

# Arbeidsnotater

S T A T I S T I S K S E N T R A L B Y R Å

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

0 77/7

15. februar 1977

STIANALYSE I SOSIOLOGISK FORSKNING:  
ET EKSEMPEL FRA LEVEKÅRSUNDERSØKELSEN 1973

av

Petter Laake og Arne S. Andersen<sup>\*)</sup>

## INNHold

	Side
1. Innledning .....	1
2. Generelt om stianalyse .....	2
3. Et eksempel på en stianalyse: Bruk av legetjenester .....	7
3.1. Oversikt over begreper og variable i undersøkelsen av bruk av legetjenester .....	8
3.2. En modell for bruk av legetjenester .....	9
4. Beskrivelse av utvalg og data .....	11
5. Resultater .....	12
6. Diskusjon av resultater .....	18
Referanser .....	20
Vedlegg: Estimering av parametre i lineære strukturlikninger .....	22

\*) Vi vil takke spesielt Sigurd Høst og Ib Thomsen for deres konstruktive merknader til et første utkast av dette notatet. Kommentarene har ført til betydelige endringer i det opprinnelige utkastet.

## 1. Innledning

Analyse av sosiologiske undersøkelser krever vanligvis at en bruker mange ulike teknikker for å sammenfatte dataene i en enkel og tolkbar form. Slike metoder kan variere fra konstruksjon av enkle tabeller eller grafiske framstillinger til mer kompliserte statistiske analyseteknikker.

I dette notatet vil vi studere en av de statistiske analyseteknikker som har til formål å analysere ikke-eksperimentelle data. Bakgrunnen for at vi er interessert i å studere en slik metode er selvfølgelig at data som framkommer i en sosiologisk undersøkelse, nettopp er ikke-eksperimentelle. De vanligst brukte metodene, slik som kontingenstabellanalyse og regresjonsanalyse, er brukt i mange sammenhenger i Byrået, og de har til formål å studere avhengigheten eller samvariasjonen mellom de variablene vi studerer. I dette notatet vil vi studere regresjonsmetoden litt nærmere og prøve å begrunne at regresjonsanalyse ikke alltid vil være en egnet metode til å studere samvariasjonen mellom variable. Vi vil derfor studere andre metoder som etter vår mening vil være bedre egnet enn regresjonsmetoden til analyse av ikke-eksperimentelle data.

Bakgrunnen for at sosiologer ofte er interessert i å studere slike modeller er at det ut fra sosiologisk teori ofte kan være mulig å etablere årsaks- virkningsrelasjoner mellom de variablene en studerer. I motsetning til den klassiske regresjonsanalysen, der en studerer samvariasjonen mellom én avhengig venstreside variabel og et sett av høyreside variable, vil bruken av årsaks- virkningsrelasjonene føre til at vi bør studere hele settet av variable simultant uten å ta ut én spesiell avhengig variabel. En vanlig måte å studere variablene simultant på, er å anta at vi kan uttrykke sammenhengen mellom variablene som et sett av lineære likninger, noe som også vil vise seg å være et naturlig utgangspunkt for å studere samvariasjonen mellom alle variablene.

Tidlig ble det interesse for slike modeller i økonomisk forskning, der en kalte slike likningssystemer for simultane likningssystemer eller lineære strukturlikninger. I de siste årene har det blitt en økt interesse for bruk av lineære strukturlikninger også i sosiologisk forskning. Sosiologer har kalt denne metoden for stianalyse ("path analysis"), et uttrykk som ble innført av Wright (1934). Studerer en den nyere sosiologiske litteraturen, virker det som om mange har blitt inspirert av arbeidene til Blau & Duncan (1967) og Duncan (1966), og det finnes nå mange anvendelser innenfor forskjellige problemområder. Metoden har også blitt anvendt innenfor andre sosialvitenskaper, for anvendelser i psykologi kan en eksempelvis se Werts & Linn (1970).

Hvis vi allerede nå skal prøve å oppsummere det som stianalyse omfatter, kan en si at en på grunnlag av sosiologisk (evt. økonomisk, psykologisk o.l.) teori setter opp hvilke årsakssammenhenger som er rimelige. På grunnlag av dette lager en så en nærmere spesifisering av sammenhengene i form av et simultant likningssystem og en foretar så til slutt de numeriske beregningene av parametrene i den foreslåtte modellen. Vi vil imidlertid presisere at stianalyse ikke er en metode til å oppdage årsakssammenhenger, men en metode til å kvantifisere effektene i et gitt kausalt system.

I dette notatet vil vi prøve å gi en innføring i de vesentlige idéene i stimodeller. Vi vil bare studere modeller der sammenhengen mellom variablene kan uttrykkes som et sett av lineære likninger. Videre vil vi i det vesentlige begrense oss til modeller der "den kausale ordningen av variablene går i én retning".

I tillegg til dette er formålet også å gi et eksempel på en stimodell til analyse av sosiologiske data. Vi vil konstruere en kausal modell for bruk av legetjenester, og har tatt utgangspunkt i tankegangen i en modell av Andersen et. al. (1970). Vi har lagt vekt på å undersøke om det er andre faktorer enn behovsfaktorer, altså personenes helsetilstand, som har vesentlig betydning for bruk av legetjenester. Vi har derfor lagt vekt på hvilke helsevariable som inngår i modellen. De fleste fler-variabelanalyser av bruken av legetjenester har blitt foretatt i USA. Forskjellene mellom helsetjenestene i USA og Norge har ført til visse forskjeller i utformingen av modellen. Eksempelvis inngår inntekt og forskjellige forsikringssystemer som en viktig del av flere analyser foretatt i USA, men er ikke så aktuelle i Norge. Vi har valgt ikke å bruke inntekt, men lar sosial status inngå i modellen. Denne variabelen vil kunne avspeile virkninger av forskjeller i materielle ressurser og av eventuelle kulturelle forskjeller i bruken av legetjenester.

I den konstruerte modellen vil parametrene bli estimert ved hjelp av EDB-programmet LISREL, som er utviklet av Jøreskog & van Thillo (1973), for data fra levekårsundersøkelsen (Statistisk Sentralbyrå (1975)).

I avsnitt 2 vil vi gi en oversikt over noen enkle stimodeller. Til tross for at modellene er enkle håper vi at en del av hovedidéene i stianalyse kan bli klarlagt. Avsnittene 3, 4, 5 og 6 omhandler analysen av legetjenester i Norge. I vedlegget har vi gitt en formalisering av teorien i avsnitt 2, og av estimeringen av parametre i modellen. For å gjøre notatet så lett leselig som mulig for lesere uten bakgrunn i matematisk statistikk, er denne delen tatt med som vedlegg.

## 2. Generelt om stianalyse

Som nevnt innledningsvis er formålet med en stianalyse å studere samvariasjonen mellom interkorrelerte variable. For en del formål har det vist seg at regresjonsanalyse har vært en viktig analyseteknikk når en vil studere slik samvariasjon.

I analyse av ikke-eksperimentelle data kan imidlertid regresjonsanalyse være en utilfredsstillende måte å analysere data på. For å begrunne denne påstanden vil vi kommentere regresjonsmodellen litt nærmere. I en enkel regresjonsmodell uttrykker regresjonen av en variabel  $y$  m.h.p. en variabel  $x$  den betingete forventningen av  $y$  som en funksjon av  $x$ . Dersom en altså ønsker å studere hvordan den betingete forventningen til  $y$  varierer med  $x$ , vil regresjonsanalyse være en egnet metode. I mange tilfeller vil en imidlertid uttrykke regresjonen av variabelen  $y$  m.h.p. et sett av variable  $x_1, x_2, \dots, x_p$ . I slike tilfeller vil tolkningen av regresjonsanalysen bli komplisert. Regresjonskoeffisientene tolkes nå nemlig som effekten av en endring på én variabel når de øvrige  $x$ -variablene i modellen holdes konstant. I en ikke-eksperimentell situasjon kan imidlertid ikke variablene kontrolleres i forsøksituasjoner, og dette gjør det vanskelig å finne en rimelig tolkning av effekten som de ulike  $x$ -variablene har på  $y$  innen en regresjonsmodell. Til nå har vi latt begrepet effekt være knyttet til tolkningen av samvariasjonen mellom én variabel og et sett forklaringsvariable. I sosialvitenskapene snevrer en ofte inn begrepet effekt ved å la det være knyttet til tolkningen av årsaks- virkningssammenhenger mellom variablene. En ønsker da å utvide regresjonsmodellene til modeller der en kan studere hvordan alle variablene i variabelsettet påvirker hverandre for at en ved å bruke slike modeller kan få en enklere tolkning av hvilke effekter  $x$ -variablene har på  $y$ . Til tross for innvendingen mot de klassiske regresjonsmodellene, vil selvfølgelig slike modeller likevel være egnet i en rekke situasjoner, f.eks. der en nettopp ønsker å predikere den betingete forventningen til  $y$  for gitte verdier av  $x_1, x_2, \dots, x_p$ .

Når vi nå vil prøve å utdype idéene i det foregående litt, går vi først tilbake til bakgrunnen for behovet for nye metoder til analyse av ikke-eksperimentelle data. Sosiologer ønsker som sagt tidligere, gjerne å studere årsaksstrukturen i et sett variable. I det følgende vil vi da anta at strukturen mellom variablene kan uttrykkes ved et sett av lineære likninger. En mulig spesifisering av et slikt samband vil da være å spesifisere at hver variabel er direkte påvirket av alle de andre variablene i systemet. I et slikt likningssystem vil da hver variabel opptre både som høyre- og venstresidevariable i likningene. I de fleste tilfeller vil en imidlertid kunne anta at noen av koeffisientene i likningene er lik null. Slike antagelser må bygges på a priori kunnskap om de kausale ordningen mellom variablene i systemet. I dette notatet vil vi ikke gå nærmere inn på den formelle innføringen av begrepet kausalitet, men bare henvise interesserte lesere til f.eks. Simon (1953).

Vi vil derimot studere et enkelt eksempel på en modell mellom de fire variablene  $x_1, x_2, x_3$  og  $x_4$ .

Vi antar i det følgende at variablene er standardisert slik at de har forventning lik 0 og varians lik 1. Dette er selvfølgelig ikke en nødvendig antagelse i slike modeller. Vi har bare antatt dette for å forenkle tolkningen av de ulike effektene. Dersom vi, som over, antar at hver av variablene blir direkte påvirket av de øvrige og i tillegg antar at i hver relasjon er venstresidevariabelen ikke fullstendig bestemt av kombinasjonen av høyresidevariable, vil likningssystemet bli spesifisert ved

$$\begin{aligned}
 x_1 &= p_{12} x_2 + p_{13} x_3 + p_{14} x_4 + \varepsilon_1, \\
 x_2 &= p_{21} x_1 + p_{23} x_3 + p_{24} x_4 + \varepsilon_2, \\
 x_3 &= p_{31} x_1 + p_{32} x_2 + p_{34} x_4 + \varepsilon_3, \\
 \text{og} \\
 x_4 &= p_{41} x_1 + p_{42} x_2 + p_{43} x_3 + \varepsilon_4,
 \end{aligned}
 \tag{2.1}$$

der p-ene er ukjente parametre, og  $\varepsilon$ -ene er residualvariable. Ved å ordne variablene og dividere hele likningen med passende parametre, ser vi at alle likningene i systemet er identiske, og det er derfor umulig å "skille" mellom dem. I hver enkelt analyse må en derfor bruke å priori kunnskap til å spesifisere enkelte av parametrene i (2.1). En mulig spesifisering er å anta at visse parametre er lik 0.

