene i mars 1996.

Anslagene til sosialstønadstall for året 1995 ble satt til 173000 tilfeller og 4,1 milliarder kroner i Ukens statistikk nr. 11/96. Dette skjedde etter en skjønnsmessig vurdering hvor vi tok mest hensyn til det enkle estimeringsopplegget i utvalgsmode

Forklaringen på hvorfor vi ikke valgte å forholde oss til *et* bestemt estimeringsopplegg som må gi optimalt troverdige estimater var at opplysningene som konklusjonene i de forrige kapitlene bygger på er noe diffuse. Vi vurderte, med utgangspunkt fra variasjonskoeffisientene til forsøkene i både 1994 og 1995, at det multivariate opplegget for rateestimering rent generelt var bedre enn det enkle. Sammenligningene med «fasitverdiene» i de samme forsøkene, de s.k. relative avvikene, pekte ikke alltid i samme retning som variasjonskoeffisientene. En grunn til dette kan selvfølgelig være at estimatene i noen tilfeller er skjeve. I kapittel 4.2.4 resonnerer vi oss imidlertid frem til at denne skjevheten burde være forholdsvis beskjeden. Vi nevnte også (i kapittel 4.2.3) at estimatene faktisk varierer tilfeldig, og at en simulering med bare et forsøk faktisk *kan* gi dårlig «treff».

Prognosene for 1995 års tall i begynnelsen av mars måned fremgår av **tabellene 4.6.a** og **4.6.b**. Her kan vi se at variasjonskoeffisientene er lavest for de multivariate modellene, uansett om det handler om antallet tilfeller eller utbetalte kronetall (den samme konklusjonen trakk vi for de to foregående årstallene i forsøkene som vi har beskrevet). Det virker også som om utvalgsmoell III, enkel estimering, kommer rimelig godt ut, særlig for kronetall (for de to foregående årstallene var forholdet faktisk omvendt).

Tabell 4.6.a

**Sosialstønad, antall tilfeller, hele landet, for 1995.
Prognose laget i uke 10-11/1996.**

**Ant. tilfeller
(endelige 94-tall):
177657**

Utvalgsmoell:	Enkel est.	pr. 100 innb.	Endring i 95	cv
I	173671	3,97	-2,24 %	1,83 %
II	172144	3,94	-3,10 %	1,45 %
III	172852	3,96	-2,70 %	1,98 %
Utvalgsmoell:	Multiv.est	pr. 100 innb.	Endring i 95	cv
I	173747	3,98	-2,20 %	0,96 %
II	173716	3,98	-2,22 %	0,97 %
III	-	-	-	-

Tabell 4.6.b.

**Sosialstønad i kronetall (mkr), hele landet, for 1995.
Prognose laget i uke 10-11/1996.**

**N_innb96:
4369216
Kronetall (ende-
lige 94-tall, mkr):
3964**

Utvalgsmoell:	Enkel est.	kr pr. innb	Endring i 95	cv
I	4072	932,0	2,72 %	2,46 %
II	4121	943,2	3,96 %	1,98 %
III	4098	937,9	3,38 %	1,71 %
Utvalgsmoell:	Multiv.est	kr pr. innb	Endring i 95	cv
I	4160	952,1	4,94 %	1,54 %
II	4158	951,7	4,89 %	1,57 %
III	-	-	-	-

5. Generell oppsummering.

Vi har gjort et forsøk på å lage prognosetall for det totale antall sosialstønadstilfeller og total utbetalt sosialstønad i kronetall i Norge. Vi har sett på forholdene i 1995 ved å bruke tall fra 1994 slik at vi kan sammenligne med det «virkelige» utfallet. For å kontrollere materialets «stabilitet» over tid har vi også sett på tilsvarende forhold i 1994. Endelig har vi, i begynnelsen av mars i år, laget anslåtte prognosetall som er publisert i Ukens statistikk nr. 11/96. Disse bygger på foreløpige 1995-tall, og her har vi tillempet de konklusjonene og erfaringene som vi har trukket utfra simuleringsforsøkene for årene før.

Problemet har vært hvordan vi skal prøve å finne et estimeringsopplegg som gir optimal tilpassing når vi har et datamateriale som inneholder *foreløpige tall* for et (skjevt) utvalg av kommuner. De foreløpige tallene er opplysninger som kan inneholde feil på grunn av at de ikke har gjennomgått rutinene for endelig retting når fristen for publisering av de anslåtte prognosene har løpt ut. Utvalget består av kommuner som har levert slike opplysninger som vi kan anta holder en akseptabel standard. Som frist har vi brukt 1. mars i forsøkene.

Som metode har vi valgt å bruke to forskjellige estimeringsmodeller: Den ene er konvensjonell *enkel rateestimering* hvor raten mellom utvalgssummene for sosialstønad i inneværende år og foregående år blir blåst opp med populasjonssummen for foregående år. Den andre er *multivariat rateestimering*, som innebærer at vi «veker sammen» et antall enkle rateestimatorer som har en felles responsvariabel. Vi har her valgt å vekte sammen tre enkle modeller som svarer til tre forklaringsvariabler. Som responsvariabel har vi her valgt å bruke proporsjonstallene for sosialstønad (hhv. tilfeller og kronetall pr. innebygger) i inneværende år, mens proporsjonstall for sosialstønad i de to foregående årene samt andelen enslige mødre blant barnefamilier ialt er forklaringsvariabler. Det multivariate modellvalget er dannet med utgangspunkt i en vurdering som blant annet bygger på et antall forskjellige trinnvise regresjonsutkjøringer av materialet.

