Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Dep, Oslo 1. Tlf.*(02) 41 38 20

IO 77/44 30. november 1977

ESTIMERING AV ENGELDERIVERTE VED MANGLENDE INNTEKTSDATA

av

Odd Skarstad[1)]

INNHOLD

1) Jeg vil takke dosent Arne Amundsen, forsker Erik Biørn, konsulentene Grete Dahl, Helge Herigstad og Petter Laake for nyttige kommentarer under arbeidet.

I. EMNE FOR ANALYSEN

Dette arbeidet handler om estimering av sammenhenger mellom inntekt og forbruksutgifter på tverrsnittsdata fra forbruksundersøkelser. Det er egenskaper ved forbruksundersøkelsesdata - med derav følgende begrensninger og muligheter for analyse - som oppmerksomheten er konsentrert om.

Datamaterialer i forbruksundersøkelser blir vanligvis innhentet ved at husholdningene fører regnskap over forbruksutgiftene i en kortere periode (én eller noen få uker). Dessuten blir de gjerne intervjuet om anskaffelse av sjeldenkjøpsvarer over en lengre periode (f.eks. privatbil).

Av praktiske grunner er det neppe mulig å foreta noen særlig mer fullstendig kartlegging av forbruket i de enkelte husholdningene, f.eks. ved å anvende en meget lengre regnskapsperiode. Erfaring viser at frafallet blant husholdningene i forbruksundersøkelser ofte er ubehagelig stort selv ved en kort regnskapsperiode, og det er neppe grunn til å tvile på at det som regel ville øke ved overgang til lengre periode.

Forbruksundersøkelser gir altså et nokså kort "glimt" av forbruket hos de deltakende husholdningene. Det kan ikke overraske at slike korte glimt er en ulempe ved analyse av data, fremfor mer fullstendige opplysninger om forbruket. Det er vesentlig å være klar over dette ved bruk av et materiale. Foruten svakheter som skyldes en kort regnskapsperiode, må en selvsagt også regne med andre feilkilder, f.eks. at en del utgifter kan være uteglemt under bokføringen.

Framstillingen er knyttet til datamateriale fra norske forbruksundersøkelser. Men det betyr neppe at synspunkter og resultater i dette arbeidet behøver å være uten interesse under arbeid med andre undersøkelser. Dette gjelder vel selv om disse måtte ha et noe annerledes opplegg. Problemer med å utnytte et datamateriale og tolke resultater kan være prinsipielt nokså like selv om f.eks. lengden på bokføringsperioden er noe forskjellig.

I våre forbruksundersøkelser er det et hovedproblem at man mangler eller har mangelfulle inntektsopplysninger for husholdningene. I denne analysen har vi sett på mulighetene for å estimere sammenhenger mellom inntekt og utgifter til forskjellige vare- og tjenestegrupper ved manglende inntektsoppgaver.

I kapittel II er det gitt noen eksempler på bruk av total forbruksutgift som klassifieringsvariabel, dvs. som variabel for gruppering av husholdninger etter økonomisk nivå. Det kan i vesentlig grad skyldes tilfeldigheter hvorvidt en husholdning har store eller små utgifter i løpet av en kort registreringsperiode, og dermed skyldes det også delvis tilfelgheter om den havner i gruppen for lavt eller høyt materielt nivå. I kapittel II er det også gitt resultater fra forbruksundersøkelsene og fra et konstruert talleksempel på et datamateriale.

I kapittel III har vi forutsatt at det gjelder en lineær sammenheng mellom forbruksdisponibel inntekt og forbruk. Vi ønsker å estimere parametrene i relasjonen. Imidlertid antas det at inntektsdata mangler. Vi har derfor undersøkt egenskapene ved estimatoren for den marginale konsumtilbøyelighet (for forskjellige varegrupper) ved bruk av total forbruksutgift i stedet for inntekt som forklaringsmiddel. Dette er belyst både gjennom et simulert datamateriale og analytisk. Deretter har vi studert estimatorer for den marginale konsumtilbøyelighet via mellomregninger med total forbruksutgif (eller en lignende størrelse) som "forklaringsvariabel".

II. GRUPPERING AV HUSHOLDNINGER ETTER TOTAL FORBRUKSUTGIFT

1. Innledning

Ved analyser av forbruksundersøkelser er det ofte interesse for å dele inn husholdningene etter en eller annen indikator for økonomisk evne. Dette kan bl.a. ha sammenheng med ønske om å sammenlikne "rike" med "fattige" befolkningsgrupper.

En vanlig brukt indikator på økonomisk nivå er inntekt. Inntekt kan defineres på mange forskjellige måter (f.eks. brutto-, nettoinntekt ned skatteligningen). Et inntektsbegrep som ofte er nærliggende å nytte er inntekt etter at skatter og avgifter er betalt, dvs. inntekt som er disponibel til forbruk og sparing. Det er imidlertid ofte vanskelig å skaffe pålitelige data.

På grunn av manglende eller ufullstendige inntektsopplysninger blir undertiden total forbruksutgift nyttet som mål på økonomisk nivå. En deler altså inn husholdningene i grupper etter totalutgift og studerer forbruket og dets sammensetning innen gruppene. I neste avsnitt er det gitt noen resultater fra et datameteriale. I avsnittet som følger deretter, blir tolkningen av resultatene drøftet.

2. Noen resultater fra forbruksundersøkelsen 1973

Som et eksempel på "sammensetningen" av forbruket betraktes todelingen:

- visse varer som kjøpes ofte, her kalt dagligvarer[1)]
- alle ander forbruksutgifter

Tabell 1 viser utgifter til disse to varegruppene i forbruksundersøkelsen 1973. Husholdningene er gruppert etter total forbruksutgift. Vi betrakter to grupper av husholdninger, dem med total forbruksutgift henholdsvis under 10 000 kroner og over 80 000 kroner pr. år.

Tabell 1. Gjennomsnittlig utgift pr. husholdning i grupper for total forbruksutgift, etter vare- og tjenestegruppe

Vare- og tjenestegruppe	Alle		Total forbruksutgift: Under 10 000 kroner		Total forbruksutgift: 80 000 kroner og over	
	Kroner	Prosent	Kroner	Prosent	Kroner	Prosent
Total forbruksutgift	37 964	100,0	6 930	100,0	104 423	100,0
Dagligvarer	11 065	29,1	3 418	49,3	21 709	20,8
Andre utgifter	26 899	70,9	3 512	50,7	82 714	79,2
Tallet på husholdninger ..	3 362		246		220	

Gjennomsnittlig total forbruksutgift pr. husholdning er bortimot 38 000 kroner pr. år. Det er imidlertid interessant å legge merke til at variasjonsbredden i materialet er stor. 246 husholdninger, eller vel 7 prosent, hadde under 10 000 kroner i årlig utgift - med et gjennomsnitt på bare knapt 7 000. 220 husholdninger, dvs. knapt 7 prosent, hadde total forbruksutgift på over 80 000 kroner pr. år, med et gjennomsnitt på over 104 000 kroner. Videre ser vi at dagligvareandelen er høy for husholdninger med lav total forbruksutgift og lav for husholdninger med høy total forbruksutgift. I neste avsnitt blir resultatene forsøkt tolket.

3. Tolkning av resultatene - belyst ved regneeksempel

Som nevnt gir data fra forbruksundersøkelsen et kort og meget ufullstendig glimt av forbruket i de enkelte husholdningene. Det er særlig viktig å holde klart for seg at oppgaveperioden er kort. Det er f.eks. ikke å vente at utgifter i løpet av en vilkårlig uttrukket to-ukersperiode skal gi noe korrekt bilde av den enkelte husholdnings forbruk i det lange løp. De registrerte utgiftsbeløpene vil som regel avvike fra "langtidsgjennomsnittet" for husholdningen. Avvikene kan selvsagt være enten positive eller negative.

Vi vil nå illustrere litt nærmere denne egenskapen ved hjelp av et regneeksempel. Som vi snart skal komme tilbake til, mener vi at eksemplet - så forenklet og "rendyrket" som det er - likevel på en bra måte gjenspeiler en side ved det norske forbruksdatamaterialet som er vesentlig for den som skal analysere data.

Vi betrakter et tenkt tilfelle hvor 20 husholdninger deltar i en forbruksundersøkelse. For å gjøre eksemplet enklest mulig forutsettes at alle disse husholdningene har samme inntekt og i det lange løp samme forbruk, og at observerte forskjeller mellom husholdningene i observerte forbruksutgifter skyldes tilfeldige (kortsiktige) avvik. Inntekten antas å være ukjent.

Vi deler inn utgiftene i to grupper:

1. Utgifter med relativt små tilfeldige avvik (varer som kjøpes ofte, her kalt "dagligvarer").
2. Utgifter med store tilfeldige avvik (sjeldenkjøpsvarer, bl.a. kjøp av store varige forbruksgoder, her kalt "andre utgifter").

1) Matvarer, drikkevarer og tobakk.

Det forutsettes at utgifter til dagligvarer (betegnet X_1) er lik en konstant 10 for alle husholdningene med tillegg av et stokastisk ledd U_1, altså for husholdning nr. i

$$X_{1i} = 10 + U_{1i} \qquad (i = 1, 2, \ldots, 20)$$

Leddet U_1 antas å være normalfordelt med forventning 0 og standardavvik 1. Tilsvarende forutsettes andre utgifter (betegnet X_2) å ha fordelingen

$$X_{2i} = 10 + U_{2i} \qquad (i = 1, 2, .., 20)$$

hvor U_2 er et normalfordelt ledd med forventning 0 og standardavvik 3. Når det gjelder avhengighet mellom leddene U, U_2 er disse forutsatt å være ukorrelerte. Det er neppe realistisk å anta at leddene i et faktisk forbruksmateriale er ukorrelerte. Det er imidlertid ett element i hver av leddene som antakelig med nonenlunde reimelighet kan antas å være ukonelerte, nemlig det som har med tilfeldig variasjon over tiden å gjøre (det beror for en stor del på "rene" tilfeldigheter om utgifter til f.eks. "andre utgifter" har vært store innenfor en vilkårlig kortere regnskapsperiode). Dette drøftes forøvrig nærmere i neste avsnitt.