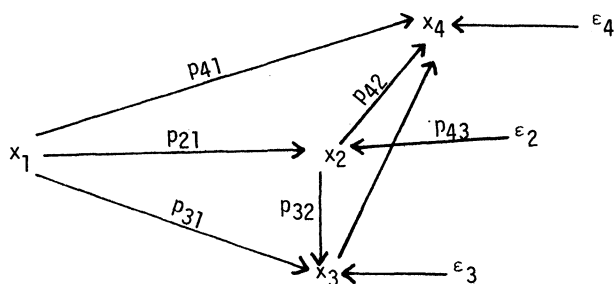
La oss nå, som et eksempel på en a priori spesifisering av parametre, anta at modellen (2.1) kan reformuleres til

$$\begin{aligned}
 x_2 &= p_{21} x_1 + \varepsilon_2, \\
 x_3 &= p_{31} x_1 + p_{32} x_2 + \varepsilon_3, \\
 \text{og} \\
 x_4 &= p_{41} x_1 + p_{42} x_2 + p_{43} x_3 + \varepsilon_4.
 \end{aligned}
 \tag{2.2}$$

Vi ser at siden noen av parametrene i likningssystemet (2.1) er satt lik 0, vil ikke alle variablene i systemet påvirke de øvrige. Vi antar videre at residualene  $\varepsilon_2$ ,  $\varepsilon_3$  og  $\varepsilon_4$  er stokastisk uavhengig av  $x_1$ , og at residualene er innbyrdes uavhengige. Den intuitive tolkningen av (2.2) blir da at variabelen  $x_2$  ikke påvirker  $x_1$ ,  $x_3$  påvirker hverken  $x_1$  eller  $x_2$ , og  $x_4$  påvirker ingen av variablene  $x_1$ ,  $x_2$  og  $x_3$ . Hver av de tre likningene i (2.2) representerer altså en relasjon mellom variablene som karakteriserer strukturen i sammenhengen mellom variable. Dette begrunner da navnet på slike likningssett, strukturlikninger. Som nevnt innledningsvis skiller slike modeller seg fra regresjonsmodellen i det vi (2.1) og (2.2) har spesifisert et sett av lineære relasjoner mellom variablene. Analogt med det som er vanlig i regresjonsanalyse, vil vi kalle variablene i likningene for henholdsvis venstreside- og høyresidevariable.

Vi ser at den kausale ordningen mellom variablene i (2.1) er slik at enhver variabel står i en ikke-tilbakevirkende relasjon med de andre variablene. Enhver variabel som vi altså bringer inn i modellen, kan være en effektvariabel til de andre variablene, men den kan ikke samtidig være en årsaksvariabel til de samme variablene. Antagelsene om at residualene er uavhengig av  $x_1$  og at residualene er innbyrdes uavhengige, fører til at enhver residual er uavhengig av alle de øvrige variablene i den likningen der residualen opptrer. Når disse antagelsene er oppfylt og i tillegg at residualene er innbyrdes uavhengige sier vi at likningssystemet er hierarkisk og at modellen er rekursiv.

Stimodeller blir ofte representert ved et diagram. I et slikt diagram betegner enveispiler fra en variabel til en annen at den ene variabelen er en årsaksvariabel, og den andre variabelen er en effektvariabel. De rette linjene betegner altså en direkte påvirkning fra en variabel til en annen. Symbolene over pilene "måler" denne effekten og kalles stikoeffisienter. Alle disse symbolene har dobbel fotskrift. Den første fotskriften betegner variabelen i hodet av pilen (effektvariabelen), og den andre betegner årsaksvariabelen. Korrelasjoner som vi lar være uanalysert, betegnes vanligvis ved en krum dobbeltpil. I et stidiagram blir modellen (2.2) symbolisert slik:



(2.3)

I diagrammet (2.3) ser vi at en forandring i variabelen  $x_1$  gir en direkte forandring i variablene  $x_2$ ,  $x_3$  og  $x_4$ . I tillegg vil vi få en indirekte påvirkning på  $x_3$  via variabelen  $x_2$  og en indirekte påvirkning på  $x_4$  via  $x_2$  eller  $x_3$ . En forandring i  $\epsilon_2$  vil forandre  $x_2$  direkte og  $x_3$  og  $x_4$  indirekte. Ved å gå gjennom dette resonnementet for alle variablene, ser vi altså at den kausale ordningen mellom variablene er spesifisert slik at  $x_4$  ikke er årsak til noen andre variable,  $x_3$  er hverken årsak til  $x_1$  eller  $x_2$ , og  $x_2$  er ikke årsak til  $x_1$ . Kausaliteten i dette systemet går altså fra  $x_1$  til  $x_4$  uten noen tilbakevirkende kausaliteter.

Vi vender nå tilbake til strukturlikningene (2.2). Siden variablene er standardiserte, innser vi at korrelasjonen mellom  $x_i$  og  $x_j$  kan uttrykkes som

$$\rho_{ij} = \sum_k p_{ik} \rho_{jk} \text{ for } i > j. \quad (2.4)$$

Eksempelvis kan altså korrelasjonen mellom  $x_1$  og  $x_4$  uttrykkes som

$$\rho_{41} = p_{41} + p_{42} \rho_{21} + p_{43} \rho_{31}. \quad (2.5)$$

Siden  $\rho_{21}$  og  $\rho_{31}$  på tilsvarende måte kan uttrykkes ved stikoeffisienter fra strukturlikningene, får vi altså at

$$\rho_{21} = p_{21} \rho_{11} = p_{21}, \quad (2.6)$$

og

$$\begin{aligned} \rho_{31} &= p_{31} + p_{32} \rho_{21} \\ &= p_{31} + p_{32} p_{21}. \end{aligned}$$

Ved å innføre (2.6) i (2.5), finner vi at

$$\rho_{41} = p_{41} + \underbrace{p_{21}p_{42} + p_{31}p_{43} + p_{21}p_{32}p_{43}}_{\text{indirekte effekt}} \quad (2.7)$$

direkte effekt

Fra diagrammet (2.3) og strukturlikningene (2.2) ser vi at korrelasjonen mellom  $x_1$  og  $x_4$  kan splittes i ett ledd som representerer den direkte effekten og ett ledd som representerer den indirekte effekten av  $x_1$  på  $x_4$ . Parameteren  $p_{41}$  "måler" den direkte effekten av  $x_1$  på  $x_4$ , og  $p_{21}p_{42}$  "måler" den indirekte effekten av  $x_1$  på  $x_4$  gjennom  $x_2$ . Tilsvarende representerer  $p_{31}p_{43}$  effekten som  $x_1$  har på  $x_4$  gjennom  $x_3$  og  $p_{21}p_{32}p_{43}$  effekten som  $x_1$  har på  $x_4$  ved at  $x_1$  har virkning via variablene  $x_2$  og  $x_3$ .

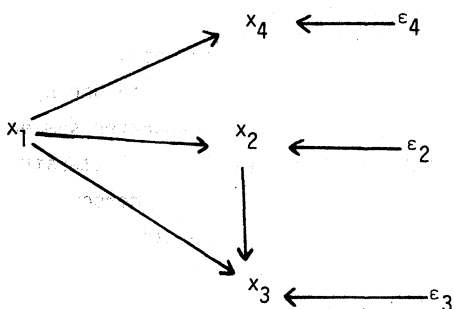
Analogt med bruken av (2.4) til å uttrykke korrelasjonen mellom  $x_1$  og  $x_4$  finner vi at korrelasjonen mellom  $x_2$  og  $x_4$  kan uttrykkes som

$$\rho_{42} = p_{42} + p_{43}p_{32} + p_{41}p_{21} + p_{43}p_{31}p_{21}. \quad (2.8)$$

I denne likningen uttrykker  $p_{42}$  den direkte påvirkningen av variabelen  $x_2$  på  $x_4$ . Hvert enkelt av de andre leddene på høyre side i likningen finner man ved å gå tilbake fra variabel  $x_4$  langs alle mulige stier og derfra fram til variabel  $x_2$  langs alle stier som fører til  $x_2$ . Duncan (1966) kaller de tre siste leddene på høyre side i (2.8) for indirekte effekter. Vi innså at det var naturlig å kalle de tre høyresideleddene i (2.8) for indirekte effekter innser vi på følgende måte: Leddet  $p_{43}p_{32}$  forklares ved at  $x_3$  er en "mellomliggende" årsaksvariabel for  $x_2$  og  $x_4$ , og det representerer altså den indirekte effekten av  $x_2$  på  $x_4$ . Leddet  $p_{41}p_{21}$  er derimot den delen av korrelasjonen mellom  $x_2$  og  $x_4$  som skyldes at  $x_1$  er årsaksvariabel for både  $x_2$  og  $x_4$ . Tilsvarende representerer  $p_{43}p_{31}p_{21}$  den delen av korrelasjonen som skyldes at  $x_1$  også er årsaksvariabel for  $x_3$  som igjen er årsaksvariabel for  $x_4$ . Det er derfor naturlig å kalle leddene  $p_{41}p_{21}$  og  $p_{43}p_{31}p_{21}$  for virkningene av en bakenforliggende årsaksvariabel. Det er altså bare ett av leddene på høyre side i likheten (2.7) som representerer den indirekte effekten av  $x_2$  på  $x_4$ . Visuelt vil dette framgå dersom vi studerer stidiagrammet (2.3). En grundigere innføring i oppsplittingen av den totale effekten i direkte og indirekte effekter og effekter av bakenforliggende variable er gitt av f.eks. Finney (1972) eller Alwin & Hauser (1975).

Likning (2.4) er et av hovedteoremene i stianalyse, og det viser hvordan korrelasjonen mellom variablene kan splittes opp i komponenter som representerer effekten av de ulike variablene i strukturlikningene.

I eksempel (2.2) har vi antatt at det finnes direkte påvirkninger mellom alle variablene i den gitte ordningen. I en del tilfeller kan en ut fra den teorien en behandler i tillegg anta at visse variable i ordningen ikke er direkte avhengig av andre variable i ordningen. Hvis vi eksempelvis antar at  $x_2$  ikke har noen direkte påvirkning på  $x_4$ , vil  $p_{42} = 0$ . La oss i tillegg anta at også  $p_{43} = 0$ , som betyr at  $x_3$  heller ikke har noen direkte påvirkning på  $x_4$ . Stidiagrammet for denne situasjonen er da



(2.9)

Då gir (2.8) at

$$\rho_{42} = p_{41}p_{21}, \quad (2.10)$$

som sammen med (2.5) og (2.6) gir at

$$\rho_{42} = \rho_{41}\rho_{21}. \quad (2.11)$$

Uten at det er nødvendig å anta noe om fordelingen til  $x_1$ ,  $x_2$  og  $x_4$ , vil den partielle korrelasjonen mellom  $x_2$  og  $x_4$  gitt  $x_1$ , kunne uttrykkes ved

$$\rho_{42} - \rho_{41} \rho_{21} = \frac{\rho_{42} - \rho_{41} \rho_{21}}{\sqrt{1-\rho_{41}^2} \sqrt{1-\rho_{21}^2}},$$

som ved innsetting av (2.11) gir at

$$\rho_{42} - \rho_{41} \rho_{21} = 0.$$

Til nå har vi antatt at residualene er ukorrelerte. La oss nå anta at  $\epsilon_2$  og  $\epsilon_4$  er korrelerte. Da kan det vises at korrelasjonen mellom  $\epsilon_2$  og  $\epsilon_4$  kan uttrykkes som

$$\rho_{\epsilon_1 \epsilon_2} = \frac{\rho_{42} - \rho_{41} \rho_{21}}{\sqrt{1-\rho_{41}^2} \sqrt{1-\rho_{21}^2}},$$

slik at

$$\rho_{\epsilon_1 \epsilon_2} = \rho_{42} - \rho_{41} \rho_{21}$$

Hvis vi i en analyse finner at  $\rho_{42} - \rho_{41} \rho_{21} \neq 0$ , betyr dette at modell (2.9) ikke tilpasser dataene tilfredsstillende. Vi kan da endre modellen ved å innføre korrelasjon mellom residualene  $\epsilon_2$  og  $\epsilon_4$ . En mulig forklaring på at residualene er korrelerte er at det finnes variable som ikke er tatt med i spesifiseringen av relasjonene for  $x_2$  og  $x_4$ . Slike variable vil være medvirkende faktorer for korrelasjonen mellom residualene.

En annen reformulering av modellen (2.9) kan være at en lar variable som allerede er med i systemet, være med i spesifiseringen av relasjonene for  $x_2$  og  $x_4$ . En mulig reformulering kan da være å anta at  $\rho_{42} \neq 0$  og  $\rho_{43} \neq 0$ . Når vi innfører stikoeffisientene  $\rho_{42}$  og  $\rho_{43}$ , er dette ekvivalent med en kausal tolkning av at  $\rho_{42} - \rho_{41} \rho_{21} \neq 0$ . Den alternative modellen, der en antok at residualene var korrelerte, gir vi ingen kausal tolkning. Valget mellom alternative modeller kan derfor ofte bli avgjort av i hvilken grad en har a priori kunnskap om relasjonene mellom variablene.