Oslo kommune har havnet i en spesiell situasjon, først og fremst på grunn av dets størrelse og inndelingen i bydeler: De enkelte bydelene er ansvarlige for innsending av skjema, og det har medført at ikke alle sender inn opplysningene i tide. Vi får altså en situasjon hvor vi bare kjenner et begrenset utvalg av bydeler som har sendt inn data for Oslo. Derfor har vi brukt tre forskjellige *utvalgsmodeller*: I *modell I* har vi laget en førestimering ved å «blåse opp» de enkelte bydeler som har levert opplysninger om sosialstønad med det totale Oslotallet for sosialhjelp i året før. I *modell II* har vi utelukket Oslo, mens *modell III* bygger på enkel summering av tall for bydelene i Oslo. Til modell III har vi bare kunnet bruke den enkle rateestimeringsmodellen, til de andre både den enkle og den multivariate. Derved har vi sammenlignet fem forskjellige estimeringsopplegg.

Vi har ikke trukket noen entydig konklusjon. Imidlertid har vi sett tegn som peker i retning av at den multivariate rateestimeringsmodellen generelt gir bedre presisjon enn den enkle. Dessuten har vi gjort vurderingen at utvalgsmodell III er å foretrekke hvis vi velger å bruke enkel rateestimering. Etter å ha ført diskusjoner med seksjonen for leveårsstatistikk har vi funnet at valget kanskje først og fremst står mellom to opplegg: *modell II, multivariat estimering* og *modell III, enkel estimering*. Det var også med utgangspunkt i begge disse oppleggene vi lagde prognosene som ble publisert i Ukens statistikk i uke 11, år 1996.

6. Litteratur

- SSB (1994) *Standard for kommuneklassifisering*, NOS C 192
- SSB (1996) «Økte utbetalinger til sosialhjelp», Ukens statistikk nr. 11
- Sosial og helsedepartementet (1994) *Styrings og informasjonshjulet, sammenligningstall for kommunene i 1994*, rundskriv I-20/94
- Weisberg, Sanford (1980) *Applied Linear Regression*, John Wiley & Sons, New York
- Cochran, William G (1977) *Sampling Techniques*, third edition, John Wiley & Sons, New York

7. Appendiks.

7.1 Modeller for estimering av prognosetall.

7.1.1 Enkel rateestimering.

Opplegget for enkel rateestimering (i hensikten å lage totaltall) kan skrives som

$$\hat{Y} = \hat{R}\hat{X} = \frac{\bar{y}}{\bar{x}}\hat{X}$$

hvor \bar{x} og \bar{y} er utvalgsmiddelverdiene for variablene X og Y , mens \hat{R} er en tilnærmedesvis forventingsrett estimator, X er totalsummen av X for hele populasjonen og \hat{Y} er parameteret som vi søker: den estimerte totalsummen av responsvariabelen Y .

Dersom man er interessert i å lage en imputasjon for den enkelte kommune i , vil rateestimatoren få formen

$$\hat{y}_i = \hat{R}x_i = \frac{\bar{y}}{\bar{x}}x_i,$$

hvor x_i er størrelsen på variabel X i kommune i og \hat{y}_i er estimatet på størrelsen til variabel Y i densamme kommunen.

Den estimerte variasjonskoeffisienten^v til den enkle rateestimatoren er

$$cv_R = \sqrt{\frac{N^2(1 - \frac{n}{N})}{(\sum_{i=1}^N \hat{y}_i)^2 n(n-1)} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

hvor N er populasjonsstørrelsen og n er utvalgsstørrelsen.

Variasjonskoeffisienten som vi presenterer, med korreksjonsfaktor for en endelig populasjon, baserer seg på et enkelt, tilfeldig utvalg. Man bør her være oppmerksom på at vårt utvalg ikke er trukket tilfeldig. Vi har imidlertid vurdert at variasjonskoeffisienten, slik den er blitt definert her, er bruklig for vårt formål å lage en sammenligning av de forskjellige oppleggen ved hensyn på kvalitet. Denne saken blir drøftet i kapittel 4.

^v For direkte sammenligninger mellom modellene bruker vi et standardisert mål for standardavvik, den s.k. variasjonskoeffisienten, som er definert som kvadratroten av variansen til den estimerte parameteren dividert på den samme parameteren, dvs. $cv = \frac{\sqrt{V(\theta)}}{\theta}$. Man bør imidlertid være

oppmerksom på at vi har beregnet variasjonskoeffisienten som et standardisert mål for *utvalgsstandardavviket*, slik at utvalgsstørrelse og utvalgsfraksjon *må* være identisk dersom man vil sammenligne den statistiske usikkerheten til forskjellige opplegg.

7.1.2 Olkin's modell for multivariat ratestimering.

Olkin's modell for multivariat ratestimering innebærer at man først lager konvensjonal, enkel ratestimering ved å bruke hjelpeinformasjon fra et antall forklaringsvariabler som er høyt korrelerte med responsvariabelen. Deretter prøver man å finne noen parametre som optimalt vektet sammen de enkle estimatene, slik at den statistiske usikkerheten (utvalgsvariansen) blir minimert. Vi har valgt å se litt både på den teoretiske modellen (som forutsetter kjennskap til populasjonen, en hypotetisk tanke, og som altså bare er et ledd i å forklare hva som skjer), og den utvalgsbaserte modellen (som ble brukt i vår anvendelse i dette notatet).