Ved hjelp av tilfeldig genererte tall har vi simulert 20 uavhengige observasjonssett (tabell 2).

Tabell 2. Utgifter til dagligvarer, andre utgifter og total forbruksutgift for hver observasjon. Regneeksempel

Observasjon nr.	Total forbruksutgift	Dagligvarer	Andre utgifter
1	21,85	10,32	11,53
2	18,14	8,74	9,40
3	19,17	10,55	8,62
4	20,71	8,94	11,77
5	16,31	9,79	6,52
6	9,11	7,81	1,30
7	18,86	8,11	10,75
8	24,03	9,80	14,23
9	21,92	10,57	11,35
10	19,86	10,82	9,04
11	20,06	9,82	10,24
12	20,54	8,38	12,16
13	19,74	11,09	8,65
14	18,21	9,14	9,07
15	15,44	8,74	6,70
16	17,66	9,43	8,23
17	20,84	9,97	10,87
18	26,08	9,84	16,24
19	11,39	8,96	2,43
20	23,48	9,70	13,78

Gjennomsnitt X_1 = 9,53
Standardavvik X_1 = 0,87

Gjennomsnitt X_2 = 9,64
Standardavvik X_2 = 3,62

Total forbruksutgift i dette regneeksemplet varierer fra 9,11 opp til 26,08. En får altså betydelige variasjoner i utgifter som følge av variasjonen i U_1 og U_2. Den totale forbruksutgiften

20 og standardavvik $\sqrt{1^2 + 3^2} = \sqrt{10} \approx 3{,}2$, siden U_1, og U_2 er forutsatt å være ukonelerte. Som vi husker fra tabell 1 har der ca. 7 prosent av husholdningene en total forbruksutgift på under 10 000 kroner, og nesten like amnge over 80 000 kroner pr. år. Det er etter vår mening vanskelig å trekke noen konklusjoner om velstandsfordelingen blant norske husholdninger på grunnlag av nevnte tabell. Med den registreringsmåten som ligger til grunn for datamaterialet mener vi at en må regne med meget betydelige tilfeldige variasjoner i forbrukstallene, og at tabellen derfor vanskelig kan nyttes til dette formål. Det er antakelig generelt vanskelig å nytte et slikt materiale til å anslå fraktiler i fordelingen av det private forbruket (kanskje bortsett fra medianen).

Tabell 3 viser regneeksemplets utgiftstall og budsjettandeler for husholdninger med forskjellig total forbruksutgift.

Tabell 3. Gjennomsnittlig utgift pr. husholdning i grupper for total forbruksutgift, etter utgiftsgruppe. Regneeksempel

Utgiftsgruppe	Alle	Total forbruksutgift	
		20 og under	Over 20
		Beløp	
Total forbruksutgift	19,17	16,72	22,16
Dagligvarer	9,53	9,38	9,70
Andre utgifter	9,64	7,34	12,46
		Prosent	
Total forbruksutgift	100,0	100,0	100,0
Dagligvarer	49,7	56,1	43,8
Andre utgifter	50,3	43,9	56,2
Tallet på husholdninger ...	20	11	9

Andelen til dagligvarer er omtrent 50 prosent for alle husholdninger sett under ett. (Ved et økende antall observasjoner vil den selvsagt gå mot akkurat 50 prosent.)

For husholdninger med total forbruksutgift under 20 er dagligvareandelen 56,1 prosent, mot 43,8 prosent for dem med total forbruksutgift over 20. (Andre utgifter varierer selvsagt motsatt.) Dette kommer her av at en lav total forbruksutgift gjerne er en følge av lave andre utgifter mens en høy totalutgift helst skyldes at gruppen andre utgifter er høy. Dette har selvsagt sammenheng med at variasjonsbredden for gruppen andre utgifter (for U_2) er forutsatt å være stor, mens den er liten for dagligvarer. (Variansen på U_2 er stor; på U_1 er den liten.)

I tabell 1 ble det vist at dagligvareandelen er høy for husholdninger med lav total forbruksutgift, men lav for husholdninger med høy totalutgift. (Variansberegninger som er foretatt viser at variasjonen er langt mindre for dagligvareutgiftene enn for andre utgifter.) Vi tror at den samme "mekanismen" som i regneeksemplet kan ha influert vesentlig på tallene og bidratt til at forskjellene i budsjettandelene er blitt såvidt markerte. Dette indikerer at man bør være litt forsiktig med å nytte resultatene til å beskrive forskjeller i forbrukeratferden for husholdninger på forskjellig økonomisk nivå.

I stedet for en tabellanisk analyse av det simulerte materialet, kunne man være interessert i å foreta regresjonsanalyse, nemlig av sammenhengen mellom økonomisk nivå og utgiftsbeløp til henholdsvis dagligvarer og andre utgifter. Som man husker var det en forutsetning ved simuleringen av forbrukstallene for de 20 husholdningene at de alle har samme inntekt og forbruk, bortsett fra tilfeldige avvik i forbruket. Uten at man behøver å drøfte noen modellspesifikasjon, er det lett å forstå - også intuitivt - at det ikke er mulig å estimere noen sammenheng mellom inntekt og forbruksutgifter når man har et materiale hvor alle husholdningene har samme inntekt.

Vi antar som eksempel at det forutsettes en lineær sammenheng mellom inntekt og forbruksutgifter. I mangel av inntektsdata kan det være nærliggende å nytte total forbruksutgift som forklaringsvariabel for henholdsvis dagligvarer og andre utgifter. Det kan f.eks. tenkes at følgende sammenheng blir spesifisert:

$$X_1 = a_1 + b_1 (X_1 + X_2) + \text{stokastisk restledd (hvor } a_1 \text{ og } b_1 \text{ betegner konstanter).}$$

Minste kvadratersestimatoren på ligningen blir

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{1i} + X_{2i} - \bar{X}_1 - \bar{X}_2)}{\sum_{i=1}^{n} (X_{1i} + X_{2i} - \bar{X}_1 - \bar{X}_2)^2}$$

hvor $\bar{X}_1$ og $\bar{X}_2$ betegner gjennomsnitt, og antall observasjoner er n. Ved å sette inn for

$$X_1 = 10 + U_1$$

og $X_2 = 10 + U_2$

og la antall obervasjoner vokse over alle grenser vil $\hat{b}$ konvergere mot grenseverdien

$$\text{Plim } \hat{b}_1 = \frac{\sigma^2_{u_1}}{\sigma^2_{u_1} + \sigma^2_{u_2}}$$

hvor σ_{u_1} og σ_{u_2} betegner (teoretisk) standardavvik for u_1 og u_2.

I dette eksemplet blir grenseverdien for den engeldervierte (den marginale konsumtilbøyeligheten) for dagligvarer lik $\frac{1}{10}$. Samme beregning for gruppen andre utgifter gir en engelderivert lik $\frac{9}{10}$. Vi finner m.a.o. at de estimerte marginale konsumtilbøyeligheter ved et tilstrekkelig antall observasjoner blir proporsjonale med variansene på restleddene for de forskjellige utgiftsgruppene.

Regneeksemplet er "ekstremt", særlig på den måten at alle husholdningene er forutsatt å ha samme inntekt. Resultatene kan derfor ikke på noen måte tas som tallmessig uttrykk for hvor betenkelig det er å nytte totalutgiften som grupperings-/forklaringsvariabel i praksis. Poenget har vært i fremheve et moment som neppe alltid blir tilstrekkelig påaktet ved bruk av datamateriale fra forbruksundersøkelser og ved tolkning av resultatene.

4. Forbrukeratferd og datamateriale

I avsnitt 2 betraktet vi en tabell fra forbruksundersøkelsen 1973, og i avsnitt 3 en omtrent tilsvarende tabell fra et simulert datamateriale på 20 husholdninger. Som et ledd i vurderingen av om talleksemplet gir en bra beskrivelse av virkeligheten, er det naturlig å starte med en drøfting av forbrukernes atferd. Dette gjelder spørsmålet om "på hvilken måte" forbrukerne tilpasser forbruket til en gitt inntekt.

Stort sett har det vel vært vanlig å anta at tilpassingsmekanismen kan tenkes å foregå på to prinsippielt forskjellige måter[1]:

- forbrukerne fastlegger sin totale forbruksutgift når de har en gitt inntekt til disposisjon. Innenfor denne rammen fordeler de så midlene mellom kjøp av de forskjellige vare- og tjenestegruppene.
- forbrukerne fastlegger forbruket av de forskjellige vare- og tjenestegruppene ut fra en gitt inntekt. Totalutgiften blir så summen av de foskjellige vare- og tjenestegruppene.

Det er ikke lett å finne holdepunkter for hva som er den typiske konsumentatferden. Det kan selvsagt [også] tenkes kombinasjoner av de to måtene, f.eks. at husholdningen først setter av mer eller mindre på forhånd bestemte beløp til "nødvendige" og "faste" utgifter, f.eks. matvare- og

1) Se f.eks. Econometrica 27 (1959) p. 121-129, R. Summers, med kommentar av S. J. Prais.

boligutgifter, og at det som så blir igjen av inntekten (etter evt. sparing) deretter fordeles mellom de øvrige vare- og tjenestegruppene.

Det er vel imidlertid nærliggende at husholdningen må ha en nokså stramt avgrenset total ramme for forbruksutgiftene in mente når den foretar de forskjellige anskaffelser, eller foretar handlinger som medfører betalingsforpliktelser. Dette gjelder når man betrakter et tidsrom av en viss lengde. Og nettopp tidshorisonten er vesentlig ved studium av forbruksatferden. På svært kort sikt setter ikke inntekten noen streng grense for forbruket, f.eks. som følge av evt. tidligere oppsparte midler, opptak av lån, redusering av utgiftene i den nærmest påfølgende tiden osv.

Dersom man antar at husholdningene først fastlegger totalutgiften og deretter fordeler midlene mellom de forskjellige vare- og tjenestegruppene, burde det stort sett være en viss negativ korrelasjon mellom de tilfeldige avvikene i utgiftene til gruppene - i regneeksemplet mellom U_1 og U_2. Dersom f.eks. en husholdning har uvanlig høye utgifter til dagligvarer i en periode, skulle altså dette ofte gi seg utslag i uvanlig lave andre utgifter.