I dette avsnittet har vi antatt at alle variablene er standardiserte med forventning lik 0 og varians lik 1. Det har i lengre tid pågått en diskusjon i litteraturen om en i stianalyse bør nytte standardiserte eller ustandardiserte variable. Det synes å være to motstridende synspunkter som er framlagt av eksempelvis Tukey (1954) og Blalock (1967). Det mest vanlige i sosiologisk litteratur synes å være at en bør bruke standardiserte variable og standardiserte stikoeffisienter dersom en ønsker å lage inferens om en bestemt bestand. Dersom en imidlertid ønsker å sammenlikne bestander for å bestemme om de kausale sammenhenger i bestandene er identiske, bør en bruke ustandardiserte variable.

I et senere arbeid av Mayer & Younger (1976) viser imidlertid forfatterne at estimatorene for standardiserte stikoeffisienter er inkonsistente. Stikoeffisientene som bygger på standardiserte variable og som er mest vanlig i sosiologisk litteratur, er altså ikke "gode" estimater for de standardiserte koeffisientene. Vi har derfor valgt, slik som Tukey (1954) anbefaler, å nytte ustandardiserte variable i den senere analysen.

Vi har til nå ikke beskrevet hvordan en estimerer parametrene i en stimodell. Siden estimeringsmetoden vil bli beskrevet langt mer teknisk enn de øvrige avsnittene i notatet, har vi valgt å la dette være med som et vedlegg. Vi vil imidlertid allerede nå orientere om at de numeriske beregningene er foretatt ved hjelp av EDB-programmert LISREL som er utarbeidet av Jøreskog & van Thillo (1973) og implementert for bruk i Byrået. Programmet gir sannsynlighetsmaksimeringsestimater for parametrene i generelle lineære strukturlikninger.

For en del stimodeller finnes det andre EDB-programmer som også er egnet til numeriske beregninger. Ett av disse er implementert i Byrået og beskrevet av Garaas (1975) i et arbeidsnotat. Siden LISREL bygger på estimering etter sannsynlighetsmaksimeringsprinsippet, vil imidlertid dette programmet etter vår mening være det som er best egnet til å teste og eventuelt videreutvikle modellene vi studerer. Denne bruken av programmet vil vi komme tilbake til i den senere dataanalysen. I et senere notat vil vi gi en oversikt over hvilke typer av modeller som kan behandles av LISREL og vise en del av anvendelsesområder for modellene.

I dette notatet vil vi ikke gå nærmere inn på begrepet identifiserbarhet. En innføring i identifikasjonsproblemet i simultane likningsystemer er gitt i f.eks. Amundsen (1973). Vi vil nøye oss med å orientere om at alle de relasjonene som vi har studert i dette avsnittet og som vi vil studere senere, vil gi identifiserbare parametre.

I dette avsnittet har vi diskutert noen av hovedidéene i stianalyse, og vi oppsummerer nå de antagelsene som har vært nødvendige å gjøre for å få en enkel tolkning av samvariasjonen mellom variablene:

- (i) relasjonene mellom variablene er lineære i parametrene,
  - (ii) det finnes ingen tilbakevirkende årsakssammenhenger mellom variablene,
  - (iii) variablene er uten målefeil,
  - (iv) residualene er uavhengige av høyresidevariablene i hver likning der residualene opptrer,
  - (v) residualene har forventning null og kovariansmatrisen er diagonal.
- og

Noen av disse antagelsene er ikke nødvendige for å utføre en generell stianalyse. I den mest generelle formen er det nødvendig bare å ha oppfylt (i) og (iv), (iv) da i en noe annen versjon, og i tillegg at alle parametrene er identifiserbare, mens (ii), (iii) og (v) er unødvendige.

### 3. Et eksempel på en stianalyse: Bruk av legetjenester

Som eksempel på en konkret stianalyse har vi valgt å utprøve en modell for bruk av legetjenester. Det eksisterer en ganske lang forskningstradisjon omkring bruk av legetjenester. En har også sett forsøk på å stille opp modeller for bruk av legetjenester, og stianalyse har flere ganger vært nyttig (Anderson (1973), Galvin & Fan (1975), Wan & Soifer (1974)). Det finnes i materialet fra levekårsundersøkelsen tilstrekkelig mange av de variable som en har vist er viktige faktorer i bruken av legetjenester, til at det er mulig å sette opp en meningsfull modell.

Studier av bruken av legetjenester utgjør bare en del av legetjenesteforskningen. I undersøkelser av bruken av legetjenester er en opptatt av omfanget av forbruk av legetjenester totalt og i noen grad også av forbruket av ulike typer legetjenester. I forskningen omkring legetjenester forøvrig har en vært opptatt av kvaliteten av helsetjenester, av effekten av den behandling legene gir, en har vært opptatt av hvilke funksjoner legetjenesten faktisk har, av forholdet lege - pasient o.a.

En har til nå samlet en ganske omfattende viten om bruken av legetjenester. En kjenner samvariasjonen mellom en lang rekke variable og bruken av legetjenester (Aday (1972)), og vet en del om ulikheten i bruk mellom ulike grupper i samfunnet.

Utover den generelle problemstilling å finne fram til hvilke faktorer som virker inn på bruken, og hvordan denne innvirkningen skjer, synes den problemstilling som egentlig ligger til grunn, å være spørsmålet om over- og underforbruk av legetjenester. Det finnes neppe noen norm for hva som er riktig bruk av lege, og det er derfor vanskelig å formulere presise hypoteser om over- og underforbruk. Likevel ligger det til grunn for de fleste studier av bruk av legetjenester forestillinger om at det finnes visse faktorer: kulturelle, sosialpsykologiske, kostnadsmessige, organisatoriske, som skaper forskjeller i bruken av legetjenester mellom ulike grupper, og som vil virke hemmende på tilbøyeligheten til å søke lege i visse grupper. I USA har spesielt den positive sammenheng en har funnet mellom variable som inntekt, sosioøkonomisk status, sosial klasse og bruk av legetjenester vært viet stor oppmerksomhet (Monteiro (1973), Galvin & Fan (1975), Bice et.al. (1972)).



### 3.1. Oversikt over begreper og variable i undersøkelsen av bruk av legetjenester

Den avhengige variabel, forbruket av legetjenester, har blitt målt som en binær variabel: kontakt, ikke kontakt med lege i et visst tidsrom. De fleste synes imidlertid å ha brukt tallet på kontakter i et visst tidsrom som mål. Bice bemerker at undersøkelser av bruk av legetjenester i de fleste tilfeller har hatt lite differensierte mål på legekontakt. En har ikke skilt mellom legekontakt med forskjellig formål (diagnose og behandling, vaksineringskontroll o.l.), og heller ikke etter kontaktsted eller -måte (legekontor, hjemmebesøk, legevakt, telefonkonsultasjon) (Bice et. al. (1972)).

Når det gjelder hvilke faktorer som virker inn på bruk av legetjenester, er det langt fra avklart hvilke enkeltfaktorer som er viktige i en forklaring av bruk, og det synes heller ikke å være enighet om hvilken vekt en skal tillegge ulike typer av faktorer (Berkanovic & Reeder (1974)).

Det er gjort flere forsøk på å klassifisere faktorer som virker inn på bruk av legetjenester. Vi skal her følge klassifikasjonen til Andersen et.al. (1970).

1. Behovsvariable. For de legekontakter som har diagnose og behandling som formål, er det klart at personens helsetilstand er viktig. Det er langt fra alle undersøkelser av legebruk som nytter mål for helsetilstanden, men i studier der en forsøker å lage modeller for bruk av legetjenester inngår vanligvis mål for helsetilstanden. Helsetilstanden er imidlertid et så komplisert fenomen at det neppe er mulig å beskrive den ved ett eller to mål. Bruk av bare ett mål på helsetilstand, f.eks. tallet på dager med nedsatt aktivitet p.g.a. sykdom, kan føre til tvilsomme utsagn og feilaktige slutninger. Som eksempel på dette, har en Monteiros (1974) diskusjon av legekontakter i lavinntektsgrupper og andre inntektsgrupper: "When illness is not present, i.e. when the respondent denies any restricted activity days, the lower income respondents report a higher utilization than the higher income respondents. Those lower income respondents who contact a physician when need is evidently not present .....". Han overser helt at sykdommer, f.eks. kroniske sykdommer ikke alltid resulterer i nedsatt aktivitet.

Som mål på helsetilstand har en foruten tallet på dager med nedsatt aktivitet på grunn av sykdom, nyttet: antall syketilfelle fra en gitt liste over sykdommer (Andersen et.al. (1970)), respondentens egen vurdering av sin totale helsetilstand (Wan & Soifer (1974)), og antall syketilfelle oppstått i en gitt periode (Galvin & Fan (1975)). Behovet for ulike helsetjenester har en forsøkt å estimere ved å tildele behovsvekter til et stort antall sykdomsgrupper (Kalimo & Sievers (1968)).

2. Den andre klassen av faktorer som mange betrakter som viktige for bruk av legetjenester, er kulturelle og sosialpsykologiske faktorer. En har undersøkelser av sammenhengen mellom legekontakter og et meget stort antall av slike variable (Aday (1972), McKinlay (1972)). Vi skal ikke her gå inn på resultatene av disse undersøkelsene, men gi enkelte eksempler på variable som har vært undersøkt. Virkningen av gruppeforskjeller i definisjon og vurdering av symptomer, symptomsensitivitet (Anderson & Bartkus (1973)), forskjeller i kunnskap om og holdning til egen helse, holdninger til helsevesenet, virkninger av kulturelle forskjeller mellom lege og pasient, av venners og naboers kunnskap om og holdninger til helse har vært undersøkt. En svakhet ved mange undersøkelser av slike variable er at en ikke har sett virkningen av disse variable på legekontakt i sammenheng med virkningen av variable av andre typer. I en gjennomgang av sammenhengen mellom legekontakter og sosioøkonomisk status konkluderer Bice et.al. (1972) med at eksisterende undersøkelser ikke tyder på at kulturelle og sosialpsykologiske faktorer kan forklare sammenhengen mellom legekontakter og sosioøkonomisk status.

3. Den tredje type faktorer er de som knytter seg til individets situasjon og organiseringen av helse-tjenester. Dette er dels faktorer i individets situasjon som påvirker det en i vid forstand kan kalle kostnadene ved legebesøk. I de amerikanske undersøkelsene blir inntekt ansett for den viktigste av disse faktorene, men også faktorer som familiestørrelse og avstand til lege har vært trukket inn. Av de organisatoriske faktorer som har vært undersøkt er det først og fremst ordningen med betaling av legetjenester, om personen har sykeforsikring eller går inn under offentlige ordninger for legehjelp (Medicare, Medicaid) som har blitt undersøkt. Virkningen av en organisasjonsmessig kompleks helse-tjeneste på bruken av legetjenester er vanskelig å påvise i utvalgsundersøkelser. Det blir likevel referert at inntrykket av en rekke mer kvalitative undersøkelser er at det er problemer av organisasjonsmessig karakter med å få legehjelp (Bice et.al. (1972)).

4. Sosiodemografiske variable. Variable som alder, kjønn, husholdningsstørrelse inngår vanligvis i undersøkelser av legekontakt. Slike variable inngår i beskrivelsen av bruken av legetjenester i viktige grupper i samfunnet. En del slike variable inngår imidlertid også i modeller for bruk av legetjenester.

### 3.2. En modell for bruk av legetjenester

Vi skal ta tankegangen i en modell av Andersen et.al. (1970) som utgangspunkt for formulering av en modell for legekontakt. Tankegangen er at en person med et behov for legehjelp vil søke slik hjelp avhengig dels av en rekke kulturelle og sosialpsykologiske faktorer ("predisposing factors") som bestemmer "tilbøyeligheten" til å søke legehjelp, dels av en rekke situasjonsfaktorer og organisasjonsmessige faktorer ("enabling factors"), som er avgjørende for de reelle kostnader (tid, penger, fysisk og psykisk energi m.m.) ved å søke legehjelp.