Teoretisk modell, hvor populasjonen er kjent: Prinsippet for multivariat ratestimering ifølge Olkin's metode er, at vi først lager en rateestimator for hver uavhengig variabel og deretter veier sammen de estimerte tallene. For estimering av totaltall vil dette gi oss modellen

$$\hat{Y}_{\text{tot}} = W_1 \frac{\bar{y}}{\bar{X}_1} N\bar{X}_1 + W_2 \frac{\bar{y}}{\bar{X}_2} N\bar{X}_2 + \dots + W_p \frac{\bar{y}}{\bar{X}_p} N\bar{X}_p ,$$

ved bibetingelsen $\sum_{j=1}^p W_j = 1$,

hvor W_1, W_2, \dots, W_p er parametre^{vi} som vi ønsker å estimere, $\bar{X}_1, \dots, \bar{X}_p$ uttrykker utvalgsmiddelerverdier, $\bar{X}_1, \dots, \bar{X}_p$ er populasjonsmiddelerverdier og N er det totale antall observasjoner.

Dersom formålet er å lage prognosetall for hver enkelt observasjon i , vil denne modellen få formen

$$\hat{y}_i = W_1 \frac{\bar{y}}{\bar{X}_1} x_{1i} + W_2 \frac{\bar{y}}{\bar{X}_2} x_{2i} + \dots + W_p \frac{\bar{y}}{\bar{X}_p} x_{pi} .$$

Estimatoren til parametermengden $\{W_1, W_2, \dots, W_p\}$ fåes ved at vi minimerer variansen til \hat{Y}_{tot} som, uttrykt i matriseform, er

$$V(\hat{Y}_{\text{tot}}) = k \begin{matrix} \mathbf{W}' & \mathbf{V} & \mathbf{W} \\ [1 \times p] & [p \times p] & [p \times 1] \end{matrix}$$

^{vi} Summen av parametrene må være lik 1, og det forklares ved følgende: den enkle rateestimatoren for

den j 'te parameteren er tilnærmet forventingsrett, det vil si $E\left(\frac{\bar{y}_j}{\bar{X}_j} N\bar{X}_j\right) = \hat{Y}_j$. Av dette følger, at

forventningen til den multiple rateestimatoren $E(\hat{Y}_{\text{tot}}) = \sum_{j=1}^p W_j \hat{Y}_j$. Vilkaaret for at \hat{Y}_{tot} skal være

tilnærmet forventingsrett må altså være at summen av parametrene W_1, W_2, \dots, W_p er lik 1, det vil si

$$\sum_{j=1}^p W_j = 1 .$$

hvor $k = \frac{N^2(1 - \frac{n}{N})}{n}$ er en skalær som har betydning for beregning av *utvalgsvariansen*^{vii} (men som er helt uinteressant i forbindelse med estimeringen av $\{W_1, \dots, W_p\}$),

$$\mathbf{W}_{[p \times 1]} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{bmatrix} \text{ er en vektor med parametermengden } \{W_1, \dots, W_p\},$$

$$\text{og } \mathbf{V}_{[p \times p]} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{bmatrix} \text{ er en kovariansmatrise, hvor hver enkelt celledverdi } v_{kl} \text{ er}$$

definert som

$$v_{kl} = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \frac{\bar{y}}{\bar{x}_k} x_{ki})(y_i - \frac{\bar{y}}{\bar{x}_l} x_{li}).$$

Dette vil da gi

$$\mathbf{W}_{[p \times 1]} = \frac{\begin{pmatrix} \mathbf{V}^{-1} & \mathbf{1} \\ [p \times p] & [p \times 1] \end{pmatrix}}{\begin{pmatrix} \mathbf{1} & \mathbf{V}^{-1} & \mathbf{1} \\ [1 \times p] & [p \times p] & [p \times 1] \end{pmatrix}},$$

$$\text{hvor } \mathbf{1}_{[p \times 1]} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \text{ og } \mathbf{1}_{[1 \times p]} = [1 \quad 1 \quad \cdots \quad 1],$$

det vil si, den estimerte parametervektoren \mathbf{W} er lik kolonnesummen til den inverterte kovariansmatrisen dividert på totalsummen til densamme (slik at summen av parameterverdiene i \mathbf{W} alltid er lik 1).

Variasjonskoeffisienten for den teoretiske modellen av Olkin's metode vil bli

$$CV_{OR} = \sqrt{\frac{1}{\sum_{i=1}^N \hat{y}_i^2} V(\hat{Y}_{tot})}$$

^{vii} Uttrykket $(1 - n/N)/n$ korrigerer for et utvalg som er av endelig størrelse ($< \infty$), mens termen N^2 bare skal ses som en tilpassing til at vi har brukt estimering av totaltall i den matematiske modellen (uten N^2 hadde vi fått variansen til middelerdien).

Modell for estimasjoner basert på utvalgsdata: I virkeligheten kjenner vi bare et *utvalg* av observasjoner (121 kommuner av 435 den 1. mars 1995 f.eks.). Vi ønsker altså å estimere parametermengden $\{\hat{W}_1, \hat{W}_2, \dots, \hat{W}_p\}$, hvorved også her gjelder at parametersummen er lik 1, det vil

si $\sum_{j=1}^p \hat{W}_j = 1$, slik at

$$\hat{Y}_{\text{tot}} = \hat{W}_1 \frac{\bar{y}}{\bar{x}_1} N\bar{x}_1 + \hat{W}_2 \frac{\bar{y}}{\bar{x}_2} N\bar{x}_2 + \dots + \hat{W}_p \frac{\bar{y}}{\bar{x}_p} N\bar{x}_p,$$

eller, for imputering av enkelte observasjonsverdier,

$$\hat{y}_i = \hat{W}_1 \frac{\bar{y}}{\bar{x}_1} x_{1i} + \hat{W}_2 \frac{\bar{y}}{\bar{x}_2} x_{2i} + \dots + \hat{W}_p \frac{\bar{y}}{\bar{x}_p} x_{pi}.$$