I kap III avsnitt 10 er det gjort forsøk på å estimere korrelasjonen mellom de stokatiske leddene i forbruksmaterialet for 1973. Resultatene der tyder ikke på negative korrelasjoner mellom leddene. Legg merke til den begrensning som ligger i utsagnet; man finner ikke negative korrelasjoner når utgiftene er registrert som i forbruksundersøkelsene, med blant annet en kort regnskapsperiode. Dette sier imidlertid ikke så mye om hvordan husholdningenes innkjøpsatferd kan tenkes å arte seg betraktet over et lengre tidsrom. Det kunne her ha vært relevant i skille mellom "forbruksteori" og "kjøpsteori".

Med den korte registreringsperioden som er i våre forbruksundersøkelser mener vi simuleringen i avsnitt 2 fremhever et poeng som det er av stor betydning å ta i betraktning ved analyse av forbruksmaterialet som vi samler inn.

Det er ikke foreløpig gjort noe særlig forsøk på å presisere hva slags "avvik" det har vært snakk om, f.eks. om avvikene bør forstås som såkalte restledd (avvik for eksakt relasjon) etter målefeil. Dette er hittil utelatt, i det vi mener at det ikke er noe hovedpoeng. Uansett hva man kaller avvikene, bør konklusjonen være at man må være varsom med å nytte totalutgiften som grupperingsvariabel (forklaringsvariabel) ved analyse av våre forbruksdata. Advarselen gjelder uansett analysemåte, f.eks. såvel tabelloppstillinger som regresjonsanalyser.

I kapittel III vil vi se nærmere på bruk av total forbruksutgift som forklaringsvariabel ved forbruksanalyser - når inntekstsdata mangler.

III. ESTIMERING NÅR DET IKKE FORELIGGER INNTEKTSDATA FOR DE INDIVIDUELLE HUSHOLDNINGENE

1. Innledning

Det ble i kapittel II manet til forsiktighet med bruk av total forbruksutgift for gruppering etter økonomisk nivå ved analyser av utgifter til forskjellige vare- og tjenestegrupper. Det er imidlertid litt for lettvint bare å avfeie en metode som ubrukbar. Det er ikke så sjelden - etter vår erfaring - at inntektsdata rett og slett mangler for de husholdninger man har forbruksdata for. Og hva skal man da gjøre? I slike situasjoner er det ikke alltid like interessant å diskutere særlig grundig om en fremgangsmåte er god eller mindre god i en viss "absolutt forstand". Målet må alltid være å klare seg så godt man kan med det datamateriale som er tilgjengelig.

Regneeksemplet i kapittel II er noe "spesielt" forsåvidt som en der tenker seg at alle husholdningene har samme forbruk (bortsett fra tilfeldige avvik). Slike tenkte eksempler kan ofte være effektive for å avsløre om en metode generelt sett er holdbar, altså om den fører til sanne resultater på alle tenkelige datasett. Eksemplene behøver imidlertid ikke alltid gi et korrekt bilde av en metodes mangelfullhet når den bare nyttes på "rimelige" datasett dvs. datasett som man kan komme ut for i praksis.

I det følgende vil vi søke å finne anslag på inntektsfordelingen for populasjonen av norske husholdninger, som utvalget til forbruksundersøkelsen 1973 er trukket blant. Deretter simuleres et prøvemateriale på 300 observasjoner med den samme relative inntektsfordelingen. Prøvematerialet simuleres under forutsetning av at modellen er en enkel lineær relasjon mellom innetkt og forbruksutgifter

med altenative (kjente) talleksempler på koeffisientene, foruten tilfeldige tall for restleddene. På dette prøvematerialet blir det foretatt "prøveestimeringer" med total forbruksutgift som forklaringsvariabel, hvoretter estimatene på koeffisientene sammenlignes med de sanne (kjente) koeffisientene som er lagt til grunn ved simuleringen av materialet. Hovedhensikten med dette er å få et visset inntrykk av hvor galt man (med vårt materiale) kommer ut ved i nytte total forbruksutgift i stedet for inntekt som forklaringsvariabel.

Deretter vil det analystisk bli sett nærmere på mulighetene for å estimere inntektsderiverte når en har kjennskap til inntektens variasjon i populasjonen (uten å kjenne inntekten for den enkelte husholdning i utvalget). Det er også foretatt noen empiriske beregninger på data fra 1973-undersøkelsen.

2. Modellen

Når en skal nytte tallmateriale til å beregne "styrken" i sammenhengen mellom variable, er det først nødvendig å gjøre forutsetninger om hvilken eller hvilke typer sammenhenger som gjelder. Slike forutsetninger kan selvsagt gjøres uten å skjele til noe datamateriale. En behøver f.eks. ikke å ta hensyn til om variabelverdiene i modellen er tilgjengelige eller ikke for å kunne spesifisere en modell. Det er en annen sak at selve estimeringen i høy grad kan bli vanskeliggjort som følge av manglende eller mangelfulle data.

Det er ikke så sjelden å se at modeller "justeres" etter data. En ønsker f.eks. å estimere sammenhengen mellom inntekt og forbruk, men mangler brukbare inntektsdata. En bestemmer seg da kanskje for å nytte total forbruksutgift i stedet for inntekt. Det vil ofte være ønskelig å undersøke hva en slik "justering" innebærer, spesielt om den innebærer en "annen" modell.

Vi innfører følgende symboler:

X_j' betegner faktisk (ikke observerbart) årsforbruk av vare- og tjenestegruppe j

X_j " beregnet årsutgift på grunnlag av bl.a. 14 dagers regnskapsføring av husholdningen (observert forbruk)

$X_j'' = X_j - X_j'$

X_j'' betegner altså avviket mellom observert årsforbruk og det faktiske årsforbruket av vare- og tjenestegruppen. X_j'' er et stokastisk ledd med forventning null; og oppfattes som en tilfeldig målefeil i forbruket. Det forutsettes at det ikke er andre typer målefeil i forbruksdataene.

Begrepet faktisk årsforbruk kan gis flere tolkninger. Årsforbruket kan f.eks. oppfattes som den faktiske årsutgiften. Men det kan også tolkes mer "bokstavelig", som direkte anvendelse av ikke-varige forbruksgoder (som ikke behøver være identisk med anskaffelsene), slitasje på varige gjenstander etc. Uten å gjøre noe poeng ut av dette, vil vi forenkle resonnementet ved å si at årsforbruket er identisk med den faktiske årsutgiften.

Vi forutsetter at utgiften til en vare- og tjenestegruppe kan uttrykkes som en lineær funksjon av forbruksdisponibel inntekt (etter skatt), (R):

$$X_j' = a_j + b_j R + E_j \qquad \text{(III.2.1.)}$$

hvor a_j og b_j betegner konstanter og E_j et restledd. Ved å erstatte den uobserverbare X_j' med X_j, kan dette skrives

$$X_j = a_j + b_j R + (E_j + X_j'') \qquad \text{(III.2.2.)}$$

For enkelhets skyld innføres symbolet

$$U_j = E_j + X_j'' \qquad (j = 1, 2, ..,m)$$

og dermed

$$X_j = a_j + b_j R + U_j \qquad \text{(III.2.3.)}$$

U_j består altså av to ledd. Det ene, E_j, kan tolkes som et restledd i tradisjonell forstand, mens det andre, X_j'' er målefeilleddet.

Når det gjelder fordelingsegenskapene på restleddet U_j, forutsettes det foreløpig forventning null og konstant varians. Videre forutsettes det uavhengighet mellom restleddene, dvs. kovar $(U_j, U_k) = 0$ $(j \neq k)$. Forutsetningene blir drøftet nærmere i avsnitt 10 i dette kapitlet. Vi vil

der innføre visse modifikasjoner angående restleddsegenskapene.

En må regne med at det i praksis er adskillige andre forhold enn inntekten som influerer på forbruket og dets sammensetning, f.eks. type husholdning. For at modellen skal gi en god beskrivelse av virkeligheten kan den bare anvendes innenfor husholdningsgrupper som er noenlunde homogene m.h.t. disse andre variablene, f.eks. når det gjelder husholdningstype.

3. Inntektsfordelingen i forbruksundersøkelsesmaterialet 1973

For å kunne simulere et datamateriale noe nær likt forbruksmaterialet må man kjenne eller være i stand til å anslå inntektsfordelingen for husholdningene i utvalget eller for populasjonen som utvalget er trukket blant, dvs. alle private norske husholdninger. Dette siste kunne tenkes gjort på grunnlag av inntektsstatistikk. Den innteksdefinisjonen som er nyttet i vår inntektsstatistikk, er imidlertid ikke særlig godt egnet for dette formål.

Det hadde vært ønskelig å ha kunnet anslå fordelingen av husholdningene etter den forbruksdisponible inntekten. Dette har imidlertid ikke vært mulig. I stedet har vi måttet nøye oss med å anslå fordelingen etter en beregnet faktisk total forbruksutgift. I vedlegg 1 er det gjort rede for hvordan dette er gjort. Det er imidlertid vanskelig å vite om den benyttede metode er den rimligste eller beste. Poenget er bare at fordelingen av husholdningene tilnærmet skal tilsvare den man finner i den norske befolkningen. (Måten å anslå fordelingen på er ikke i seg selv noe poeng i denne artikkelen.)