Vi skal legge vekten på behovsfaktoren og på kostnadsfaktorer. Vi har ikke noen direkte mål for disponerende kulturelle og sosialpsykologiske faktorer. Problemstillingen blir derfor om kostnadsfaktorer har noen virkning på legekontakt for et gitt behov for legehjelp, og eventuelt hvilke kostnadsfaktorer som er viktige. Blir fordelingen av legetjenester i samfunnet alene bestemt av folks behov for legetjenester, eller er det andre faktorer i tillegg som skaper ulikhet i bruken av slike tjenester?

#### 1. Kontakt med lege

Bruken av legetjenester måles her ved tallet på kontakter med lege i løpet av en 12-månedersperiode. En har bare regnet med kontakter i forbindelse med egne sykdommer eller skader, og ikke kontakter i forbindelse med sykehusopphold, rutineundersøkelser o.l. Legekontaktene omfatter kontakter uten hensyn til kontaktsted og kontaktmåte.

#### 2. Behov for legetjenester

Vi skal bruke personens helsetilstand som mål på behovet for legehjelp. Helse er imidlertid et svært sammensatt og komplisert fenomen, og det finnes neppe noe enkelt mål på helsetilstanden som vil kunne sammenfatte alle måter helsen vil ha betydning for bruk av lege på.

Den oppfatning av helse som ligger til grunn for vår modell ser en persons helsetilstand som sammensatt av tilstanden på tre områder: psykiske lidelser og langvarige og medfødte fysiske lidelser.

Data av den siste typen er praktisk talt ikke registrert i levekårsundersøkelsen, noe vi oppfatter som en stor svakhet for spesifiseringen av vår modell. Psykiske lidelser er representert ved en indeks som blir beskrevet senere (s. 11). Den andre variabelen som er nyttet til beskrivelse av helsetilstanden i modellen, er antall sykedager. Helseproblemer av alle tre typer vil kunne slå ut i sykedager, men vi antar at bare når det gjelder akutte fysiske lidelser, vil variabelen fange opp de fleste lidelser som vil medføre behov for lege.

Det er altså på forhånd klart at spesifisering av helsetilstanden i vår modell har en rekke svakheter.

#### 3. Kostnader ved bruk av legetjenester

Vi skal bruke tre mål for situasjonsfaktorer og organisasjonsmessige faktorer som har betydning for kostnadene ved legekontakt. Det er: reisetid til legekantor, arbeids- og familieforpliktelsener.

Den første faktor er reisetid til legekantor. Vi skal knytte tre kommentarer til bruk av reisetid til legekantor som mål på kostnaden i form av "tapt" tid ved bruk av lege.

For det første er reisetiden et mål på kostnaden ved bruk av lege i de tilfelle der personen besøker legen. Dersom besøk av legen som andel av alle legekontakter, øker med økende reisetid, vil det gjøre verdien av variabelen reisetid mindre.

For det andre vil reisetiden bare utgjøre en del av kostnaden i form av tapt tid ved legebesøk, også ventetiden vil telle. Dersom ventetiden er korrelert med reisetiden, vil også dette bidra til at reisetid ikke blir noe godt mål på tidskostnaden ved legebesøk.

En tredje kommentar gjelder forutsetningen i modellen om en lineær sammenheng mellom reisetid og bruk av lege. Det virker rimelig å anta at for reisetider under en viss grense vil reisetiden være uten betydning. I så fall vil effekten av reisetid ha form av en slags terskeeffekt, reisetiden får først betydning over en viss grense.

Den andre faktoren, arbeidsforpliktelser, er i modellen representert ved antall timer i uken i inntektsgivende arbeid. En antar at inntektsgivende arbeid, og spesielt arbeid på heltid, vil representere en hindring for kontakt med lege, bl.a. fordi legebesøk for mange lønnstakere vil innebære en direkte kostnad i form av tapt arbeidsfortjeneste. Systemet med syke- og friskmelding ved lengre sykdom vil kunne resultere i noe større legekontakt for lønnstakere enn for andre, og altså virke i motsatt retning av effekten ovenfor.

Den tredje faktoren er familieforpliktelser i form av pleie av barn, syke og gamle. Denne faktoren oppfattes på samme måte som arbeidsforpliktelser, som en hindring for bruk av lege. Som en grov indikator, bruker vi antall barn i husholdningen. En svakhet er det at antall barn stort sett bare vil være en hindring for kvinnene. En har her et tilfelle av en interaksjonseffekt mellom kjønn og antall barn.

Som tidligere nevnt har vi ikke noen direkte mål for disponerende faktorer. Spørsmålet er om noen av de variable som inngår i modellen kan betraktes som indirekte mål for kulturelle eller sosialpsykologiske faktorer som disponerer personen for å kontakte lege. Vi mener at det er svært nærliggende å anta at dette gjelder for sosial status. Konklusjonen av en gjennomgang av sammenhengen mellom sosioøkonomisk status og en del disponerende faktorer er imidlertid at det ikke er noe som tyder på at det finnes en slik sammenheng (Bice et.al. (1972)).

Av sosiodemografiske faktorer inngår bare kjønn og alder i modellen. Den vesentligste funksjon til disse variable er å være "kontrollvariable".

Vi skal til slutt kort ta opp muligheten for feilspesifiseringer av modellen. En form for feilspesifisering består i at den modellen som er spesifisert kan være korrekt, men bare for en del av populasjonen. Dersom legekontakter for en del av befolkningen må forklares etter en helt annen modell, vil forklaring av legekontakter i hele befolkningen med en modell i beste fall gi et forenklet bilde av virkeligheten, i verste fall vil en få et helt feilaktig bilde av bruken av legetjenester.

En undersøkelse gjennomført av en lege i Nord-Norge gjør det sannsynlig at modellen som er nyttet her, er for enkel (Telje (1974)). Det ble foretatt en spesiell legeundersøkelse av personer som ikke hadde vært i kontakt med lege de siste 5 årene. Undersøkelsen viste stor hyppighet av helseproblemer, også mer alvorlige helseproblemer. Gruppen er altså karakterisert ved et stort underforbruk av legetjenester.

I vårt materiale vil denne gruppen være en del av gruppen som ikke har vært i kontakt med lege de siste 12 månedér, en gruppe som for største delen ikke går til lege fordi de har god helse. En kan kanskje si at vår modell er egnet til å forklare forskjeller mellom antall legekontakter for personer som er tilbøyelige til å bruke lege dersom de blir syke. Dersom det finnes en gruppe av ekstreme underbrukere av legetjenester slik som Telje fant, vil en slik gruppe vanskelig kunne avsløres ved den analysen vi har gjennomført.

#### 4. Beskrivelse av utvalg og data

Analysen er gjennomført for data fra levekårsundersøkelsen i 1973.

Utvalget til levekårsundersøkelsen ble trukket tilfeldig blant personer 17 år og eldre som tilhørte de 4 707 husholdninger som ble oppsøkt for intervju i Forbruksundersøkelsen 1973. I alt ble det trukket ut 3 874 personer, og en oppnådde ontervju med 2 966 personer. For en nærmere beskrivelse av utvalgsmetode og frafall henvises leseren til publikasjon NOS A 720 (Statistisk Sentralbyrå (1975)).

Vi skal i dette avsnittet gi en nærmere beskrivelse av de variablene som inngår i modellen.

##### Legekontakt ( $X_1$ )

Alle personer i utvalget har blitt stilt spørsmål om antall ganger de har vært i kontakt med lege på grunn av egne sykdommer eller skader i løpet av de siste 12 månedene. En har ikke regnet med kontakt i forbindelse med sykehusopphold, rutineundersøkelser o.l. 52 prosent av de spurte hadde vært i kontakt med lege, og det største antall kontakter som ble oppgitt var 75 kontakter i løpet av de siste 12 månedene. Gjennomsnittlig antall legekontakter var 2,3.

Den lange registreringsperioden kan føre til systematiske avvik og dermed skjeve estimater. En kan imidlertid vurdere estimatene i levekårsundersøkelsen mot estimatene i Helseundersøkelsen 1968. Tallet på kontakter med lege utenom sykehus er der 72 pr. 1 000 personer i en 14-dagers periode. Antar en at en 14-dagers periode er representativ for årsgjennomsnittet, gir dette 1,9 kontakter pr. år pr. person. Dette er altså noe lavere enn det en fant i levekårsundersøkelsen, men samsvaret må sies å være tilfredsstillende.

##### Antall sykedager ( $X_2$ )

Dette er en enkel additiv indeks som er basert på svarene på tre spørsmål. Indeksen bygger på antall dager på sykehus, psykiatrisk klinikk, sykestue, eller annen helseinstitusjon i løpet av de siste 12 måneder, antall dager sengeliggende utenfor helseinstitusjon i samme periode og antall dager ute av stand til å utføre daglige gjøremål i samme periode, uten at man har vært sengeliggende. Disse tre opplysninger er summert til antall sykedager i løpet av de siste 12 måneder. Denne summering er foregått uten å gi forskjellige vektore til de ulike typene av aktivitetsnedsettelse.

##### Psykisk velvære ( $X_3$ )

Denne variabelen bygger på en indeks som er konstruert på grunnlag av svarene på fire spørsmål. Det gjelder spørsmål om en har vært plaget av tretthet i flere uker i løpet av de siste 12 måneder, om en har vært plaget av angst, uro, nervøsitet og depresjon i denne periode, om en ofte har vanskeligheter med å sove, og om en "ofte føler seg anspent i fritiden slik at en ikke får sluppet ordentlig av". En har talt opp hvor mange av disse plagene intervjupersonen har nevnt. Dette gir en indeks med verdier fra 0 til 4, med 4 for de med minst psykisk velvære. Som et mål på psykisk velvære kan en ikke regne med at variabelen oppfyller kravene til en intervallskala. En bør derfor være noe forsiktig med å tolke resultatet.

Vi finner det imidlertid rimelig å anta at variabelen psykisk velvære er en ordinalvariabel. En kan ikke regne med at resultatet er invariant overfor transformasjoner av måleenheten. En kunne selvfølgelig studere effekten av ulike transformasjoner for å få et inntrykk av hvor følsomt resultatet er overfor slike transformasjoner. Vi vil her nøye oss med en generell merknad ut fra formen på fordelingen til variabelen. Denne fordeling er svært skjev. Prosenten som har indeksverdier 0-4 er henholdsvis 68, 19, 8, 3 og 1 prosent. I likhet med et resonnement av Blau & Duncan (1967) i et tilsvarende tilfelle kan vi si at på grunn av den skjeve fordeling vil en endring av avstanden mellom de to høyeste variabelverdier måtte være meget drastisk for å påvirke resultatet vesentlig.

#### Arbeidstid ( $X_4$ )

Variablen angir den samlede arbeidstid i inntektsgivende arbeid, medregnet overtid og ekstraarbeid hjemme, i uken forut for intervju-tidspunktet. For midlertidig fraværende har en antatt en arbeidstid på 30 timer.

#### Antall barn i husholdningen ( $X_5$ )

Variablen angir antall barn under 16 år i husholdningen. Barna er ikke nødvendigvis intervju-personens barn, men kan også være barnebarn, søsken osv.

#### Sosial status ( $X_6$ )

Grupperingen etter sosial status bygger på en klassifisering utarbeidd ved Institutt for anvendt sosialvitenskapelig forskning (INAS) av Skrede (1971). Denne klassifiseringen bygger hovedsakelig på data om menn fra folketellingen i 1960. I klassifiseringen knyttes en statusverdi til hvert enkelt yrke. Inndelingen i sosialgrupper har tidligere vært nevnt i Byrået i den statistiske analysen av Yrke og dødelighet 1970-73, og leseren kan finne en nærmere beskrivelse av grupperingen i denne publikasjonen (Statistisk Sentralbyrå (1976)).

For sysselsatte menn og for sysselsatte kvinner med arbeidstid over 21 timer knytter sosial status seg til nåværende yrke. For andre kvinner knytter sosial status seg til ektefelles yrke. Dersom det ikke finnes noen ektefelle, blir kvinnens sosiale status bare bestemt dersom hun er sysselsatt (arbeidstid 1-21 timer).

Sosial status for skoleelever og studenter og personer i militærtjeneste knytter seg til hovedforsørgers yrke under oppveksten.