Estimatoren vil bli funnet, i prinsippet, på samme måte som i tilfellet med den teoretiske «populasjonsbaserte» modellen, ved at vi minimerer variansen til \hat{Y}_{tot} ,

$$v(\hat{Y}_{\text{tot}}) = k \underset{[1 \times p]}{\hat{W}'} \underset{[p \times p]}{\hat{V}} \underset{[p \times 1]}{\hat{W}}$$

$$\text{hvor } k = \frac{N^2(1 - \frac{n}{N})}{n} \text{ (det vil si det samme som i den teoretiske modellen),}$$

$\hat{W}_{[p \times 1]}$ er en vektor som inneholder estimatene til parametervektoren $\mathbf{W}_{[p \times 1]}$,

og $\hat{V}_{[p \times p]}$ er en kovariansmatrise hvor hvert enkelt cellederdi er definert som

$$\hat{v}_{kl} = \frac{1}{(n-1)} \sum_{i=1}^n (y_i - \frac{\bar{y}}{\bar{x}_k} x_{ki})(y_i - \frac{\bar{y}}{\bar{x}_l} x_{li}),$$

og derved vil formelen for estimatoren i dette tilfellet bli

$$\underset{[p \times 1]}{\hat{W}} = \frac{\underset{[p \times p]}{\hat{V}^{-1}} \underset{[p \times 1]}{\mathbf{1}}}{\underset{[1 \times p]}{\mathbf{1}} \underset{[p \times p]}{\hat{V}^{-1}} \underset{[p \times 1]}{\mathbf{1}}}.$$

Den estimerte variasjonskoeffisienten vil her bli beregnet som

$$CV_{\text{OR}} = \sqrt{\frac{1}{N} \frac{v(\hat{Y}_{\text{tot}})}{(\sum_{i=1}^n \hat{y}_i)^2}}.$$

En kommentar til vår anvendelse av Olkin's metode: I vår anvendelse ble det brukt *proporsjonstall*. Vi hadde en modell hvor *andelen* sosialstønadstilfeller pr. innbyggere i 1994 var avhengig av tilsvarende *andel* i 1993 og 1992 samt *andelen* enslige mødre av totalt antall barnefamilier i hver kommune. Disse tallene summerte vi, og deretter beregnet vi vektene W_{PSOS93} , W_{PSOS92} og W_{ENSMOR} etter oppskriften over.

Til selve formålet å lage prognoser hadde vi to beregningsmåter å velge blant. Alternativ 1 var å beregne middelverdien til PSOS94 ved å bruke formelen

$$EST_{multipl}(PSOS94) = W_{PSOS93} \frac{1}{N} \sum_{i=1}^N PSOS93_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n PSOS93_i} \right) + W_{PSOS92} \frac{1}{N} \sum_{i=1}^N PSOS92_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n PSOS92_i} \right) + W_{ENSMOR93} \frac{1}{N} \sum_{i=1}^N ENSMOR93_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n ENSMOR93_i} \right).$$

(«kortnavnen» til alle kjennemerkene ble introdusert for leseren i kapittel 3).

Dette ville imidlertid by på en del problemer ettersom alle kommunene i landet uansett innebyggerantall ville fått samme vekt. Alternativ 2, som innebærer at vi først beregner proporsjonstallene på landsbasis, det vil si

$$TPSOS93 = \frac{\sum_{i=1}^N SOS93_i}{N_INNB94_{Hele_landet}},$$

$$TPSOS92 = \frac{\sum_{i=1}^N SOS92_i}{N_INNB93_{Hele_landet}}$$

og

$$T_ENSMOR93 = \frac{\sum_{i=1}^N ENSMOR93_i \times BARNFA93_i \times NFAM93_i}{\sum_{i=1}^N BARNFA93_i \times NFAM93_i}$$

hvor $N_INNB93_{Hele\ land}$ og $N_INNB94_{Hele\ land}$ angir det totale innbyggerantall i Norge pr. den 1.1.93 og den 1.1.94, og $NFAM93_i$ angir antallet familjer (kilde: SSB) i kommune i ,

og deretter putter inn disse i formelen

$$EST_{multipl}(PSOS94_i) = W_{PSOS93} TPSOS93_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n PSOS93_i} \right) + W_{PSOS92} TPSOS92_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n PSOS92_i} \right) + W_{ENSMOR93} T_ENSMOR93_i \left(\frac{\sum_{i=1}^n HPSOS94_i}{\sum_{i=1}^n ENSMOR93_i} \right).$$

vurderte vi som mer formålstjenlig i denne sammenhengen. Vi gikk derved «rundt» problemet med at f.eks. Utsira kommune med sine drøye 200 innbyggere ellers hadde fått like stor vekt som f.eks. Oslo eller Bergen.

For estimering av totale kronetall brukte vi et tilsvarende opplegg.

7.2 Modelltilpassing ved trinnvis regresjon og «common sense».

For å finne frem til en *multivariat* ratestimeringsmodell med høy forklaringsverdi har vi testet noen forskjellige variabeltransformasjoner og sett på hvordan disse oppfører seg i en trinnvis regresjonsmodell. Vi har både sett på regresjonsmodeller med og uten konstantledd (intercept). Disse blir presentert nedenfor.

Alle kjøringene er blitt utført med SAS. De kritiske p-verdiene for aksept og forkastelse av variabler i modellene («sl_entry» og «sl_stay» under modellspesifikasjonen i SAS-programmet) er blitt satt til henholdsvis 0,05 og 0,10.

Vi har testet de uttransformerte sosialstønadstallene pluss to forskjellige transformeringsmodeller. Disse to er *relative forandringer fra et år til neste* og *antall stønadstilfeller per innebygger*.