Uttrykt ved symbolene i avsnitt II.2. får vi faktisk total forbruksutgift

$$R = \sum_{j=1}^{m} X_j' \qquad \text{(III.3.1.)}$$

Dette begrepet vil senere bli lagt til grunn ved estimering av utgiftsderiverte/elastisiteter. Definisjonssammenhengen innebærer at

$$X_j = a_j + b_j \sum_{j=1}^{m} X_j + E_j \quad (j = 1, 2, \ldots, m)$$

og

$$\sum_{j=1}^{m} E_j = 0$$

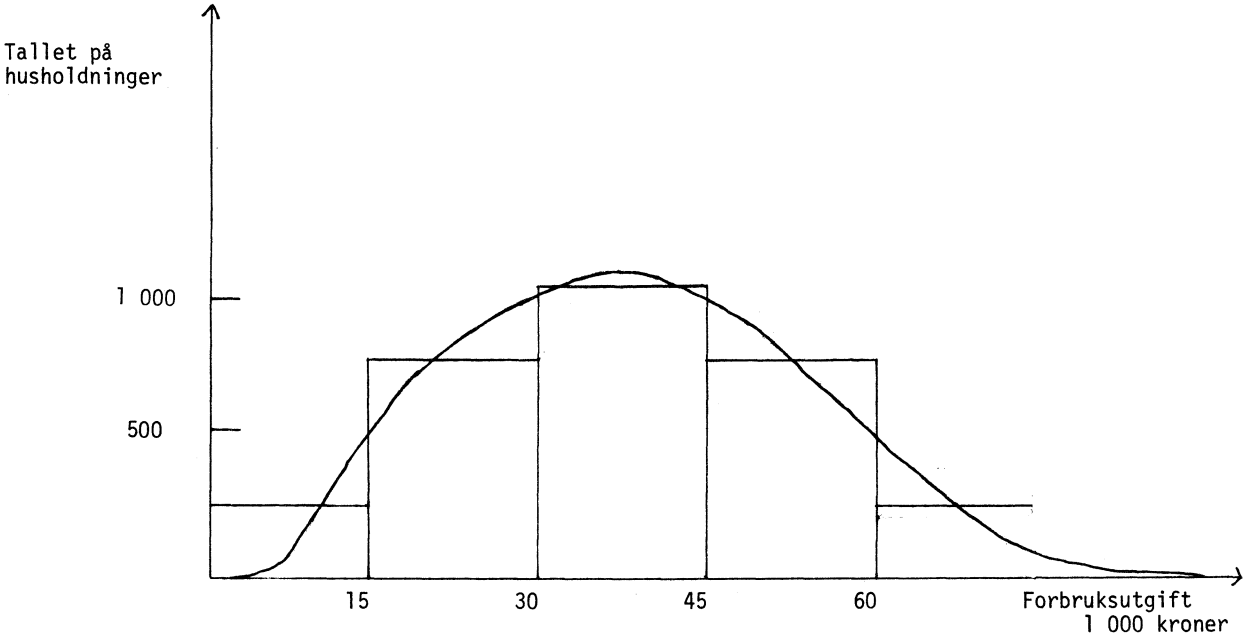
Når R er definert på denne måten, vil en få avhengighet mellom leddene E_j. (Ved to vare- og tjenestegrupper: $E_1 = - E_2$. Dette må antas å berøre rimeligheten av å forutsette at U-ene er uavhengige av hverandre. Dette er behandlet i avsnitt 11 i dette kapitlet.)

Tabell 4 og figuren med histogramm viser den estimerte fordelingen ved forbruksundersøkelsesmaterialet for 1973.

Tabell 4. Tallet på husholdninger etter beregnet faktisk forbruksutgift pr. år

Faktisk forbruksutgift pr. år	Tallet på husholdninger	
	Absolutte tall	Prosent
I alt	3 362	100,0
Under kr 15 000	252	7,5
Kr 15 000 - 30 000	879	26,1
" 30 000 - 45 000	1 133	33,7
" 45 000 - 60 000	879	26,1
" 60 000 og over	219	6,5

Figur. Beregnet inntektsfordeling



4. Simulering av prøvemateriale

Vi betrakter modellen i avsnitt 2 (III,2.3) med den forskjell at vi benytter faktisk årlig total forbruksutgift i stedet for forbruksdisponibel inntekt (jfr. III, 3.1.). I forbindelse med simulering av prøvematerialet har vi sørget for å få en relativ fordeling av husholdningene etter inntekt lik den estimerte fordelingen ved forbruksundersøkelsen 1973. Vi har delt inn forbruket i to vare- og tjenestegrupper, den gruppen man for øyeblikket er konsentrert om, og resten. I det følgende er (de observerte) utgiftsbeløpene symbolisert ved henholdsvis X_1, X_2 og total forbruksutgift ved $X = X_1 + X_2$. Restleddene i det simulerte materialet (U_1 og U_2) er forutsatt å være normalfordelte, med forventning null og konstant varians. Det forutsettes videre at restleddene er ukorelerte med hverandre.

Vi velger nivåer på a, b og standardavvik på restleddene, og simulerer deretter X_1- og X_2-verdiene for alle 300 enhetene via

$$X_1 = a_1 + b_1 R + U$$

$$X_2 = a_2 + b_2 R + U_2,$$

ved hjelp av tilfeldige tall for leddene U_1 og U_2. Av (III, 3.1.) følger at

$$X = X_1 + X_2 = (X_1' + X_1'') + (X_2' + X_2'') = R + (X_1'' + X_1'')$$

Siden X_1'' og X_2'' har forventing null, blir

$$b_1 + b_2 = 1 \qquad \text{og } a_1 + a_2 = 0$$

Gjennomsnittlig inntekt er satt lik 40 i prøvematerialet. Standardavviket på restleddet for total forbruksutgift er satt lik 20. Dette tilsvarer omtrent forholdet mellom standardavviket på restleddet til total forbruksutgift og nivået for gjennomsnittlig total forbruksutgift i forbruksundersøkelsesmaterialet 1973 (se vedlegg 1 og 2). Variansene på restleddene til utgiftsgruppene betegnes henholdsvis $\sigma_{U_1}^2$ og $\sigma_{U_2}^2$, hvor variansen på restleddet til total forbruksutgift σ_U^2 er

$$\sigma_U^2 = \sigma_{U_1}^2 + \sigma_{U_2}^2 \quad \text{eller}$$

$$\sigma_U = \sqrt{\sigma_{U_1}^2 + \sigma_{U_2}^2}$$

Vi har her gitt a_1, b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$ følgende alternative verdier:

$$a_1 = -10,\ 0,\ 10,$$

$$b_1 = 0{,}00,\ 0{,}05,\ 0{,}20,\ 0{,}50$$

$$\frac{\sigma_{U_1}}{\sigma_U} = 0{,}05,\ 0{,}20,\ 0{,}50,\ 0{,}70$$

og har simulert tall for X_1, X_2 (og dermed X) for alle kombinasjoner av disse verdiene. Noen av kombinasjonene er vel nokså sannsynlige i praktiske tilfelle, mens andre er mer ekstreme. Dette blir kommentert i neste avsnitt.

5. Estimater på de simulerte datasettene

Vi danner funksjoen

$$X_1 = \alpha_1 + \beta_1 X + V_1 \qquad \text{(III. 5.1.)}$$

hvor α_1 og β_1 betegner konstanter og V_1 et restledd[1]. Vi nytter minste kvadraters estimater for α_1 og β_1 (henholdsvis $\hat{\alpha}_1$ og $\hat{\beta}_1$) som anslag på a_1 og b_1. Tabell 5 viser resultatene fra det simulerte materialet ved forskjellige verdier på a_1, b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$.

Tabell 5. Verdier på $\hat{\alpha}_1$ og $\hat{\beta}_1$ ved forskjellige verdier på a_1, b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$. 300 observasjoner

a_1=	b_1=	$\frac{\sigma_{U_1}}{\sigma_U}$ = 0,05		0,20		0,50		0,70	
		$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\alpha}_1$	$\hat{\beta}_1$
-10	0,00	-10,1	0,003	-11,1	0,026	-15,7	0,137	-20,7	0,262
"	0,05	- 9,0	0,024	-10,0	0,047	-14,5	0,158	-19,6	0,284
"	0,20	- 5,6	0,088	- 6,6	0,111	-11,1	0,223	-16,1	0,349
"	0,50	1,3	0,215	0,3	0,238	- 4,3	0,351	- 9,3	0,478
0	0,00	- 0,1	0,003	- 1,1	0,026	- 5,7	0,137	-10,7	0,262
"	0,05	1,0	0,024	0,0	0,047	- 4,5	0,158	- 9,6	0,284
"	0,20	4,4	0,088	3,4	0,111	- 1,1	0,223	- 6,1	0,349
"	0,50	11,3	0,215	10,3	0,238	5,7	0,351	0,7	0,478
10	0,00	9,9	0,003	8,9	0,026	4,3	0,137	- 0,7	0,262
"	0,05	11,0	0,025	10,0	0,047	5,5	0,158	0,4	0,284
"	0,20	14,4	0,088	13,4	0,111	8,9	0,223	3,9	0,349
"	0,50	21,3	0,215	20,3	0,238	15,7	0,351	10,7	0,478

Det første vi legger merke til er at estimatet $\hat{\beta}_1$ er upåvirket av konstantleddet a_1. Dette kan også lett vises analystisk. Derimot varierer $\hat{\alpha}_1$ med nivået på b_1. Vi er her primært interessert i å undersøke hvordan $\hat{\beta}_1$ synes å fungere som estimator på b_1 ved forskjellig nivå på b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$.

(Forøvrig, dersom man først har funnet et bra anslag på b_1, er det lett å innse hvordan a_1 kan anslås, nemlig ved $\hat{a}_1 = \bar{X}_1 - \hat{b}_1 \bar{X}$)

Vi legger merke til at $\hat{\beta}_1$ gir bra anslag på b_1 ved visse kombinasjoner av b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$, nemlig

$b_1 = 0{,}05$ og $\frac{\sigma_{U_1}}{\sigma_U} = 0{,}20$ $\qquad (\hat{\beta}_1 = 0{,}047)$

$b_1 = 0{,}20$ og $\frac{\sigma_{U_1}}{\sigma_U} = 0{,}05$ $\qquad (\hat{\beta}_1 = 0{,}223)$

$b_1 = 0{,}50$ og $\frac{\sigma_{U_1}}{\sigma_U} = 0{,}70$ $\qquad (\hat{\beta}_1 = 0{,}478)$

1) V_1 kan skrives $V_1 = (a_1 - d_1) + (b_1 - \beta_1) R + (U_1 - \beta_1 (X_1'' + X_2''))$

Alle disse kombinasjonene tilsvarer omtrent at

$$(\frac{\sigma_{U_1}}{\sigma_U})^2 \approx b_1, \text{ eller } \frac{\sigma^2_{U_1}}{b_1} \approx \sigma^2_U.$$

Dette gjelder selvsagt hvilken som helst vare- og tjenestegruppe. Det ser altså ut som at forholdet mellom restleddsvariansen og den engelderiverte må være omtrent den samme for alle vare- og tjenestegruppene dersom en skal kunne regne med at teknikken gir pålitelige resultater. (Jamfør også omtalen av regresjonsanalysen i kap. II, avsnitt 3.) Det er ikke lett å si noe sikkert om i hvilken grad dette spesielle kravet er oppfylt i vårt forbruksmateriale. Restleddsvariansen avhenger i stor grad av husholdningenes innkjøpsrutiner for de forskjellige vare- og tjenestegruppene. Det er vel i alle fall vanskelig å kunne godta at det alltid gjelder noen slik sammenheng mellom innkjøpsrutiner og marginal konsumtilbøyelighet. Og hvis denne forutsetningen er dårlig oppfylt, kan estimatene bli meget unøyaktige.