Sosial status for arbeidsuføre, pensjonister og personer uten arbeid knytter seg til eventuelt tidligere yrke.

Personer med ubestemt sosial status er holdt utenom analysen.

### 5. Resultater

Tabell 1 viser korrelasjonene mellom de variable som inngår i modellen for bruk av legetjenester. Korrelasjonene mellom tallet på legekontakter og de forskjellige variable er gjennomgående små. Korrelasjonene med helsevariablene skiller seg klart ut som de mest betydelige. Tallet på legekontakter korrelerer ikke med reisetid til legekontor, og korrelasjonene med arbeidstid og antall barn er også små.

Korrelasjonsmatrisen viser også en viss korrelasjon mellom de uavhengige variable. Spesielt korrelasjonene mellom de to helsevariablene og mellom alder og antall barn og mellom kjønn og arbeidstid er av en viss størrelse.

I alle modellene i denne analysen har vi transformert observasjonene av hver variabel slik at de er målt i forhold til sitt gjennomsnitt. Dette fører til at konstantleddet er utelatt i alle de modellene vi studerer.

Neste trinn i analysen er at vi etablerer en regresjonsmodell mellom antall legekontakter og de øvrige variablene. La da regresjonslikningen være gitt ved at

$$X_1 = \alpha_{12} X_2 + \alpha_{13} X_3 + \alpha_{14} X_4 + \alpha_{15} X_5 + \alpha_{16} X_6 + \alpha_{17} X_7 + \alpha_{18} X_8 + \alpha_{19} X_9 + v_1 \quad (5.1)$$

Tabell 1. Korrelasjonsmatrise for variablene i modellen

	Tallet på kontakter med lege $X_1$	Tallet på sykedager $X_2$	Psykisk velvære $X_3$	Reisetid til legekontor $X_4$	Arbeidstid, timer pr. uke $X_5$	Tallet på barn i husholdningen $X_6$	Sosial status $X_7$	Alder $X_8$	Kjønn $X_9$
Tallet på kontakter med lege	1.000								
Tallet på sykedager .....	0.379	1.000							
Psykisk velvære .....	0.334	0.242	1.000						
Reisetid til legekontor .....	0.016	0.015	0.020	1.000					
Arbeidstid, timer pr. uke ....	-0.108	-0.134	-0.180	-0.079	1.000				
Tallet på barn i husholdningen	-0.083	-0.092	-0.042	-0.044	0.037	1.000			
Sosial status .....	-0.033	-0.075	-0.050	-0.075	0.069	0.071	1.000		
Alder .....	0.090	0.157	0.138	0.086	-0.076	-0.384	-0.115	1.000	
Kjønn .....	0.022	-0.078	0.154	0.044	-0.386	0.060	0.002	-0.131	1.000
Gjennomsnitt .....	1.84	14.49	1.25	17.31	23.23	1.15	1.29	43.61	1.46
Standardavvik .....	3.71	47.26	1.82	21.08	21.51	1.22	0.90	17.12	0.50

Tabell 2. Estimer for regresjonskoeffisienter i modellen (5.1) og koeffisientenes standardavvik\*)

	Regresjonskoeffisient	Standardavvik
Tallet på sykedager ( $X_2$ ) .....	0.025	0.002
Psykisk velvære ( $X_3$ ) .....	0.519	0.039
Reisetid til legekontor ( $X_4$ ) .....	(0.001)	0.003
Arbeidstid, timer pr. uke ( $X_5$ ) .....	(-0.003)	0.004
Tallet på barn i husholdningen ( $X_6$ ) .....	-0.146	0.060
Sosial status ( $X_7$ ) .....	(0.031)	0.077
Alder ( $X_8$ ) .....	(-0.003)	0.004
Kjønn ( $X_9$ ) .....	(0.001)	0.152
Multipel korrelasjonskoeffisient .....	0.46	

\*) Parentes omkring estimer betyr at koeffisientene er ikke-signifikante på 5 prosent nivå (beregnet ved en t-test).

En slik enkel regresjon viser "effektene" av den enkelte variabel når de andre høyresidevariablene "holdes konstant". En slik tolkning forutsetter at en del betingelser er oppfylt, bl.a. forutsetningen om linearitet og additivitet mellom høyresidevariablene. Vi skal ikke gå nærmere inn på disse forutsetninger her.

Ifølge modellen er det helsevariablene som har størst effekt på tallet på legekontakter. Av de øvrige variable er det bare kostnadsvariabelen antall barn som gir et signifikant bidrag til forklaring av tallet på legekontakter.

Den vanlige tolkning av kvadratet av den multiple korrelasjonskoeffisient gir som resultat at den enkle regresjonsmodell "forklarer" 21 prosent av variasjonen i tallet på legekontakter.

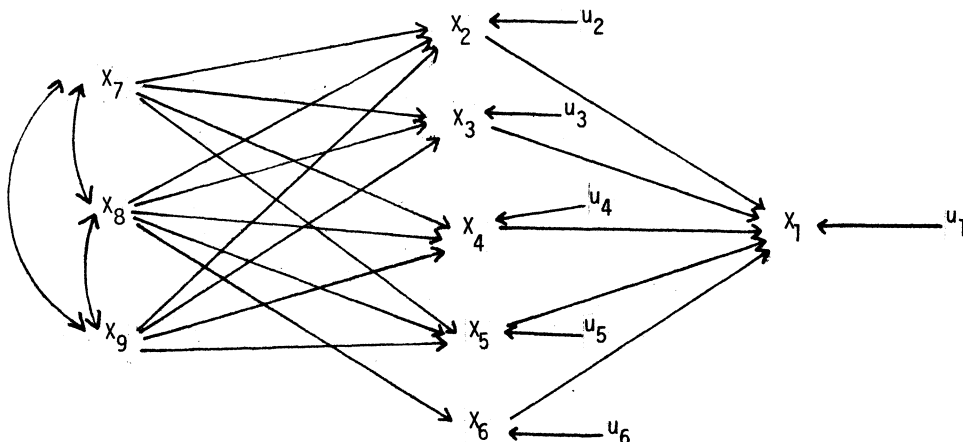
Den estimerte regresjonsmodell vil gjøre det mulig å forutsi antall legekontakter gitt bestemte verdier på variablene  $X_2$ - $X_9$ . Regresjonsanalyse er derfor en egnet metode til å studere samvariasjonen mellom en venstresidevariabel og et sett høyresidevariabler. Dersom en er interessert i årsaks-virkningsforhold mellom variable, vil den enkle regresjonsmodellen være til nytte dersom en ikke klarer, eller etter formålet i analysen ikke anser det som nødvendig, å konstruere årsaks-virkningsforhold mellom høyresidevariablene. Hvis det er mulig å konstruere slike relasjoner, mener vi at bruken av dem vil lette tolkningen av samvariasjonen i en vesentlig grad.

Vi skal anta at det er mulig å forklare kontakten med lege ved en modell av form som et rekursivt likningssystem i de 9 variable i tabell 1. Vi vil i første omgang undersøke følgende modell.

$$\begin{aligned} X_1 &= \beta_{12} X_2 + \beta_{13} X_3 + \beta_{14} X_4 + \beta_{15} X_5 + \beta_{16} X_6 + u_1, \\ X_2 &= \beta_{27} X_7 + \beta_{28} X_8 + \beta_{29} X_9 + u_2, \\ X_3 &= \beta_{37} X_7 + \beta_{38} X_8 + \beta_{39} X_9 + u_3, \\ X_4 &= \beta_{47} X_7 + \beta_{48} X_8 + \beta_{49} X_9 + u_4, \\ X_5 &= \beta_{57} X_7 + \beta_{58} X_8 + \beta_{59} X_9 + u_5, \\ \text{og} \quad X_6 &= \beta_{68} X_8 + u_6, \end{aligned} \tag{5.2}$$

der residualene er antatt å være innbyrdes uavhengig.

Denne modellen tar ikke sikte på å forklare variasjoner i sosial status ( $X_7$ ), alder ( $X_8$ ) og kjønn ( $X_9$ ). Slike variable som en vanligvis sier er "bestemt utenfor modellen", kalles eksogene variable. De øvrige variablene i modellen kalles endogene variable. De endogene variable, med unntak av antall legekontakter, er bare funksjoner av de eksogene variablene. Vi har altså ikke antatt relasjoner mellom de endogene variable i modellen, med unntak av relasjonene med  $X_1$ . Under har vi gitt et stidiagram for denne modellen:



Estimatene for stikoeffisientene i denne modellen er vist i tabell 3. Estimatene for stikoeffisientene i likningen for  $X_1$ , er naturlig nok omtrent de samme som estimatene for regresjonskoeffisientene i modell (1). Vi skal forøvrig ikke gå noe særlig inn på resultatene i denne modellen.

I tillegg til at vi får beregnet estimater for koeffisientene i modellen (5.2) gir også beregningsprogrammet en observator som måler graden av tilpasningen til modellen. La da nullhypotesen  $H_0$  være definert ved parameterspesifiseringen i modellen (5.2) og la alternativet være at vi ikke har pålagt noen restriksjoner. Da vil observatoren under nullhypotesen være  $\chi^2$  - fordelt med 21 frihetsgrader (se vedlegg). Den beregnede observator er 227.0, som gir klar forkastning av hypotesen om at modellen gir tilfredsstillende tilpasning.

Forkastningen av modellen gjør at vi må prøve å finne fram til en alternativ modell som vil gi bedre tilpasning. I dette arbeidet vil vi bruke modellen (5.2) som et utgangspunkt for en modellforbedring. Et naturlig utgangspunkt for forbedringen kan være å studere nærmere antagelsen om at residualene i modellen er ukorrelerte.

Tabell 3. Estimater for stikoeffisientene og for standardavvik i modellen (5.2)\*)

Venstresidevariabel	Høyresidevariabel	Stikoeffisient	Standardavvik	Multipel korrelasjonskoeffisient
Tallet på legekontakter ( $X_1$ )	Tallet på sykedager ( $X_2$ )	0.024	0.001	
	Psykisk velvære ( $X_3$ )	0.516	0.037	
	Reisetid til legekantor ( $X_4$ )	(0.001)	0.003	
	Arbeidstid ( $X_5$ )	(-0.003)	0.003	
	Antall barn ( $X_6$ )	-0.130	0.056	0.456
	Tallet på sykedager ( $X_2$ )	Sosial status ( $X_7$ )	-3.083	1.067
Alder ( $X_8$ )		0.394	0.056	
Kjønn ( $X_9$ )		-5.605	1.925	0.177
Psykisk velvære ( $X_3$ )	Sosial status ( $X_7$ )	(-0.065)	0.041	
	Alder ( $X_8$ )	0.017	0.002	
	Kjønn ( $X_9$ )	0.638	0.073	0.224
Reisetid til legekantor ( $X_4$ )	Sosial status ( $X_7$ )	-3.517	0.476	
	Alder ( $X_8$ )	0.094	0.025	
	Kjønn ( $X_9$ )	2.303	0.858	0.181
Arbeidstid ( $X_5$ )	Sosial status ( $X_7$ )	1.321	0.450	
	Alder ( $X_8$ )	-0.154	0.024	
	Kjønn ( $X_9$ )	-17.365	0.812	0.410
Antall barn i husholdningen ( $X_6$ )	Alder ( $X_8$ )	-0.027	0.001	0.384

\*) Parentes omkring estimater betyr at koeffisientene er ikke-signifikante på 5 prosent nivå (beregnet ved en t-test).



Korrelasjon mellom to residualer vil en finne dersom modellen på en eller annen måte er feilspesifisert. Det kan skje f.eks. ved at det finnes bakenforliggende variable som ikke går inn i modellen, men som likevel samvarierer med de to variablene med korrelerte residualer, eller ved at det er påvirkning mellom de to variablene, enten direkte eller indirekte gjennom en tredje variabel.

Estimeringsprogrammet gir muligheter til å undersøke om residualene er korrelerte. Vi har undersøkt om det finnes korrelasjon mellom residualene, og funnet at residualene  $u_2$  og  $u_3$ ,  $u_2$  og  $u_6$ ,  $u_2$  og  $u_5$ ,  $u_3$  og  $u_5$  og  $u_4$  og  $u_5$  er korrelerte, og vi vil nå prøve å utnytte denne informasjonen i det videre arbeidet med modellbyggingen.