Vi har ekskludert Alta og Utsira i regresjonskjøringene. Dette skyldes to ting: Alta hadde ikke levert noen opplysninger om sosialstønad i 1993, og derfor mangler vi tall for året før. Vi vurderte også at Utsira, som er en svært liten kommune med drøyt 200 innbyggere, ville gi en del problemer på grunn av folkemengden. Totalt har vi, etter disse justeringene, et materiale som består av 433 kommuner.

Modell I: Den uttransformerte sosialstønadsvariabelen: dette er den enkleste modellen. Her har vi satt inn SOS94 som en avhengig variabel, mens SOS93, SOS92, BARNFA93, ENSMOR93 og kommuneklasse i form av dummyvariablene K1--K7 er blitt brukt som «kandiderende» forklaringsvariabler:

(a): Regresjon med konstantledd:

Uvektet modell med konstantledd

Stepwise Procedure for Dependent Variable SOS94

Step 1 Variable SOS93 Entered R-square = 0.99935135 C(p) = 38.69168392

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	1117709606.1742	1117709606.1742	664026	0.0001
Error	431	725472.45862317	1683.23076247		
Total	432	1118435078.6328			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	12.82062841	2.03065449	67095.00884106	39.86	0.0001
SOS93	0.96767981	0.00118751	1117709606.1742	664026	0.0001

Bounds on condition number: 1, 1

Step 2 Variable K5 Entered R-square = 0.99937640 C(p) = 22.62958111

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	1117737623.6878	558868811.84390	344558	0.0001
Error	430	697454.94499177	1621.98824417		
Total	432	1118435078.6328			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	9.16922673	2.17839031	28737.02012641	17.72	0.0001
SOS93	0.96736985	0.00116809	1112440503.1704	685850	0.0001
K5	20.97414329	5.04653496	28017.51363140	17.27	0.0001

Bounds on condition number: 1.004093, 4.016372

Step 3 Variable K7 Entered R-square = 0.99938879 C(p) = 15.69463185

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	1117751483.3732	372583827.79107	233820	0.0001
Error	429	683595.25957291	1593.46214353		
Total	432	1118435078.6328			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	7.44283876	2.23709367	17638.05255650	11.07	0.0010
SOS93	0.96618465	0.00122554	990391315.04978	621534	0.0001
K5	23.45360683	5.07212250	34070.73970490	21.38	0.0001
K7	22.47509091	7.62071369	13859.68541886	8.70	0.0034

Bounds on condition number: 1.141862, 9.898168

Step 4 Variable ENSMOR93 Entered R-square = 0.99939880 C(p) = 10.47865326

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	1117762676.6313	279440669.15783	177871	0.0001
Error	428	672402.00147170	1571.03271372		
Total	432	1118435078.6328			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-13.18498803	8.04091571	4224.09512020	2.69	0.1018
SOS93	0.96557414	0.00123819	955388346.47330	608128	0.0001
K5	24.01624996	5.04070786	35662.56387463	22.70	0.0001
K7	21.81643128	7.57091180	13045.36622537	8.30	0.0042
ENSMOR93	119.50385442	44.77094571	11193.25810122	7.12	0.0079

Bounds on condition number: 1.164815, 17.55453

Step 5 Variable K1 Entered R-square = 0.99940433 C(p) = 8.49251001

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	1117768859.8440	223553771.96880	143282	0.0001
Error	427	666218.78881683	1560.23135554		
Total	432	1118435078.6328			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-11.41934910	8.06216081	3130.17622359	2.01	0.1574
SOS93	0.96546397	0.00123517	953253188.76348	610969	0.0001
K1	-9.70317076	4.87417461	6183.21265487	3.96	0.0471
K5	21.27106295	5.20918807	26015.20860744	16.67	0.0001
K7	19.13217808	7.66438110	9722.17979250	6.23	0.0129
ENSMOR93	125.63615642	44.72298575	12312.79659759	7.89	0.0052

Bounds on condition number: 1.179585, 28.11184

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable SOS94

Step	Variable Entered	Number Removed	Partial R**2 In	Model R**2	C(p)	F	Prob>F
1	SOS93		0.9994	0.9994	38.6917	664026.366	0.0001
2	K5		0.0000	0.9994	22.6296	17.2736	0.0001
3	K7		0.0000	0.9994	15.6946	8.6978	0.0034
4	ENSMOR93		0.0000	0.9994	10.4787	7.1248	0.0079
5	K1		0.0000	0.9994	8.4925	3.9630	0.0471

(b): Regresjon uten konstantledd:

Uvektet modell uten konstantledd

Stepwise Procedure for Dependent Variable SOS94

Step 1 Variable SOS93 Entered R-square = 0.99933443 C(p) = 79.94594849

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	1190020388.5325	1190020388.5325	648637	0.0001
Error	432	792567.46746421	1834.64691543		
Total	433	1190812956.0000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
SOS93	0.96947415	0.00120375	1190020388.5325	648637	0.0001

Bounds on condition number: 1, 1

Step 2 Variable ENSMOR93 Entered R-square = 0.99939843 C(p) = 32.81522310

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	1190096598.9707	595048299.48534	358014	0.0001
Error	431	716357.02931842	1662.08127452		
Total	433	1190812956.0000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
SOS93	0.96721179	0.00119346	1091646200.7803	656795	0.0001
ENSMOR93	76.92767375	11.36060602	76210.43814579	45.85	0.0001

Bounds on condition number: 1.085034, 4.340134

Step 3 Variable K5 Entered R-square = 0.99942133 C(p) = 17.23813850

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	1190123864.1354	396707954.71179	247550	0.0001
Error	430	689091.86464179	1602.53922010		
Total	433	1190812956.0000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
SOS93	0.96696646	0.00117339	1088289113.9865	679103	0.0001
K5	20.46850990	4.96233979	27265.16467663	17.01	0.0001
ENSMOR93	58.02608689	12.05979914	37100.10047637	23.15	0.0001