I neste avsnitt er det foretatt en analytisk betraktning av sammenhengen mellom $\hat{\beta}_1$ og parameteren b_1.

6. Den engelderiverte for en varegruppe og bruk av total forbruksutgift som forklaringsvariabel, belyst analytisk

Vi tar utgangspunkt i modellen (III.2.3.) med 2 vare- og tjenestegrupper og relasjonen (III.5.1.)

$$X_1 = \alpha_1 + \beta_1 X + V_1$$

Estimering ved minste kvadraters metode gir følgende estimator for β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

hvor fotskift i betegner husholdning m i og n er antall observasjoner. Vi setter inn

$$X_{1i} = a_1 + b_1 R_i + U_{1i},$$

$$X_{2i} = a_2 + b_2 R_i + U_{2i} \quad \text{og}$$

$$X_{1i} + X_{2i} = X_i \qquad (i = 1, 2, .., n),$$

og får

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} [b_1 (R_i - \bar{R}) + (U_{1i} - \bar{U}_1)] [(b_1 + b_2)(R_i - \bar{R}) + (U_{1i} - \bar{U}_1) + (U_{2i} - \bar{U}_2)}{\sum_{i=1}^{n} [b_1 + b_2)(R_i - \bar{R}) + (U_{1i} + U_{2i} - \bar{U}_1 - \bar{U}_2)]^2}$$

Det er visse problemer med å finne eksakte uttrukk for forventning og varians lik $\hat{\beta}_1$ siden det er stokastikk både i telleren og nevneren. Vi vil derfor her nøye oss med å studere de asymtotiske egenskapene. Vi forutsetter at sentralmomentet for inntekten - M^2_R - konvergerer mot en konstand (m^2_R) når antall observasjoner vokser over alle grenser, altså

$$\lim M^2_R = m^2_R$$

Siden U_1 og U_2 er antatt uavhengige av hverandre og uavhengige av R, vil $\hat{\beta}_1$ med økende antall observasjoner gå mot grenseverdien[1)]

$$P \lim \hat{\beta}_1 = \frac{b_1 (b_1 + b_2) m_R^2 + \sigma_{U_1}^2}{(b_1 + b_2)^2 m_R^2 + \sigma_U^2}$$

Ved å sette inn $b_1 + b_2 = 1$ får vi

$$P \lim \hat{\beta}_1 = \frac{b_1 m_R^2 + \sigma_{U_1}^2}{m_R^2 + \sigma_U^2} \qquad \text{(III.6.1.)}$$

Dersom $\sigma_{U_1}^2$, σ_U^2 og m_R^2 (eller i det minste brøkene $\frac{\sigma_{U_1}^2}{m_R^2}$ og $\frac{\sigma_U^2}{m_R^2}$ er kjent eller kan anslås (symboler for anslagene:

$$\hat{\sigma}_{U_1}^2, \hat{\sigma}_U^2 \text{ og } \hat{m}_R^2)$$

er det nærliggende å nytte uttrykket

$$\hat{b}_1 = \frac{(\hat{m}_R^2 + \hat{\sigma}_U^2) \cdot \hat{\beta}_1 - \hat{\sigma}_{U_1}^2}{\hat{m}_R^2}$$

som estimator for $b_1 \cdot \hat{b}_1$ vil være en asymtotisk forventningsrett og konsistent estimator dersom anslagene er riktige.

Samme beregning for den andre varegruppen gir

$$\hat{b}_2 = \frac{(\hat{m}_R^2 + \hat{\sigma}_U^2) \cdot \hat{\beta}_2 - \hat{\sigma}_{U_2}^2}{\hat{m}_R^2}$$

I uttrykkene for $\hat{b}_1$ og $\hat{b}_2$ ligger det innebygget at

$$\hat{b}_1 + \hat{b}_2 = 1.$$

Dette følger av at $\hat{\beta}_1$ og $\hat{\beta}_2$ er konstruert slik at

$$\hat{\beta}_1 + \hat{\beta}_2 = 1$$

sammen med at

$$\hat{\sigma}_{U_1}^2 + \hat{\sigma}_{U_2}^2 = \hat{\sigma}_U^2.$$

I tabell 6 er det vist eksempel på hvilke estimater for b_1 denne estimeringsteknikken kan gi ved det (endelige) antall observasjoner vi har i det simulerte prøvematerialet, nemlig 300 observasjoner (med de kjente samme verdiene på $\sigma_{U_1}^2$, σ_U^2 og M_R^2).

---

1) Framstillingen avviker ikke på noe vesentlig punkt fra den som er gitt av R. Summers i Econometrica (1959) p. 121-126.

Tabell 6. Estimerte verdier for b_1 ($\hat{b}_1$) ved forskjellige verdier på b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$ · 300 observasjoner

b_1=	$\frac{\sigma_{U_1}}{\sigma_U}$			
	0,05	0,20	0,50	0,70
	$\hat{b}_1$	$\hat{b}_1$	$\hat{b}_1$	$\hat{b}_1$
0,00	0,003	0,0009	0,001	-0,012
0,05	0,050	0,055	0,048	0,035
0,20	0,190	0,196	0,189	0,178
0,50	0,470	0,476	0,473	0,464

Vi finner i dette tilfelle at estimatet $\hat{b}_1$ ligger i "nonenlunde rimelig" nærhet av parameteren b_1 for alle kombinasjoner av b_1 og $\frac{\sigma_{U_1}}{\sigma_U}$ selv med bare 300 observasjoner. (Med et større antall observasjoner ville anslagene selvsagt blitt bedre. Det samme gjelder dersom variansene på restleddene hadde vært mindre.)

Etter at vi har funnet estimatet for b_1, kan estimatet for a_1 finnes ved

$$\hat{a}_1 = \bar{X}_1 - \hat{b}_1\bar{X}$$

Det springende punktet i spørmålet om å nytte $\hat{b}_1$ og b_2 som estimatorer er selvsagt hvorvidt $\sigma_{U_1}^2$, σ_U^2 og m_R^2 lar seg anslå.

For nærmere å vurdere denne estimeringsteknikken, har vi i vedlegg 3 sett på formelen for varianser til $\hat{b}_1$. Det asymtotiske formeluttrykket blir

$$\operatorname{var} \hat{b}_1 = \frac{(b_1^2 \hat{m}_R^2 + \hat{\sigma}_{U_1}^2)(\hat{m}_R^2 + \hat{\sigma}_U^2) - (b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{n\,\hat{m}_R^4}$$

Det kan være nyttig å undersøke hva teknikken med bruk av total forbruksutgift som høyresidig variabel medfører for variansen på $\hat{b}_1$ sammenlignet med om man hadde kjent R for de individuelle husholdninger og dermed kunne ha estimert direkte på ligningen

$$X_1 = a_1 + b_1 R + U_1$$

I sistnevnte tilfelle får man følgende minste kvadratestimator:

$$b_1^* = \frac{\sum_{i=1}^{n} (X_{1_i} - \bar{X}_1)(R_i - \bar{R})}{\sum_{i=1}^{n} (R_i - \bar{R})^2}$$

Denne estimatoren er forventningsrett og har varians

$$\operatorname{var} b_1^* = \frac{\sigma_{U_1}^2}{n\,m_R^2}$$

når vi setter inn $\lim \sum_{i=1}^{n} (R_i - \bar{R})^2 = n \cdot m_R^2$. Det er ikke så lett å sammenligne uttrykkene. Vi vil derfor se på tilfellene med talleksemplene på b_1, m_R^2, σ_U^2 og $\sigma_{U_1}^2$ fra tabellene foran. Tabell 7 viser

forholdet mellom standardavvikene, dvs. $\sqrt{\text{var } \hat{b}_1}$ $\sqrt{\text{var } b_1^*}$.[1)]

Tabell 7. Talleksempler på forholdet $\sqrt{\text{var } \hat{b}_1}$ / $\sqrt{\text{var } b_1^*}$ fra det simulerte materialet.

b_1 =	$\frac{\sigma_{U_1}}{\sigma_U}$			
	0,05	0,20	0,50	0,70
0,00	1,57	1,47	1,39	1,27
0,05	1,87	1,45	1,35	1,24
0,20	4,47	1,67	1,30	1,14
0,50	10,63	2,72	1,39	1,06

Tallene i tabellen viser at teknikken med å bruke total forbruksutgift som høyresidig variabel i mange tilfelle øker standardavvikene på estimatene for b_1 vesentlig. Dette tyder på at det ikke er "like bra" å estimere b_1 via totalutgiften som direkte med hensyn på inntekten (hvis man også kjenner inntekten). I prinsippet kan kjennskapen til et slikt forhold være av betydning når en under planlegging av en undersøkelse skal vurdere utvalgsstørrelse, ekstra kostnader ved å skaffe inntektsopplysninger e.l.

Forutsetningen for å kunne estimere b_1 via bruk av total forbruksutgift er som sagt foruten at sentralmomentet for inntekten (M_R^2) er kjent at restleddsvariansene ($\sigma_{U_1}^2$, $\sigma_{U_2}^2$) også er kjente eller kan anslås. Det vil forekomme at estimater på variansene i visse tilfelle er kjente, f.eks. gjennom erfaring fra andre undersøkelser. Via kjennskap til innkjøpsrutiner for de forskjellige vare- og tjenestegruppene er det kanskje også mulig å foreta kvalifiserte gjetninger på variansene. Men nokså ofte vil det være forbundet med store vanskeligheter å lage gode anslag. I neste avsnitt er det vist hvordan b_1 og b_2 og restleddsvariansene i prinsippet kan estimeres simultant dersom sentralmomementet for inntekten er kjent på forhånd.