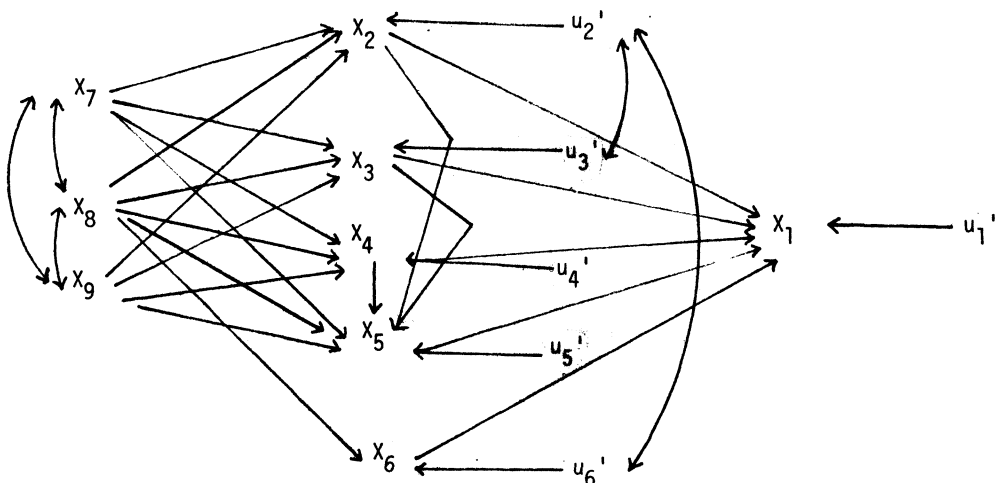
Som før nevnt kan korrelasjon mellom to residualer tolkes som en mangel ved spesifiseringen av de to tilsvarende likningene i modellen. Dersom linearitets- og additivetsforutsetningene er oppfylt, betyr det at en eller flere variable, bør tilføyes i de to likningene. Det kan være nye variable som ikke allerede inngår i modellen, eller det kan være eksogene variable som inngår i modellen eller endogene variable som inngår allerede. Vi skal i det følgende skille mellom to hovedgrupper av spesifikasjonsmangler som årsak til korrelerte residualer. Den ene består i at korrelasjonen i residualene skyldes at det finnes en eller flere felles bakenforliggende variable som ikke er tatt med i modellen. Den andre består i at de to venstresidevariablene med korrelerte residualer virker på hverandre, enten direkte eller indirekte, uten at dette er tatt hensyn til i modellen. Slike årsaks-sammenhenger kan enten gå en eller begge veier.

Ut fra de resultater vi fant for korrelasjoner mellom residualene har vi foretatt en ny spesifisering av modellen. Spesifiseringen er foretatt uten å føye til nye variable, og er gitt ved at

$$\begin{aligned} X_1 &= \beta'_{12} X_2 + \beta'_{13} X_3 + \beta'_{14} X_4 + \beta'_{15} X_5 + \beta'_{16} X_6 + u_1', \\ X_2 &= \beta'_{27} X_7 + \beta'_{28} X_8 + \beta'_{29} X_9 + u_2', \\ X_3 &= \beta'_{37} X_7 + \beta'_{38} X_8 + \beta'_{39} X_9 + u_3', \\ X_4 &= \beta'_{47} X_7 + \beta'_{48} X_8 + \beta'_{49} X_9 + u_4', \\ X_5 &= \beta'_{52} X_2 + \beta'_{53} X_3 + \beta'_{54} X_4 + \beta'_{57} X_7 + \beta'_{58} X_8 + \beta'_{59} X_9 + u_5', \\ \text{og} \\ X_6 &= \beta'_{68} X_8 + u_6', \end{aligned} \tag{5.3}$$

der alle residualene, unntatt residualparene  $(u_2', u_3')$  og  $(u_2', u_6')$ , er antatt å være innbyrdes uavhengige.

Som for modell (5.2) har vi også tegnet et enkelt stidiagram for denne modellen.



Den eneste likning som er endret, er likningen for arbeidstid ( $X_5$ ). I modell (5.2) har vi funnet at residualene  $u_2$ ,  $u_3$  og  $u_4$  var korrelert med residualen  $u_5$ . Vi fant også korrelasjoner mellom  $u_2$  og  $u_6$  og  $u_2$  og  $u_3$ , gitt som henholdsvis  $r(u_2, u_6) = 0,04$ , og  $r(u_2, u_3) = 0,24$ . Særlig korrelasjonen mellom residualene i relasjonene for sykedager og psykisk velvære var forholdsvis betydelig. Vi har likevel valgt ikke å endre likningene for  $X_2$ ,  $X_3$  og  $X_6$  for å ta hensyn til dette. Det skyldes at vi ikke har kunnet finne holdepunkter for å avgjøre hva som skaper korrelasjon mellom residualene  $u_2$  og  $u_3$  og mellom  $u_2$  og  $u_6$ . Vi har ikke noe grunnlag for å si f.eks. at tallet på sykedager påvirker den psykiske velvære, eller at den psykiske velvære påvirker antall sykedager (fravær fra arbeidet p.g.a. psykiske vansker), eller at korrelasjoner mellom henholdsvis tallet på sykedager og psykisk velvære og en tredje variabel gir opphav til korrelasjonen mellom residualene. En har naturligvis også muligheten av en blanding av disse tre effekter.

Tabell 4. Estimer for stikoeffisientene og for standardavvik i modellen (5.3)\*).

Venstresidevariabel	Høyresidevariabel	Stikoeffisient	Standardavvik	Multipel korrelasjonskoeffisient
Tallet på legekontakter ( $X_1$ )	Tallet på sykedager ( $X_2$ )	0.024	0.001	0.456
	Psykisk velvære ( $X_3$ )	0.516	0.037	
	Reisetid til legekontor ( $X_4$ )	(0.001)	0.003	
	Arbeidstid ( $X_5$ )	(-0.003)	0.003	
	Antall barn ( $X_6$ )	-0.130	0.056	
Tallet på sykedager ( $X_2$ )**)	Sosial status ( $X_7$ )	-3.083	1.067	0.177
	Alder ( $X_8$ )	0.394	0.056	
	Kjønn ( $X_9$ )	-5.605	1.924	
Psykisk velvære ( $X_3$ )**)	Sosial status ( $X_7$ )	(-0.065)	0.041	0.224
	Alder ( $X_8$ )	0.017	0.002	
	Kjønn ( $X_9$ )	0.638	0.073	
Reisetid til legekontor ( $X_4$ )	Sosial status ( $X_7$ )	-3.517	0.476	0.181
	Alder ( $X_8$ )	0.094	0.025	
	Kjønn ( $X_9$ )	2.303	0.858	
Arbeidstid ( $X_5$ )	Tallet på sykedager ( $X_2$ )	-0.059	0.009	0.443
	Psykisk velvære ( $X_3$ )	-0.855	0.229	
	Reisetid til legekontor ( $X_4$ )	-0.046	0.019	
	Sosial status ( $X_7$ )	0.923	0.448	
	Alder ( $X_8$ )	-0.112	0.024	
	Kjønn ( $X_9$ )	-17.047	0.816	
Antall barn i husholdningen ( $X_6$ )**)	Alder ( $X_8$ )	-0.027	0.001	0.384

\*) Parenteser omkring estimer betyr at koeffisientene er ikke-signifikante på 5-prosent nivå (beregnet med en t-test).

\*\*\*) Residualene  $u_2'$  og  $u_3'$ ,  $u_2'$  og  $u_6'$  er korrelerte. Estimatene for korrelasjonskoeffisientene er henholdsvis  $r(u_2', u_3') = 0,24$  og  $r(u_2', u_6') = 0,04$ .

Når det gjelder korrelasjonene mellom  $u_2$  og  $u_5$ ,  $u_3$  og  $u_5$ ,  $u_4$  og  $u_5$  er det vår påstand at de to første av disse korrelasjonene må forklares med at helsen, målt ved tallet på sykedager og psykisk velvære, virker inn på arbeidstiden (selv når en har kontrollert for andre variable).

Videre er det vår påstand at reisetid til legekontor virker inn på arbeidstid. Dette siste er uttrykk for en forenkling. Tankegangen bak er at bostedets urbaniseringsgrad virker inn på både arbeidstid (via bl.a. næringsstrukturen på stedet) og på reisetid til legekontor. En må her merke seg at arbeidstid både er uttrykk for yrkesfrekvensen på stedet og for yrkesaktives arbeidstid. Forenklingen består i at vi betrakter reisetid til legekontor som et mål på urbaniseringsgrad. Mer korrekt ville det vært i spesifiseringen av likningssystemet å la et mål for urbaniseringsgrad ( $X_{10}$ ) inngå både i likningen for  $X_4$  og likningen for  $X_5$ .

Tabell 4 viser estimatene for stikoeffisientene i denne modellen. Den beregnede  $\chi^2$ -observatoren blir for denne modellen 3,70. Siden vi under hypotesen om at parametrene er spesifisert ved modellen (5.3), finner at  $\chi^2$ -observatoren har 16 frihetsgrader, gir det numeriske resultatet at vi ikke kan forkaste denne hypotesen. Vi stopper derfor i vårt arbeid i å prøve å forbedre modellen ytterligere, og vil i neste avsnitt diskutere resultatene vi har kommet fram til.

## 6. Diskusjon av resultater

Modellen ble forenklet med utgangspunkt i den antagelse at det er tre sett av faktorer som virker inn på bruk av legetjenester, nemlig behovsfaktorer, kostnadsfaktorer og disponerende faktorer. Vår interesse knytter seg til om andre faktorer enn behovsfaktorer har noen virkning på bruk av legetjenester, hvilke faktorer dette eventuelt er og hvor stor virkning de har.

Modellen er svak når det gjelder disponerende faktorer. Bare sosial status i vår modell er i en del tilfeller brukt som disponerende faktor (Andersen et al. (1970, s. 31)). Tabell 2 viste imidlertid at sosial status ikke hadde noen direkte virkning på antallet legekontakter. Det er imidlertid stor uenighet innen helsetjenesteforskningen om disponerende faktorer har noen virkning (se f.eks. Berkanovic & Reeder (1974) og Bice & White (1969)).

Tabell 3 viser at det er helsevariablene i modellen: tallet på sykedager og psykisk velvære som har de største virkningene på legekontakter. Disse helsevariablene er i modellen tenkt som mål på behovet for legetjenester. Dette resultatet stemmer overens med resultatet fra de fleste undersøkelser der en har undersøkt den samtidige effekt av behovsfaktorer og andre faktorer på bruken av legetjenester.

Av de tre kostnadsfaktorene som inngår i modellen, er det bare tallet på barn som har en signifikant effekt på antall legekontakter.

Heller ikke når en ser på indirekte effekter av variablene, finner en grunnlag for å si at andre faktorer enn behovsfaktorer har vesentlig innvirkning på bruken av legetjenester. Sosial status har indirekte effekter gjennom reisetid til legekontor og arbeidstid. Effekten er ikke signifikant. Den indirekte effekt av reisetid til legekontor gjennom arbeidstid er heller ikke signifikant. Alder har indirekte effekt gjennom reisetid til legekontorer, arbeidstid og først og fremst gjennom antall barn i husholdningen. Disse er imidlertid små sammenliknet med de indirekte effektene av alder gjennom helsevariablene.

Vi skal nå gå nærmere inn på enkeltheten i modellene. Vi har brukt to variable som helsevariable. Det er langt fra uproblematisk å bruke slike variable som mål på behovet for legetjenester, f.eks. slik som det ville bli registrert av legen gjennom kliniske undersøkelser. Dette problemet har blitt tatt opp i meget få undersøkelser av bruk av legetjenester. Vi skal heller ikke gå nærmere inn på det her. Likevel har vi ment at en betingelse for å få et brukbart mål på behov for legetjenester, er at en nytter mål på flere av dimensjonene i helsebegrepet. Vi har derfor ikke bare nyttet tallet på sykedager som ofte har vært brukt, men også et mål på psykisk helse. Som tidligere nevnt mangler vi mål for kroniske lidelser.