Bounds on condition number: 1.26813, 10.66355

Step 4 Variable K7 Entered R-square = 0.99943179 C(p) = 11.20180854

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	1190136329.9034	297534082.47585	188645	0.0001
Error	429	676626.09659188	1577.21700837		
Total	433	1190812956.0000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
SOS93	0.96588498	0.00122600	978950132.35345	620682	0.0001
K5	22.70939041	4.98708905	32704.54856836	20.74	0.0001
K7	21.30841542	7.57944420	12465.76804991	7.90	0.0052
ENSMOR93	48.94803818	12.39223896	24607.21553752	15.60	0.0001

Bounds on condition number: 1.360506, 20.14205

Step 5 Variable K1 Entered R-square = 0.99943791 C(p) = 8.51044656

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	1190143607.0350	238028721.40699	152202	0.0001
Error	428	669348.96504039	1563.89945103		
Total	433	1190812956.0000			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
SOS93	0.96572131	0.00122317	974852948.31529	623348	0.0001
K1	-10.46267259	4.85028140	7277.13155148	4.65	0.0316
K5	19.93803134	5.12948577	23627.93227924	15.11	0.0001
K7	18.48741091	7.65983923	9110.06984642	5.83	0.0162
ENSMOR93	65.74822010	14.59202859	31750.12826780	20.30	0.0001

Bounds on condition number: 1.902458, 35.50299

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable SOS94

Step	Variable Entered	Number Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	SOS93		1	0.9993	0.9993	79.9459	648637.282	0.0001
2	ENSMOR93		2	0.0001	0.9994	32.8152	45.8524	0.0001
3	K5		3	0.0000	0.9994	17.2381	17.0137	0.0001
4	K7		4	0.0000	0.9994	11.2018	7.9036	0.0052
5	K1		5	0.0000	0.9994	8.5104	4.6532	0.0316

and.ens.mødre.pr.barnf.ialt 93,nordcd
Komm.kl.5-sentr. bl. tjen- og ind.komm.
Komm.kl.7-sentr. tjenesteytingskommuner
Kommuneklasse 1 - primærnæringskommuner

Modell II: Transformering til differanse i tid: Vi har her sett på hva som skjer dersom vi bruker forandringer i sosialstønadstallene fra et år til et annet. Vi har altså definert

$$DSOS94 = \frac{SOS94 - SOS93}{SOS93}$$

og

$$DSOS93 = \frac{SOS93 - SOS92}{SOS92}$$

DSOS94 er deretter blitt brukt som avhengig variabel, og DSOS93 pluss bakgrunnskjenmerkene BARNFA93, ENSMOR93 og kommuneklasse i form av dummyvariablene K1--K7 er blitt brukt som «kandiderende» forklaringsvariabler :

(a): Regresjon med konstantledd:

Uvektet modell med konstantledd

Stepwise Procedure for Dependent Variable DSOS94

No variable met the 0.0500 significance level for entry into the model.

(b): Regresjon uten konstantledd:

Uvektet modell uten konstantledd

Stepwise Procedure for Dependent Variable DSOS94

Step 1 Variable BARNFA93 Entered R-square = 0.03641054 C(p) = 1.34507777

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	0.80981129	0.80981129	15.83	0.0001
Error	419	21.43131367	0.05114872		
Total	420	22.24112495			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
BARNFA93	0.10102595	0.02538975	0.80981129	15.83	0.0001

Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable DSOS94

Step	Variable Entered Label	Number Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	BARNFA93 and.barnfam.pr.familie ialt 93,nordcd		1	0.0364	0.0364	1.3451	15.8325	0.0001

Modell III: Transformering til antall tilfeller per innebygger: denne modellen gir alle kommuner, uansett størrelse, lik vekt. I dette tilfellet har vi definert variablene PSOS94, PSOS93 og PSOS92 slik det fremgår av kapittel 2, og brukt PSOS94 som responsvariabel. PSOS93, PSOS92 samt familie-status- og kommunetypevariablene er på tilsvarende måte som i de to tidligere modellene «kandiderende» forklaringsvariabler:

(a): Regresjon med konstantledd:

Uvektet modell med konstantledd

Stepwise Procedure for Dependent Variable PSOS94

Step 1 Variable PSOS93 Entered R-square = 0.82392949 C(p) = 34.14887243

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	0.05635277	0.05635277	2016.88	0.0001
Error	431	0.01204237	0.00002794		
Total	432	0.06839514			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.00379302	0.00070952	0.00079850	28.58	0.0001
PSOS93	0.89775444	0.01999020	0.05635277	2016.88	0.0001

Bounds on condition number: 1, 1

Step 2 Variable ENSMOR93 Entered R-square = 0.83013597 C(p) = 19.82289856

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	0.05677726	0.02838863	1050.72	0.0001
Error	430	0.01161787	0.00002702		
Total	432	0.06839514			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.00076863	0.00103392	0.00001493	0.55	0.4576
PSOS93	0.84098728	0.02432132	0.03230449	1195.65	0.0001
ENSMOR93	0.02814702	0.00710111	0.00042449	15.71	0.0001

Bounds on condition number: 1.530792, 6.123167

Step 3 Variable K1 Entered R-square = 0.83362514 C(p) = 12.64472006

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	0.05701591	0.01900530	716.50	0.0001
Error	429	0.01137923	0.00002653		
Total	432	0.06839514			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.00108519	0.00102986	0.00002945	1.11	0.2926
PSOS93	0.82128880	0.02497711	0.02867894	1081.20	0.0001
K1	-0.00187586	0.00062539	0.00023864	9.00	0.0029
ENSMOR93	0.03238803	0.00717665	0.00054023	20.37	0.0001