7. Estimering av alle parametrene i fordelingen

Vi tar utgangspunkt i formel (III.6.1.). Videre betraktes ligning (III.5.1.):

$$X_1 = \alpha_1 + \beta_1 X + V_1$$

dvs. $$V_1 = X_1 - \alpha_1 - \beta_1 X$$

og tilsvarende for varegruppe 2. De empiriske variansene på residualene V_1 og V_2 finnes ved minste kvadraters metode på

$$s_{V_1}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \hat{\alpha}_1 - \hat{\beta}_1 X_i)^2$$

og $$s_{V_2}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_{2i} - \hat{\alpha}_2 - \hat{\beta}_2 X_i)^2$$

Ved å sette inn for X_1 og X_2 og la antall observasjoner vokse, vil $s_{V_1}^2$ og $s_{V_2}^2$ gå mot henholdsvis

1) Vi betrakter hele tiden asymtotiske grenseverdier. Siden $\hat{b}_1$ ikke er forventningsrett ved et endelig antall observasjoner, kunne det vært av interesse å studdere bruttovariansene. Dette er ikke forsøkt her.

$$P \lim S_{V_1}^2 = b_1^2 m_R^2 - 2b_1 \beta m_R^2 + \beta_1^2 m_R^2 + \sigma_{U_1}^2 - 2\beta_1 \sigma_{U_1}^2 + \beta_1^2 \sigma_U^2 \quad \text{og}$$

$$P \lim S_{V_2}^2 = b_2^2 m_R^2 - 2b_2 \beta_2 m_R^2 + \beta_2^2 m_R^2 + \tau_{U_2}^2 - 2\beta_2 \sigma_{U_2}^2 + \beta_2^2 \sigma_U^2$$

Dersom en på forhånd kjenner m_R^2, kan man i prinsippet finne anslag for b_1, b_2, $\sigma_{U_1}^2$, $\sigma_{U_2}^2$ og σ_U^2 ved å å løse ligningene:

1. $$\hat{\beta}_1 = \frac{\hat{b}_1 m_R^2 + \hat{\sigma}_{U_1}^2}{m_R^2 + \hat{\sigma}_U^2}$$

2. $$\hat{\beta}_2 = \frac{\hat{b}_2 m_R^2 + \hat{\sigma}_{U_2}^2}{m_R^2 + \hat{\sigma}_U^2} \qquad 1)$$

3. $$\hat{\sigma}_U^2 = \hat{\sigma}_{U_1}^2 + \hat{\sigma}_{U_2}^2$$

4. $$S_{V_1}^2 = \hat{b}_1^2 m_R^2 - 2 \hat{b}_1 \hat{\beta}_1 m_R^2 + \hat{\beta}_1^2 m_R^2 + \hat{\sigma}_{U_1}^2 - 2\hat{\beta}_1 \hat{\sigma}_{U_1}^2 + \hat{\beta}_1 \hat{\sigma}_U^2$$

5. $$S_{V_2}^2 = \hat{b}_2^2 m_R^2 - 2 \hat{b}_2 \hat{\beta}_2 m_R^2 + \hat{\beta}_2^2 m_R^2 + \hat{\sigma}_{U_2}^2 - 2 \hat{\beta}_2 \hat{\sigma}_{U_2}^2 + \hat{\beta}_2 \hat{\sigma}_U^2$$

hvor $\hat{b}_1$, $\hat{b}_2$, $\hat{\sigma}_{U_2}^2$, $\hat{\sigma}_{U_2}^2$ og $\hat{\sigma}_U^2$ betegner estimater for de respektive parametre. Vi har altså 5 ligninger til å anslå de 5 ukjente størrelsene. Det finnes åpenbart flere løsninger. Spørsmålet er om det er mulig å avgjøre i praksis hvilken løsning som er rimelig og om metoden generelt er anvendbar.

Ved denne teknikken er det mye som er "ukjent" og tilsvarende mye som materialet "må gi svar på". Dette stiller på en måte store krav til materialet; og det er sterkt ønskelig med mange observasjoner. Det er ikke utenkelig at man i så fall ville kunne anslå hvilken (av flere) løsninger som er den rimelige. Løsningen vil vel lettest kunne foretas ved interasjon.

Når man skal vurdere om en eller annen estimeringsmetode er tjenlig eller ikke, må dette gjøres i lys av hvilke alternativer som foreligger. Jo mer informasjon man har på forhånd (kjenner f.eks. i dette tilfelle standardavvikene på restleddene), jo bedre vil det naturligvis ofte være. (Det vil ofte kunne skje at man ikke har noen nytte av et materiale. Man har kanskje kunnskaper fra annet hold (f.eks. fra annet materiale) som er så gode eller fullstendige at et nytilkommet materiale har lite å gi i tillegg).

1) $\hat{\beta}_1 + \hat{\beta}_2 = 1$

8. Total forbruksutgift fratrukket utgifter til vedkommende vare- og tjenestegruppe som forklaringsvariabel

Vi betrakter fortsatt modell (III.2.3.) med to vare- og tjenestegrupper

$$X_1 = a_1 + b_1 R + U_1$$

$$X_2 = a_2 + b_2 R + U_2$$

I stedet for å benytte total forbruksutgift ($X = X_1 + X_2$) under estimeringen, er det mulig å bruke total forbruksutgift fratrukket utgifter til vedkommende varegruppe som forklaringsvariabel, altså

$$X_1 = \gamma_1 + \delta_1 X_2 + W_1$$

og $$X_2 = \gamma_2 + \delta_2 X_1 + W_2$$

hvor γ-ene og σ-ene betegner konstanter og W_1 og W_2 restledd[1]. Dersom en på tilsvarende måte som i avsnitt 6 estimere σ_1 og σ_2 ved minste kvadraters metode, setter inn for X_1 og X_2 og lar antall observasjoner vokse, vil uttrukkene gå mot henholdsvis

$$P\lim \hat{\delta}_1 = \frac{b_1 b_2 m_R^2}{b_2^2 m_R^2 + \sigma_{U_2}^2}$$

og $$P\lim \hat{\delta}_2 = \frac{b_1 b_2 m_R^2}{b_1^2 m_R^2 + \sigma_{U_1}^2}$$

Innsetting i 1. ligning med

$$b_2 = 1 - b_1$$

og løsning m.h.p. b_1 gir uttrykket

$$\hat{b}_1^* = \frac{2\hat{\delta}_1 + 1}{2(\hat{\delta}_1 + 1)} \pm \sqrt{\left[\frac{2\hat{\delta}_1 + 1}{2(\hat{\delta}_1 + 1)}\right]^2 - \frac{m_R^2 \hat{\delta}_1 + \hat{\delta}_2^2 \hat{\sigma}_2^2}{m_R^2 (1 + \hat{\delta})}}$$

Det kunne her ha vært av interesse å sammenligne variansen på $\hat{b}_1^*$ med variansen på $\hat{b}_1$ fra avsnitt 6. Dette har imidlertid ikke vært mulig fordi uttrykket for $\hat{b}_1^*$ har en såvidt "uhåndterlig" form.

9. Modell med flere forklaringsvariable enn inntekt

Vi tar utgangspunkt i modell (III.2.3.), bare med den forskjell at det kan være flere forklaringsvariable enn inntekt. Vi ser her på tilfellet med én variabel (y, som antas observerbar).

$$X_1 = a_1 + b_1 R + c_1 y + U_1$$

$$X_2 = a_2 + b_2 R + c_2 y + U_2$$

1) $W_1 = X_1 - (\gamma_1 + \delta_1 a_2) - \delta_1 b_2 R - \delta_1 U_2$ og tilsv. for W_2.

Den øvrige spesifikasjon av modellen er som i avsnitt 2. Det kan her være nærliggende å ta utgangspunkt i følgende ligning:

$$X_1 = \alpha_1 + \beta_1 X + H_1 Y + V_1$$

hvor H_1 betegner en konstant, og de øvrige symboler har samme betydning som i avsnitt 6 i dette kapitlet.

Det viser seg at estimering av b-ene (og c-ene) etter metoden i avsnitt 6 krever at kryssmomentet mellom R og Y i populasjonen (i tillegg til sentralmomentet for R) er kjent. Det er antakelig i praksis nokså ofte at slike kryssmomenter ikke er kjent. Dessuten blir beregningene temmelig innfløkte teknisk sett.

Det kan se ut til at evt. estimering av inntekts-/utgiftselastinntekter via teknikken med total forbruksutgift som forklaringsvariabel ofte vil måtte begrenses til modeller med bare én høyresidig variabel (dvs. inntekten).

10. Ikke konstant varians på restleddene og avhengighet mellom leddene. Modifisert modell

I modellen vår er det forutsatt at

1. variansene på restleddene er konstante (homoscedastisitet)
2. restleddene er uavhengige av hverandre.

I dette avsnittet vil vi vurdere disse forutsetningene nærmere.

Det er i forskjellige sammenhenger nokså vanlig å anta at restleddsvariasjonen ofte vil variere med nivåene på de absolutte tallstørrelsene, f.eks. med nivået på den avhengige variable. I dette tilfelle kan man tenke seg at variasjonen i U (i absolutte tall) er større innenfor husholdningsgrupper med et høyt enn med et lavt forbruk.

Vi har studert spredningen på residualene ved bruk av et beregnet inntektsmål (vedlegg 2). Dette er gjort for forskjellige inntektsklasser. Tabell 8 viser samvariasjoner mellom standardavvik på residualene og gjennomsnittlig total forbruksutgift.

Tabell 8. Standardavvik på residualen for total forbruksutgift, beregnet gjennomsnittsutgift, forholdet mellom standardavviket og den beregnede gjennomsnittsutgiften, i forskjellige utgiftsklasser. Kroner

		Beregnet total forbruksutgift				
	I alt	Under 15 000	15 000-29 999	30 000-44 999	45 000-59 999	60 000 og over
Standardavvik på residualen	21 070	13 153	15 000	19 000	25 554	34 612
Gjennomsnittsutgift	37 957	13 103	23 952	38 298	50 929	68 934
Standardavvik/gjennomsnittsutgift	0,56	1,00	0,62	0,50	0,50	0,50
Tallet på husholdninger	3 362	252	879	1 133	879	219

Tabellen viser at standardavviket på residualen øker med totalutgiften. For utgifter over 30 000 tyder tabellen på at standardavviket på residualen øker proporsjonalt med utgiften. For lavere utgiftsnivåer viser tabellen mindre relativ økning i standardavviket på residualen.