Tallet på sykedager har vært nyttet før i undersøkelser av bruken av legetjenester. I modell (5.3) fant vi at stikoeffisienten fra tallet på sykedager til bruk av legetjenester var 0.024, eller omregnet til standardiserte koeffisienter<sup>\*)</sup> 0.305. Galvin & Fan (1975) fant i en tilsvarende undersøkelse en koeffisient på 0.255. Variablene var de samme med det unntak at registreringsperioden for tallet på sykedager og tallet på legekontakter var 2 måneder.

Vi er ikke kjent med at variable av typen psykisk velvære har vært trukket inn i analysen av bruk av helsetjenester. Selv om variabelen ikke er en intervallvariabel, mener vi likevel at det ikke kan være tvil om at psykisk velvære spiller en betydelig rolle for tallet på legekontakter.

Det er flere mulige tolkninger av dette resultatet. En mulighet er at det er en nær sammenheng mellom forekomst av kroniske lidelser av primær fysisk karakter og forekomst av de symptomer som inngår i variabelen psykisk velvære (tretthet, søvnproblemer, angst og nervøsitet, anspenning i fritiden). Dersom dette er riktig, kan effekten av psykisk velvære i modellen være et uttrykk for virkningen av kroniske lidelser på legekontakt.

En annen mulig tolkning er at det mer direkte er psykiske lidelser som er årsak til legekontakter. En kan kanskje se dette i sammenheng med økning i forekomsten av psykosomatiske lidelser. Det har også vært pekt på at en del sykdomsbilder må karakteriseres som diffuse, det er vanskelig å stille diagnoser, en finner sosiale, psykiske og fysiske problemer tett sammenvevd.

Et annet forhold som kan være verdt å trekke inn når en betrakter effekten av psykisk velvære på legekontakt, er at legen har andre funksjoner enn de å stille diagnoser og helbrede sykdommer. I noen grad har legene også som funksjon å være en av de vesentligste kontaktkanaler både til behandlingsapparatet og til det sosiale hjelpeapparatet. Uttrykt på en annen måte vil legene for en del folk være de første en henvender seg til med problemer (og ikke bare medisinske problemer) som de ikke klarer å løse innen den mer private sfære, familie og vennekrets.

Av kostnadsfaktorene var det bare tallet på barn i husholdningen som hadde noen signifikant effekt på tallet på legekontakter. Utgangspunktet for å trekke denne variabelen inn i modellen var at barn i familien virket som en hindring på legebesøk. Dersom vi tolker resultatet slik, må en vente at effekten stort sett bare vil eksistere for kvinner, ikke for menn. Forutsetningen om additivitet i modellen vil ikke være oppfylt. En videre utbygging av modellen bør derfor inneholde et ledd som tar hensyn til samspillseffekten mellom kjønn og antall barn. Dersom dette viser seg å være riktig, innebærer det at effekten av antall barn i husholdningen er ganske betydelig for kvinner.

De to andre kostnadsfaktorer, arbeidstid og reisetid til legekontor, har ikke signifikante effekter på antallet legekontakter. Den geografiske tilgjengeligheten av legetjenester inngår i en rekke undersøkelser av legetjenester. Stort sett har man funnet samme negative resultat som her. I en internasjonal sammenliknende studie av bruken av legetjenester fant en at reisetid til legekontor bare hadde signifikant effekt i Jugoslavia, derimot ikke i de undersøkte områdene i England og USA (Bice & White (1969)). Som tidligere nevnt kan en mulig forklaring være at kostnadene i form av reisetid bare utgjør en del av kostnadene i form av tapt tid ved legebesøk. En annen slik kostnad ved legebesøk er ventetid hos legen.

En annen mulig forklaring er at sammenhengen mellom legekontakt og avstand til legekontor ikke er lineær. Det er mulig at reisetiden først oppfattes som en kostnad ved meget lange reisetider.

Vi skal knytte vår konklusjon til det som var utgangspunktet for formuleringen av modellen. Utgangspunktet var om andre faktorer enn behovsfaktorer har noen innvirkning på bruken av legetjenester.

En kan anta som Andersen et.al. (1970), at det er et grunnliggende mål i helsepolitikken å bygge opp en helsetjeneste der det er en "equitable distribution of health services" (Andersen et.al. (1970, s. 131-132)). Dette betyr ikke at alle skal motta helsetjenester i samme omfang, heller ikke at alle skal motta helsetjenester i det omfang de har behov for. Det betyr at helsetjenestene skal fordeles i befolkningen i forhold til de forskjellige personers behov. Dersom helsetjenestene i Norge var fordelt slik, ville en vente et resultat som var tilnærmet lik det resultat vi har funnet. Analysen gir altså ikke noen forkastning av en hypotese om at legetjenester tilnærmet er fordelt "i forhold til behov", forutsatt at de to helsevariablene som inngår i modellen er gyldige som mål på behov for legetjenester.

\*) I denne analysen har vi valgt å bruke ustandardiserte variable under estimering av koeffisientene. Ved å bruke estimater for variansen til variabler i modellen kan en lett regne om fra ustandardiserte til standardiserte koeffisienter (se f.eks. Mayer & Younger (1976)).

Referanser

- Aday, L.A. (1972): "The utilization of health services: Indices and correlates. A research bibliography." National Center for Health Services Research and Development, 1972.
- Alwin, D.F. & R.M. Hauser (1975): "The decomposition of effects in path analysis." American Sociological Review, 40, 37-47.
- Amundsen, A. (1973): "Om identifikasjon av økonometriske modeller." Statistisk Sentralbyrå, Arbeidsnotat IO 73/7.
- Andersen, R., B. Smedby & O.W. Anderson (1970): "Medical care use in Sweden and the United States. A comparative analysis of systems and behaviour." Center of Health Administration Studies, University of Chicago. Research Series 27.
- Anderson, J.G. (1973): "Causal models and social indicators: Towards the development of social systems models." American Sociological Review, 38, 285-301.
- Anderson, J.G. & D.E. Bartkus (1973): Choice of medical care: A behavioural model of health and illness behaviour. Journal of Health and Social Behaviour, 14, 348-361.
- Anderson, T.W. (1958): "An introduction to multivariate analysis." New York: John Wiley & Sons, Inc.
- Berkanovic, E. & G. Reeder (1974): "Can money buy the appropriate use of services? Some notes on the meaning of utilization data." Journal of Health and Social Behaviour, 15, 93-99.
- Bice, T.W., R.L. Eichhorn & P.D. Fox (1972): "Socioeconomic status and use of physician services: A reconsideration". Medical Care, 10, 261-271.
- Bice, T.W. & K.L. White (1969): "Factors related to the use of health services: An international comparative study." Medical Care, 7, 124-133.
- Blalock, H.M. (1967): "Path coefficients versus regression coefficients." American Journal of Sociology, 72, 675-676.
- Blau, P.M. & O.D. Duncan (1967): "The American occupational structure." New York: John Wiley & Sons, Inc.
- Duncan, O.D. (1966): "Path analysis: Sociological examples." American Journal of Sociology, 72, 1-16.
- Finney, I.M. (1972): "Indirect effects in path analysis." Sociological Methods & Research, 1, 175-186.
- Galvin, M.E. & M. Fan (1975): "The utilization of physicians' services in Los Angeles County, 1973." Journal of Health and Social Behaviour, 16, 74-94.
- Garaas, E. (1975): "Tre-trinns minste kvadraters metode. Beskrivelse av datamaskinprogrammet 3SLSPROG." Statistisk Sentralbyrå, Arbeidsnotat IO 75/41.
- Jøreskog, K.G. (1973): "A general method for estimating a linear structural equation system." I Goldberger, A.S. & O.D. Duncan (Eds) (1973): "Structural equation models in the social sciences." New York: Seminar Press.
- Jøreskog, K.G. & M. van Thillo (1973): "LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables." University of Uppsala. Research Report 73-5.
- Kalimo, E. & K. Sievers (1968): "The need for medical care: Estimation on the basis of interview data." Medical Care, 6, 1-17.
- McKinlay, J.B. (1972): "Some approaches and problems in the study of the use of services - An overview." Journal of Health and Social Behaviour, 13, 115-152.
- Mayer, L.S. & M.S. Younger (1976): "Estimation of standardized regression coefficients." Journal of the American Statistical Association, 71, 154-157.
- Monteiro, L.A. (1973): "Expense is no object ...: Income and physician visits reconsidered." Journal of Health and Social Behaviour, 14, 99-114.
- Simon, H.A. (1953): "Causal ordering and identifiability." I Hood, W.C. & T.C. Koopmans (Eds) (1973): "Studies in econometric method." New York: John Wiley & Sons, Inc.
- Skrede, K. (1971): "Sosiøkonomisk klassifisering av yrker i Norge, 1960." Institutt for anvendt sosialvitenskapelig forskning. Rapport 71-1.
- Statistisk Sentralbyrå (1975): "NOS A 720: Levekår 1973."
- Statistisk Sentralbyrå (1976): "SA nr. 21: Yrke og dødelighet 1970-1973."
- Sverdrup, E. (1964): "Lov og tilfeldighet. Bind II." Oslo: Universitetsforlaget.

- Telje, J. (1974): "Hvem går ikke til doktor?" Tidsskrift for den norske lægeforening, 26, 1 596-1 599.
- Tukey, J.W. (1954): "Causation, regression and path analysis." I Kempthorne, O et. al. (1954):  
"Statistics and mathematics in biology." Ames: Iowa State College Press.
- Wan, T.T.H. & S.J. Soifer (1974): "Determinants of physician utilization: A causal analysis."  
Journal of Health and Social Behaviour, 15, 100-108.
- Werts, C.E. & R.L. Linn (1970): "Path analysis: Psychological examples." Psychological Bulletin, 74,  
193-212.
- Wright, S. (1934): "The method of path coefficients." Annals of Mathematical Statistics, 5, 161-215.

## 7. Vedlegg: Estimering av parametre i lineære strukturlikninger

I dette vedlegget vil vi betrakte et generelt sett av lineære strukturlikninger og en metode til å estimere parametre i likningene. I det følgende vil vi anta at alle parametrene i relasjonene er identifiserbare, og vi vil derfor ikke nærmere påpeke hvor det er nødvendig å forutsette identifiserbarhet. Det er vanlig i litteraturen å skille mellom eksogene og endogene variable i lineære strukturlikninger, men i avsnitt 2 var det ikke nødvendig å innføre dette skillet. I det senere vil vi gjøre bruk av begrepene, og vi sier da at en endogen variabel er en variabel som er bestemt av den prosessen som er karakterisert av modellen, og en eksogen variabel er bestemt utenfor modellen, slik at de eksogene variable og residualene i likningene er stokastisk uavhengige.

La nå strukturlikningene være gitt ved at

$$A \underset{\sim}{y} = \varepsilon \quad (7.1)$$

der  $\underset{\sim}{y}$  er en  $(p \times 1)$ -vektor som består av observasjoner for ett individ,  $A$  er en  $(q \times p)$ -matrise av ukjente og a priori kjente parametre, og  $\varepsilon$  er en  $(q \times 1)$ -residualvektor. Vi ønsker å estimere de ukjente parametrene i  $A$ .

Variabelen  $\underset{\sim}{y}$  består av både eksogene og endogene variable. La da

$$\underset{\sim}{y} = \begin{pmatrix} \underset{\sim}{x} \\ \underset{\sim}{z} \end{pmatrix},$$

og

(7.2)

$$A = (B, \Gamma),$$

der  $\underset{\sim}{x}$  er et sett av endogene variable og  $\underset{\sim}{z}$  er et sett av eksogene variable.  $\underset{\sim}{x}$  er en  $(r \times 1)$ -vektor og  $\underset{\sim}{z}$  er en  $(s \times 1)$ -vektor, der  $p = r + s$ . Ved å innsette (7.2) i (7.1), får vi at

$$B \underset{\sim}{x} + \Gamma \underset{\sim}{z} = \varepsilon \quad (7.3)$$

Vi antar nå at residualvektoren  $\varepsilon$  er slik at

$$E \varepsilon = 0,$$

og

$$E \varepsilon \varepsilon' = \Psi$$

der  $\Psi$  er en positiv definit  $(q \times q)$ -matrise som består av ukjente eller a priori kjente parametre.