Bounds on condition number: 1.644481, 12.9416

Step 4 Variable PSOS92 Entered R-square = 0.83612788 C(p) = 8.06131609

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	0.05718708	0.01429677	545.95	0.0001
Error	428	0.01120806	0.00002619		
Total	432	0.06839514			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.00113830	0.00102349	0.00003239	1.24	0.2667
PSOS93	0.70253618	0.05266222	0.00466043	177.97	0.0001
PSOS92	0.13239260	0.05178287	0.00017118	6.54	0.0109
K1	-0.00184264	0.00062153	0.00023017	8.79	0.0032
ENSMOR93	0.02991384	0.00719615	0.00045251	17.28	0.0001

Bounds on condition number: 7.404776, 69.58777

Step 5 Variable BARNFA93 Entered R-square = 0.83760643 C(p) = 6.17203636

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	0.05728821	0.01145764	440.48	0.0001
Error	427	0.01110693	0.00002601		
Total	432	0.06839514			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.00683405	0.00306352	0.00012944	4.98	0.0262
PSOS93	0.68922692	0.05291773	0.00441254	169.64	0.0001
PSOS92	0.13922343	0.05172521	0.00018845	7.24	0.0074
K1	-0.00197924	0.00062331	0.00026228	10.08	0.0016
BARNFA93	-0.01126203	0.00571175	0.00010113	3.89	0.0493
ENSMOR93	0.02657183	0.00736956	0.00033816	13.00	0.0003

Bounds on condition number: 7.527251, 94.26819

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable PSOS94

Step	Variable Entered Label	Number Removed In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	PSOS93	1	0.8239	0.8239	34.1489	2016.8830	0.0001
2	ENSMOR93	2	0.0062	0.8301	19.8229	15.7113	0.0001
3	K1	3	0.0035	0.8336	12.6447	8.9969	0.0029
4	PSOS92	4	0.0025	0.8361	8.0613	6.5366	0.0109
5	BARNFA93	5	0.0015	0.8376	6.1720	3.8877	0.0493

(b): Regresjon uten konstantledd:

Uvektet modell uten konstantledd

Stepwise Procedure for Dependent Variable PSOS94

Step 1 Variable PSOS93 Entered R-square = 0.97688988 C(p) = 62.85929421

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	0.54279752	0.54279752	18261.1	0.0001
Error	432	0.01284087	0.00002972		
Total	433	0.55563839			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
PSOS93	0.99753637	0.00738185	0.54279752	18261.1	0.0001

Bounds on condition number: 1, 1

Step 2 Variable ENSMOR93 Entered R-square = 0.97906407 C(p) = 18.39718997

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	2	0.54400559	0.27200279	10077.8	0.0001
Error	431	0.01163281	0.00002699		
Total	433	0.55563839			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
PSOS93	0.84233805	0.02424076	0.03259024	1207.48	0.0001
ENSMOR93	0.03204289	0.00478950	0.00120806	44.76	0.0001

Bounds on condition number: 11.87588, 47.50351

Step 3 Variable K1 Entered R-square = 0.97946743 C(p) = 11.77744545

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	0.54422971	0.18140990	6837.45	0.0001
Error	430	0.01140868	0.00002653		
Total	433	0.55563839			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
PSOS93	0.82388502	0.02485848	0.02914414	1098.46	0.0001
K1	-0.00180832	0.00062218	0.00022412	8.45	0.0038
ENSMOR93	0.03767796	0.00512921	0.00143167	53.96	0.0001

Bounds on condition number: 13.85561, 83.75146

Step 4 Variable PSOS92 Entered R-square = 0.97977021 C(p) = 7.30710225

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	0.54439794	0.13609949	5194.34	0.0001
Error	429	0.01124045	0.00002620		
Total	433	0.55563839			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
PSOS93	0.70630670	0.05256749	0.00473019	180.53	0.0001
PSOS92	0.13122382	0.05178650	0.00016824	6.42	0.0116
K1	-0.00177213	0.00061846	0.00021513	8.21	0.0044
ENSMOR93	0.03548219	0.00517031	0.00123400	47.10	0.0001

Bounds on condition number: 57.52923, 509.3079

All variables left in the model are significant at the 0.1000 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable PSOS94

Step	Variable Entered Label	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	PSOS93		1	0.9769	0.9769	62.8593	18261.1085	0.0001
2	ENSMOR93		2	0.0022	0.9791	18.3972	44.7593	0.0001
3	K1		3	0.0004	0.9795	11.7774	8.4473	0.0038
4	PSOS92		4	0.0003	0.9798	7.3071	6.4208	0.0116

Generelt kan vi se et mønster i de trinnvise regresjonskjøringene. Det virker som om sosialstønadstallene i året før, i «uttransformert» form, er den beste responsvariabelen. Det er derfor naturlig å bruke en ratestimeringsmodell hvor dette kjennemerket blir brukt i utgangspunktet. Derfor er vår referansemodell, «den enkle modellen» hvor vi bruker konvensjonell univariat ratestimering, i høyeste grad forsvarlig å bruke.

Som alternativ har vi valgt å bruke *multivariat* ratestimering ifølge Olkin's metode. Det finnes en rekke kjennemerker som i høy grad simultant forklarer behovet for sosialstønad i befolkningen. Generelt sett virker det som om antallet eller andelen sosialstønadstilfeller i årene før sammen med enslige mødre og kommunetype er best egnet til slike prediksjoner. Vi synes av flere grunner at det er mest forsvarlig å bruke andelen sosialstønadstilfeller i 1993 og 1992 samt andelen enslige mødre i 1993 som responsvariable i en *multivariat* ratestimeringsmodell.