Det kan synes som om det ikke er helt rimelig å forutsette at standardavviket på restleddet er konstant. I de følgende avsnittene i dette kapitlet vil vi forutsette at standardavvik på restleddene for alle vare- og tjenestegruppene varierer proporsjonalt med inntekten. Variansen til U_1 og U_2 kan da skrives

$$\text{var } U_1 = R^2 \tau^2_{U_1}$$

og

$$\text{var } U_2 = R^2 \tau^2_{U_2}$$

hvor τ_{U_1} og τ_{U_2} betegner konstanter. For en gruppe husholdninger blir de gjennomsnittlige variansene

$$\frac{1}{n}\sum_{i=1}^{n} \text{var } U_{1i} = \frac{1}{n}\Sigma R_i^2 \tau_{U_1}^2 = (M_R^2 + \bar{R}^2)\,\tau_{U_1}^2$$

og

$$\frac{1}{n}\sum_{i=1}^{n} \text{var } U_{2i} = \frac{1}{n}\sum_{i=1}^{n} R_i^2 \tau_{U_2}^2 = (M_R^2 + \bar{R}^2)\,\tau_{U_2}^2$$

når M_R^2 og $\bar{R}$ betegner sentralmoment og gjennomsnitt for R.

I modellen i avsnitt 2 er det forutsatt at restleddene (U-ene) på utgiftene til de forskjellige vare- og tjenestegruppene er stokastisk uavhengige av hverandre. I dette avsnittet forsøker vi å vurdere rimeligheter i denne forutsetningen. (Formeluttrykkene som hittil er nyttet forutsetter uavhengighet.)

Med to vare- og tjenestegrupper får vi

$$U_1 = X''_1 + E_1$$

$$U_2 = X''_2 + E_2$$

$$E_1 = -E_2$$

Det antas at alle leddene har forventning lik null. Kovariansen mellom U_1 og U_2 blir

$$\text{kovar}\,(U_1\,U_2) = E\,(X''_1 + E_1)\,(X''_2 + E_2) =$$

$$E\,X''_1\,X''_2 + E\,X''_1\,E_2 + E\,X''_2\,E_1 + E\,E_1E_2$$

Her er $E\;E_1E_2 = -E\,E_1^2$. Vi mener at det ikke er urimelig å anta at leddene $E\,X''_1E_2$ og $E\,X''_2E_1$ ligger i noenlunde nærhet av null. Det er vanskelig å se hvilken "mekanisme" som skulle kunne gjøre dette usannsynlig. Forutsetningen kovar $(U_1U_2) = 0$ betinger dermed at

$$EX''_1X''_2 = E\,E_1^2$$

Leddet $E\,X''_1X''_2$ må altså være positivt.

Det kan antakelig reises tvil om en slik forutsetning. Det er tvert i mot ikke utenkelig at det kan være en viss negativ korelasjon mellom X"-ene, ut fra følgende tankegang: Hvis en husholdning har hatt store utgifter til f.eks. klær og skotøy i en periode, kan dette av budsjetthensyn føre til mindre utgifter til f.eks. fritidssysler i samme periode (for gitt inntekt). Dersom relativt mange husholdninger lever under et stramt budsjett (også på kort sikt), kan denne effekten tenkes å være av betydning. Det kan derfor være av interesse å se på korrelasjonen mellom restleddene i forbruksundersøkelsen 1973 for de forskjellige vare- og tjenestegruppene.

Fra forige avsnitt er det forutsatt at restleddsvariansene har følgende form:

$$\text{var } U_1 = R^2\,\tau_{U_1}^2$$

$$\text{var } U_2 = R^2\,\tau_{U_2}$$

Det er da også nærliggende å anta at tilsvarende form gjelder for kovariansen,

$$\text{kovar } U_1\,U_2 = R^2\tau_{U_1U_2}$$

hvor $\tau_{U_1U_2}$ er en konstant. Korrelasjonskoeffisienten mellom U_1 og U_2 blir

$$\text{Korr } U_1U_2 = \frac{R^2\tau_{U_1U_2}}{R\,\tau_{U_1}\,R\,\tau_{U_2}} = \frac{\tau_{U_1U_2}}{\tau_{U_1}\,\tau_{U_2}}$$

Korrelasjonskoeffisienten blir altså en konstant, uavhengig av R.

Det er ikke noen enkel oppgave å undersøke korelasjonen mellom U_1 og U_2 på en fullgod måte for å se om en modell med uavhengighet mellom leddene virker realistisk i vårt datamateriale. Det er her gjort forsøk ved å studere korrelasjonen mellom residualene slik disse er estimert i vedlegg 2.

Tabell 9 viser de empiriske korrelasjons koeffisientene mellom residualene for 9 vare- og tjenestegrupper i forbruksundersøkelsen 1973.

Tabell 9. Korrelasjonskoeffisienter mellom residualer i forbruksundersøkelsen 1973[1)]

Vare- og tjenestegruppe	Vare- og tjenestegruppe								
	0	1	2	3	4	5	6	7	8
0 Matvarer	1,00								
1 Drikkevarer og tobakk	0,29	1,00							
2 Klær og skotøy	0,17	0,17	1,00						
3 Bolig, lys og brensel	0,05	0,05	0,03	1,00					
4 Møbler og husholdningsartikler ...	0,17	0,19	0,18	0,14	1,00				
5 Helsepleie	0,05	0,00	0,03	0,05	0,01	1,00			
6 Reiser og transport	0,04	0,06	0,10	0,05	0,05	0,01	1,00		
7 Fritidssysler og utdanning	0,14	0,14	0,11	0,03	0,13	0,05	0,11	1,00	
8 Andre varer og tjenester	0,11	0,18	0,19	0,04	0,11	0,01	0,09	0,13	1,00

1) 3 362 husholdninger.

I motsetning til hva en kanskje skulle tro er alle koeffisientene positive (bortsett fra én som er lik null). Det er ikke lett å vite noe sikkert om hva som er årsaken til at korrelasjonene er positive. Det kan være nærliggende å tenke seg at den spesifiserte modellen er for "enkel", f.eks. ved at en del ikke helt uviktige forklaringsvariable inkluderes i restleddet.

Den positive korrelasjonen kan imidlertid også ha andre årsaker, f.eks. at det kan foreligge visse andre typer målefeil enn dem som er forutsatt å ligge i X"-ene. Her skal bare nevnes én bestemt type feil som vil medføre positiv korrelasjon mellom leddene: Det krever en god del innsats av husholdningene å regnskapsføre alle utgifter i to uker. Det varierer fra husholdning til husholdning hvor påpasselig man er med å få med alle utgiftene i regnskapsheftene. En god del husholdninger vil kunne få registrert omtrent alle sine utgifter, mens de regnskapsførte utgiftene vil være vesentlig mindre enn faktiske utgifter i perioden for andre. Slike variansjoner mellom husholdninger i påpasselighet med utgiftsføringen vil vel antakelig gjelde for alle vare- og tjenestegrupper. Det er nokså lett å innse at en da vil få positiv kovarians mellom restleddene for de forskjellige vare- og tjenestegruppene. (Dette kan formaliseres og vises klarere, men er ikke tatt med her.)

Det kan antakelig reises tvil om hvorvidt korrelasjonskoeffisientene mellom residualene gir et godt uttrykk for korrelasjonen mellom restleddene, med den beregningsmåten som er nyttet.

I modellen har vi delt forbruket i to vare- og tjenestegrupper: Den gruppen vi undersøker og "resten". Tabell 10 viser korrelasjonskoeffisienter mellom residualene for de forskjellige vare- og tjenestegrupper og residualene for "resten" av utgiftene (dvs. total forbruksutgift fratrukket utgifter til vedkommende vare- og tjenestegruppe).

Tabell 10. Korrelasjonskoeffisienter mellom residualer for 9 vare- og tjenestegrupper og for resten av utgiftene ved forbruksundersøkelsen 1973[1)]

Vare og tjenestegruppe	Korrelasjonskoeffisienter
0 Matvarer	0,22
1 Drikkevarer og tobakk	0,27
2 Klær og skotøy	0,24
3 Bolig, lys og brensel	0,11
4 Møbler og husholdningsartikler	0,24
5 Helsepleie	0,05
6 Reiser og transport	0,14
7 Fritidssysler og utdanning	0,22
8 Andre vare og tjenester	0,21

1) 3 362 husholdninger.

Siden det er positiv korrelasjon mellom residualene for de forskjellige vare- og tjenestegruppene, får vi selvsagt også positiv korrelasjon mellom residualen for én gruppe og for "resten" av utgiftene.

Det er et meget vanskelig problem å finne ut hva en skal gjøre dersom et datamateriale ikke "synes å stemme" med en gitt modell. For å vite hva en "egentlig" skulle gjøre måtte man kjenne grunnen til uoverensstemmelsen (f.eks. feilspesifikasjon, målefeil etc.).

En annen mulig forklaring på positiv kovarians mellom restleddene i tillegg til de nevnte er følgende: Innkjøp i husholdningene er neppe jevnt fordelt over tiden; innkjøpene varierer f.eks. med lønningsdager (f.eks. for månedslønte). Siden utgiftene til alle varegruppene regnskapsføres i én og samme 2-ukersperiode for samme husholdning, kan dette tenkes å føre til positiv korrelasjon mellom leddene (X"-ene).

Alt i alt mener vi at resultatene tyder på at det er noe betenkelig å forutsette at U-ene er ukorrelerte, og at det derfor er ønskelig å operere med en modell som tillater at det er kovarians mellom restleddene (U-ene).