Vi antar at  $B$  er normert slik at det er bare enere på hoveddiagonalen. Vi ser videre at modellen (3.3) er rekursiv dersom  $B$  er en matrise med bare nuller over hoveddiagonalen, og  $\Psi$  er en diagonalmatrise.

Hvis vi nå vender tilbake til likningssystemet (2.2), merker vi oss at  $x_1$  kan betraktes som en eksogen variabel og  $x_2$ ,  $x_3$  og  $x_4$  som endogene variable. Modellen (2.2) kan altså omskrives som

$$B \begin{pmatrix} x_2 \\ x_3 \\ x_4 \end{pmatrix} + \Gamma x_1 = \begin{pmatrix} \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

der

$$\beta = \begin{pmatrix} 1 & 0 & 0 \\ -p_{32} & 1 & 0 \\ -p_{42} & -p_{43} & 1 \end{pmatrix}, \text{ og } \lambda = \begin{pmatrix} -p_{21} \\ -p_{31} \\ -p_{41} \end{pmatrix}$$

Hvis vi i tillegg forutsetter at kovariansmatrisen til residualvektoren er diagonal, er modellen (2.2) rekursiv.

Vi vender nå tilbake til den generelle formen (7.3). Vi antar at  $\epsilon$  er multinormalfordelt og at  $z$  har en sannsynlighetstetthet som er uavhengig av parametrene i  $\beta$ ,  $\lambda$  og  $\Psi$ . Videre antar vi at vi har  $n$  sett av observasjoner  $y$  der (7.1) og dermed (7.3) gjelder. La da  $\Upsilon$ ,  $\chi$  og  $z$  være matrisene av observasjonssett slik at

$$\Upsilon = \begin{pmatrix} y_{11} & y_{21} & \dots & y_{n1} \\ y_{12} & y_{22} & \dots & y_{n2} \\ \dots & \dots & \dots & \dots \\ y_{1p} & y_{2p} & \dots & y_{np} \end{pmatrix}, \quad \chi = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots \\ x_{1r} & x_{2r} & \dots & x_{nr} \end{pmatrix} \text{ og } z = \begin{pmatrix} z_{11} & z_{21} & \dots & z_{n1} \\ z_{12} & z_{22} & \dots & z_{n2} \\ \dots & \dots & \dots & \dots \\ z_{1s} & z_{2s} & \dots & z_{ns} \end{pmatrix}$$

Videre lar vi  $\epsilon$  være gitt ved

$$\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n) = \begin{pmatrix} \epsilon_{11} & \epsilon_{21} & \dots & \epsilon_{n1} \\ \epsilon_{12} & \epsilon_{22} & \dots & \epsilon_{n2} \\ \dots & \dots & \dots & \dots \\ \epsilon_{1q} & \epsilon_{2q} & \dots & \epsilon_{nq} \end{pmatrix}$$

der kolonnene i  $\epsilon$  er uavhengige, slik at  $E \epsilon_i \epsilon_j' = 0$  for  $i \neq j$ .

Med den notasjonen som vi nå har innført, kan (7.1) utvides til å kunne skrives som

$$A \Upsilon = \epsilon, \quad (7.4)$$

og dermed kan (7.3) omformes til

$$\beta \chi + \lambda z = \epsilon \quad (7.5)$$

Dersom  $\beta$  er ikke-singulær, kan (7.5) ved en enkel omforming skrives som

$$\chi = \pi z + \mu, \quad (7.6)$$

der  $\pi = -\beta^{-1} \lambda$ , og  $\mu = \beta^{-1} \epsilon$ . Kovariansmatrisen til  $\mu = \beta^{-1} \epsilon$  er videre

$$\beta^{-1} \Psi (\beta')^{-1} = \Omega.$$

Vi vil nå finne sannsynlighetsmaksimeringsestimatorer for de ukjente parametrene i  $\beta$ ,  $\lambda$  og  $\Psi$ . Siden vi, som før omtalt, har antatt at fordelingen til  $z$  er uavhengig av parametermatrisene  $\beta$ ,  $\lambda$  og  $\Psi$ , vil sannsynlighetsmaksimeringsprinsippet føre til at vi kan studere den betingete fordelingen for  $\chi$  gitt  $z$ . Når vi i tillegg antar at  $\epsilon$  er multinormalfordelt, kan det vises, (se f.eks. Anderson (1958,



s. 179)) at logaritmen til likelihoodfunksjonen kan skrives som en konstant pluss

$$\log L = -\frac{1}{2}n \log |\Sigma| - \frac{1}{2} \text{tr} \Sigma^{-1} (\mathbf{X} - \pi \mathbf{Z}) (\mathbf{X} - \pi \mathbf{Z})' \quad (7.7)$$

Ved å innsette (7.5) i (7.7) finner vi at

$$\begin{aligned} \text{tr} \Sigma^{-1} (\mathbf{X} - \pi \mathbf{Z}) (\mathbf{X} - \pi \mathbf{Z})' &= \text{tr} \{ \mathbf{B}^{-1} \Psi (\mathbf{B}')^{-1} \}^{-1} \{ \mathbf{B}^{-1} \boldsymbol{\varepsilon} \} \{ \mathbf{B}^{-1} \boldsymbol{\varepsilon} \}' \\ &= \text{tr} \mathbf{B}' \mathbf{B}^{-1} \mathbf{B} \mathbf{B}^{-1} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' (\mathbf{B}')^{-1} \\ &= \text{tr} \Psi^{-1} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' \\ &= \sum_i \sum_j \sigma^{ij} \varepsilon_i \varepsilon_j' \end{aligned}$$

der  $\sigma^{ij}$  er  $(i, j)$ -te element i matrisen  $\Psi^{-1}$ . Innsatt i (7.6) gir dette at

$$\log L = \frac{1}{2} n \log \{ |\Psi| / |\mathbf{B}|^2 \} - \sum_i \sum_j \sigma^{ij} \varepsilon_i \varepsilon_j' \quad (7.8)$$

Sannsynlighetsmaksimeringsmetoden fører altså til at (7.8) skal maksimeres med hensyn på de ukjente parametrene i  $\mathbf{B}$ ,  $\boldsymbol{\pi}$  og  $\Psi$ .

Hvis vi antar at likningssystemet (7.5) er rekursivt, er  $\mathbf{B}$  triangulær med diagonalelementer lik 1 og  $\Psi$  er diagonal. Da blir altså  $|\mathbf{B}| = 1$ , og  $\log L$  reduseres til

$$\begin{aligned} \log L &= -\frac{1}{2} n \sum_i \log \sigma_{ii} - \frac{1}{2} \sum_i \sigma^{ii} \varepsilon_i \varepsilon_i' \\ &= \sum_i \log L_i \end{aligned} \quad (7.9)$$

Vi ser da at sannsynlighetsmaksimeringsprinsippet fører til at vi skal finne sannsynlighetsmaksimeringsestimatorene i hver likning i (7.9) for seg. I tillegg finner vi også at i et rekursivt likningssystem vil sannsynlighetsmaksimeringsestimatorene bli identiske med minste kvadraters estimatorene når estimatorene er beregnet i hver likning for seg.

Til nå har vi forutsatt at residualvektoren  $\boldsymbol{\varepsilon}$  er uavhengig av de eksogene variable i likningssystemet og at  $\boldsymbol{\varepsilon}$  er multinormalt fordelt med  $E \boldsymbol{\varepsilon} = \mathbf{0}$  og  $E \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}' = \Psi$ . I det følgende vil vi studere en alternativ estimeringsmetode. Jøreskog (1973) har nemlig foreslått en estimeringsmetode som bygger på antagelse om multinormalitet av vektoren  $\mathbf{y} = (\mathbf{X}', \mathbf{Z}')'$ . I en del praktiske eksempler vil kravet om multinormalitet være et vesentlig vanskeligere krav å få oppfylt enn kravet om at bare residualvektor er multinormalfordelt. Vi vil likevel se hvilke konsekvenser antagelsen får for estimeringen av

Som tidligere lar vi likningssystemet være gitt ved (7.5). La videre

$$E \mathbf{X} = E \mathbf{Z} = \mathbf{0},$$

og

$$E \mathbf{X} \mathbf{X}' = \boldsymbol{\Sigma}.$$

Vi innfører også rotasjonen

$$E \zeta \zeta' = \phi$$

$$E \varepsilon \varepsilon' = \Psi$$

og antar at

$$E \chi \varepsilon' = 0.$$

Ved å anvende (7.6) finner vi at

$$\bar{\zeta} = \begin{pmatrix} B^{-1} \Gamma \Psi \Gamma' (B')^{-1} + B^{-1} \Psi (B')^{-1} & - B^{-1} \Gamma \Psi \\ \phi \Gamma' (B')^{-1} & \phi \end{pmatrix} \quad (7.10)$$

Vi innfører videre  $\bar{\zeta}$  som er estimert kovariansmatrisen til  $\bar{y}$ .  $\bar{\zeta}$  kan spaltes opp som

$$\bar{\zeta} = \begin{pmatrix} \bar{\zeta}_{xx} & \bar{\zeta}_{xz} \\ \bar{\zeta}_{zx} & \bar{\zeta}_{zz} \end{pmatrix}$$

Jøreskog (1973) har nå vist at logaritmen til likelihoodfunksjonen kan skrives som en konstant pluss

$$\begin{aligned} \log L^* &= -\frac{1}{2} n \{ \log |\phi| + \text{tr} (\bar{\zeta}_{zz} \phi^{-1}) \} \\ &- \frac{1}{2} n \{ \log |\Psi| - \log |B|^2 \} - \frac{1}{2} \text{tr} \{ (B \bar{\zeta}_{xx} B' + B \bar{\zeta}_{xz} \Gamma' + \Gamma \bar{\zeta}_{zx} B' + \Gamma \bar{\zeta}_{zz} \Gamma') \Psi^{-1} \}. \end{aligned}$$

Maksimering med hensyn på  $\phi$  gir  $\hat{\phi} = \bar{\zeta}_{xx}$ . Ved å bruke at

$$B \chi + \Gamma \zeta = \varepsilon,$$

finner vi at

$$\log L^* = -\frac{1}{2} n \{ \log |\Psi| - \log |B|^2 \} - \frac{1}{2} \sum_i \sum_j \sigma^{ij} \varepsilon_i \varepsilon_j', \quad (7.11)$$

som er identisk med (7.8). Estimeringsprogrammet LISREL beregner estimater ved å minimere  $\log L^*$  ved numeriske metoder. Programmet beregner også numeriske tilnærmelser til variansen til estimatene. For å oppsummere det vi har funnet, kan vi altså si at parameterestimaterne beregnet ved LISREL er identiske med estimatene som vi finner fra (7.8), derkonstruksjonen av estimatene bare bygde på antagelse om at residualene er multinormalfordelt.

I en gitt modell vil det ofte være av interesse å teste hypoteser om parametrene i modellen. For å utvikle slike tester, benytter vi at Sverdrup (1964, s. 125) har vist at hvis  $n$  er stor, vil sannsynlighetsmaksimeringsestimatorene være tilnærmet normalfordelt med kjent momentmatrise. Vi får da at vi kan teste hypoteser om parametrene  $B$  og  $\Gamma$  ved de standard t-testene som i følge det som er nevnt ovenfor, vil være tilnærmet t-fordelt. Dersom modellen vi studerer er rekursiv, vil imidlertid teorien forenkles, og de vanlige t-testene vil være tester med de kjente statistiske egenskaper.

Når vi har beregnet sannsynlighetsmaksimeringsestimatorer, kan vi også teste tilpasningen til den modellen vi studerer. Dette gjøres ved sannsynlighetskvotetesten. La hypotesen være gitt ved den spesifisering av parametrene som modellen gir mens alternativet er at kovariansmatrisen er positiv definit. La videre  $K$  være sannsynlighetskvoten. Da kan det vises, se eksempelvis Sverdrup (1964, s. 128), at under nullhypotesen er, for store utvalg,  $-2\ln K$  tilnærmet  $\chi^2$  - fordelt med antall frihetsgrader lik differansen mellom antall ukjente parametre a priori og antall ukjente parametre under nullhypotesen. EDB-program LISREL beregner observatoren  $-2\ln K$ .