Begrunnelsen for våre to alternative modellvalg hviler på en vurdering, hvor de her presenterte trinnvise regresjonsutkjøringene utgjør *en* viktig del. Andre eksempler på kilder til vår begrunnelse er utkjøringen av tilsvarende regresjonsanalyser for året før og bruken av sosialstønad regnet i kronetall, hvor resultatene skiller seg noe ut dersom vi begynner å se på uvedkommende detaljer, men hvor helhetsbildet taler i retning av den valgte modellen. En «bokstavstolking» av de presenterte utkjøringene over hadde kanskje gitt oss litt andre konklusjoner (f.eks. at vi burde ta med en eller to dummyvariable for kommunetype i den multiple modellen, eller at den enkle modellen burde ha vært en regresjonsestimeringsmodell ettersom konstantleddet er signifikant), men vi har, på et godt erfaringsmessig grunnlag, brukt en noe «konservativ» fremgangsmåte i valget av estimeringsmodeller.

De sist utgitte publikasjonene i serien Notater

- 95/50 A.M. Kleive Holmøy: Dokumentasjonsrapport: Beregning av vektorer til Inntekts- og formuesundersøkelsene 1993. 21s.
- 95/51 V. Pedersen: Inntekts- og formuesundersøkelsen 1991: Dokumentasjon. 106s.
- 95/52 T. Kornstad: Simulering av konsum og arbeidstilbud i et livsløpsperspektiv. 32s.
- 95/53 E.M. Nielsen: Forsprosjekt for en løpende nasjonal kulturstatistikk. 28s.
- 95/54 E. Heilund: Utvalgstrekkning, usikkerhetsberegning og frafallsbehandling i inntekts- og kostnadsundersøkelsen for tannleger og fysioterapeuter. 43s.
- 95/55 B. Otnes: Egenbetaling for hjemmetjenester i kommunene i 1994 og 1995. 38s.
- 95/56 A. Langørgen: Faktorer bak kommunale variasjoner i utgifter til sosialhjelp og barnevern. 17s.
- 95/57 R.H. Kitterød: Funksjonsnivå og hjelpebehov blant brukere og ikke-brukere av pleie- og omsorgstjenesten i kommunene. 92s.
- 95/58 T. Wiersdalen Karlsen: Energimarkedet fra 1973 og fram mot 2010. 15s.
- 95/59 J.A. Sigstad Lie og L. Solheim: Gruppering av brukere av pleie- og omsorgstjenesten ved bruk av GERIX-kortet. 20s.
- 96/1 E. Vassenden: Ny lærerstatistikk (0340): Revidert dokumentasjon. 28s.
- 96/2 B. Olsen: Pleie- og omsorgstjenestene 1991-1993. 142s.
- 96/3 I.M. Smestad: Valg under usikkerhet: En analyse av eksperimentdata basert på kvalitative valghandlingsmodeller. 58s.
- 96/4 Mot et nytt system for undersøkelser av levekår: Innstilling fra en prosjektgruppe. 62s.
- 96/5 E. Nordhagen Karlsen (SSB) og S. Nestvold (SHD): Sosiale utgifter 1980-1993: Dokumentasjon av excel-datasett beregnet både på rapportering til NOSOSKO, OECD og EUROSTAT og til nasjonal statistikk. 106s.
- 96/6 M. Vik Dysterud og P. Schøning: SSB-AVLØP: Fylkesrapport 1994. 189s.
- 96/7 B. Strand: Kobling av adresseregistrene i DSF og GAB: Dokumentasjon og resultater. 34s.
- 96/8 B. Lian og K.O. Aarbu: Dokumentasjon av LOTTE-AS. 43s.
- 96/9 D. Fredriksen: Datagrunnlaget for modellen MOSART, 1993. 36s.
- 96/10 S. Grepperud og A.C. Bøeng: Konsekvensene av økte oljeavgifter for råoljepris og etterspørsel etter olje: Analyser i PETRO og WOM. 12s.
- 96/11 T. Tørstad: Bruk av Grunneiendoms-, Adresse- og Bygningsregisteret i Statistisk sentralbyrå. 39s.
- 96/12 A.C. Bøeng, O.T. Djupskås og E. Hoffart: Energistatistikk: Dokumentasjon av produksjonsrutiner. 65s.
- 96/13 B. Otnes: Statistikk om egenbetaling: Forprosjekt. 37s.
- 96/14 L. Rogstad, R. Jule, T. Vik og J.E. Vålberg: Samordnet bruk av GAB-data i SSB. 47s.
- 96/15 Å. Kaurin, E. Vinju og L. Solheim: Statistikk over avfall og gjenvinning fra deler av offentlig virksomhet. 56s.
- 96/16 K. Gerdrup: Inntektsfordeling og økonomisk vekst i norske fylker: En empirisk studie basert på data for perioden 1967-93. 45s.
- 96/17 E. Vassnes: Evaluering av arbeidsmarkeds-tiltak - bruk av registerdata. 58s.
- 96/18 A.C. Bøeng: Prisutvikling på olje ved ulike forutsetninger om utviklingen i eksogene variable: Analyser i WOM-modellen. 24s.
- 96/19 A.C. Steen: Inntekts- og kostnadsundersøkelsen for privatpraktiserende fysioterapeuter: Dokumentasjon. 67s.

Statistisk sentralbyrå

Oslo
Postboks 8131 Dep.
0033 Oslo

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

Kongsvinger
Postboks 1260
2201 Kongsvinger

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

ISSN 0806-3745



Statistisk sentralbyrå
Statistics Norway