Vi innfører nå følgende modifikasjoner av modellen i avsnitt 2:

$$\text{var } U_2 = R^2 \tau_{U_1}^2$$

$$\text{var } U_2 = R^2 \tau_{U_2}^2$$

$$\text{kovar } U_1U_2 = R^2 \tau_{U_1U_2}$$

Dette gir

$$\text{var } U = \text{var } (U_1 + U_2) = \text{var } U_1 + 2 \text{ kovar } U_1U_2 + \text{var } U_2 =$$

$$R^2 (\tau_{U_1}^2 + 2 \tau_{U_1U_2} + \tau_{U_2}^2) = R \tau_U^2$$

Med disse modifikasjonene får vi i stedet for formel (III.6.2.) uttrykket

$$\hat{b}_1 = \frac{\left[\hat{m}_R^2 + (\hat{m}_R^2 + \hat{\bar{r}}) \hat{\tau}_U^2\right] \hat{\beta}_1 - (\hat{m}_R^2 + \hat{\bar{r}}^2) (\hat{\tau}_{U_1} + \hat{\tau}_{U_1U_2})}{\hat{m}_R^2}$$

Ved å gå fram på analog måte som i vedlegg 3 kan det vises at variansen blir

$$\text{var } \hat{b}_1 = \frac{\left[b_1 \hat{m}_R^2 + (\hat{m}_R^2 + \hat{\bar{r}}^2) \hat{\tau}_{U_1}^2\right] \left[\hat{m}_R^2 + (\hat{m}_R^2 + \hat{\bar{r}}^2) \hat{\tau}_U^2\right]}{n \hat{m}_R^4}$$

$$- \frac{\left[b_1 \hat{m}_R^2 + (\hat{m}_R^2 + \hat{\bar{r}}^2)(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2}) \right]^2}{n \hat{m}_R^4}$$

I avsnitt 7 ble det vist hvordan parametrene i modellen i avsnitt 2 i prinsippet kan estimeres simultant. Det samme kan også gjøres i denne modellen. I tillegg til de ligningene som er nyttet der nyttes uttrykket for kvariansen mellom leddene V_1 og V_2:

$$SV_1V_2 = \frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \hat{\alpha}_1 - \hat{\beta}_1 X_i)(X_{2i} - \hat{\alpha}_2 - \hat{\beta}_2 X_i)$$

(Se vedlegg 4.) Dette gir en ekstra ligning som (i prinsippet kan nyttes til å estimere parameteren $\tau_{U_1U_2}$. Dersom m_R^2 og $\bar{r}^2$ på forhånd er kjente, blir det altså 6 ligninger å løse for å estimere de 6 ukjente parametrene b_1, b_2, $\tau_{U_1}^2$, $\tau_{U_2}^2$, $\tau_{U_1U_2}$ og τ_U^2. Dette må vel gjøres ved bruk av iterasjon. Det er alvorlig grunn til å reise tvil om metodens anvendbarhet i praksis (jfr. kommentarer i avsnitt 6 i dette kapitlet).

11. Empiriske resultater fra forbruksundersøkelsen 1973

Under presentasjonen av modellen i avsnitt 2 ble det antydet at den enkle funksjonsformen

$$X_j = a_j + b_j R + U_j \qquad (j = 1, 2, .., m)$$

kunne være nonenlunde realistisk, forutsatt at man betrakter grupper av husholdninger som er nokså homogene med hensyn på andre variable (enn inntekten) som kan være av betydning for forbruket.

Det er ikke problemfritt å finne fram til grupper av husholdninger som er så homogene at den enkle funksjonsformen med rimelighet kan antas å ha god mening. En kunne kanskje tenke seg alle én-personshusholdninger som et eksempel på en slik gruppe. Det er imidlertid store variasjoner i forbruksmønstret også innen denne gruppen, variasjoner som ikke primært har sammenheng med inntekten. Når det f.eks. gjelder fordelingen av utgiftene, viser enkle tabelloppstillinger riktignok store variasjoner med inntekten, men hvor inntekten neppe er hovedårsaken til variasjonene. Kvinner har f.eks. høyere utgifter til matvarer enn menn. (Dette kan vel bl.a. ha sammenheng med at mange menn har et "enkelt" kosthold.) Kvinner har også høyere utgifter til klær, mens menn har større utgifter til drikkevarer og tobakk. Dersom en ønsker å nytte den enkle funksjonsformen, burde man antakelig dele inn gruppen én-personhusholdningene i undergrupper (f.eks. etter kjønn/alder). Dette vil imidlertid medføre nokså få observasjoner i hver gruppe.

Vi har i det følgende konsentrert oss om ektepar med to barn. Dette er en noenlunde homogen gruppe av rimelig størrelse (394 husholdninger i 1973-undersøkelsen), selv om f.eks. barnas alder sikkert spiller en rolle for forbrukssammenstningen, og muligens burde vært med som forklaringsvariabel.

Vi benytter formelen

$$\hat{b}_1 = \frac{\left[M_R^2 + (M_R^2 + \bar{R}^2) \hat{\tau}_U^2 \right] \hat{\beta}_1 - (M_R^2 + \bar{R}^2)(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2})}{M_R^2}$$

under estimeringen. Bruk av dette uttrykket forutsetter som før påpekt at M_R^2, $\bar{R}^2$, $\hat{\tau}_U^2$, $\hat{\tau}_{U_1}^2$ og $\hat{\tau}_{U_1U_2}$ er kjent. I dette tilfelle har vi nyttet inntektsoppgaver fra ligningsvesenet for å estimere nevnte størrelser. Estimeringsmetoden er forklart i vedleggene 1 og 2. Variansene på restleddene er i vedleggene forutsatt konstant. Vi har nyttet estimatet for σ_U^2 som anslag på den gjennomsnittlige variansen i materialet. Estimatet for ektepar med to barn blir

$$\hat{\sigma}_U^2 = (24\ 072)^2 = (M_R^2 + \bar{R}^2) \hat{\tau}_U^2$$

Videre finner en

$$M_R^2 = 5\ 591^2$$

og $$\bar{R}^2 = 46\ 403^2$$

Dette gir

$$\hat{\tau}_U^2 = 0{,}2653$$

På tilsvarende måte finnes $\hat{\tau}_{U_1}^2$, $\hat{\tau}_{U_1U_2}$ og $\hat{\tau}_{U_2}^2$. Estimatene er gjengitt i tabell 11.

Tabell 11. Estimater på konstantene $\tau_{U_1}^2$, $\tau_{U_1U_2}$ og $\tau_{U_2}^2$ i forbruksundersøkelsen 1973. Ektepar med 2 barn[1)]

	$\hat{\tau}_{U_1}^2$	$\hat{\tau}_{U_1U_2}$	$\hat{\tau}_{U_2}^2$
0 Matvarer	0,0125	0,0082	0,2362
1 Drikkevarer og tobakk	0,0017	0,0038	0,2558
2 Klær og skotøy	0,0094	0,0115	0,2327
3 Bolig, lys og brensel	0,0469	0,0138	0,1905
4 Møbler og husholdningsartikler ...	0,0090	0,0117	0,2331
5 Helsepleie	0,0019	0,0043	0,2546
6 Reiser og transport	0,0718	0,0214	0,1505
7 Fritidssysler og utdanning	0,0141	0,0078	0,2354
8 Andre varer og tjenester	0,0054	0,0101	0,2395

1) 394 husholdninger.

Tabell 12 viser estimatet $\hat{b}_1$ (med standardavvik) og gjennomsnittlig elastisitet ($\hat{b}_1 \cdot \frac{\bar{R}}{\bar{X}}$) for de forskjellige vare- og tjenestegrupper.

Tabell 12. Estimert marginal utgiftsderivert ($\hat{b}_1$) med standardavvik (i parentes) og gjennomsnittlig elastisitet ($\hat{b}_1 \cdot \frac{\bar{R}}{\bar{X}}$) for forskjellige vare- og tjenestegrupper. Ektepar med 2 barn[2)]

Vare- og tjenestegruppe	$\hat{b}_1$		$\hat{b}_1 \frac{\bar{R}}{\bar{X}_1}$
0 Matvarer	0,081	(0,20)	0,354
1 Drikkevarer og tobakk	0,026	(0,08)	0,635
2 Klær og skotøy	0,080	(0,19)	0,812
3 Bolig, lys og brensel	0,209	(0,35)	1,377
4 Møbler og husholdningsartikler ..	0,170	(0,18)	1,794
5 Helsepleie	0,006	(0,08)	0,289
6 Reiser og transport	0,303	(0,37)	1,406
7 Fritidssysler og utdanning	0,117	(0,22)	1,252
8 Andre varer og tjenester	0,014	(0,13)	0,251

2) 394 observasjoner.

Det mest iøynefallende ved denne tabellen er de meget store estimatene for standardavvikene. Ifølge tallene er ingen av de engelderiverte signifikant positive. Estimatene på standardavvikene er i særlig grad avhengig av sentralmomentet M_R^2. Det er ikke umulig at fremgangsmåten ved estimeringen av M_R^2 har ført til en viss underestimering av denne størrelsen, noe som særlig bidrar til en overestimering av standardavvikene. En del av punktestimatene virker vel intuitivt nokså rimelige. Engelelastisiteten for matvarer er f.eks. lav, bare 0,35. Derimot er estimatene for Møbler og husholdningsartikler, Reiser og transport og Fritidssysler og utdanning høyere. Disse resultatene er ikke overraskende. Det er imidlertid grunn til å understreke at estimatene er usikre, og at en måtte ha hatt vesentlig flere observasjoner om en skulle ha fått pålitelige anslag på b-ene (bl.a. p.g.a. den korte registeringsperioden for utgiftene).

12. Noen konklusjoner

I dette kapitlet har vi sett på noen muligheter for å estimere sammenhenger mellom inntekt og forbruk når det mangler innteksdata for de aktuelle husholdninger. Som en måtte vente er det ikke uten problemer å gjennomføre slik estimering. I tilfelle en meget enkel lineær konsumfunksjon med bare inntekt som forklaringsvariabel kan estimeringen i prinsippet gjennomføres via en "mellomregning" med bruk av total forbruksutgift som forklaringsvariabel, forutsatt at inntektsfordelingen og restleddsvariansene er kjent eller kan anslås. Det er også mulig (i prinsippet) å foreta en simultan estimering av inntekts-/utgiftsderiverte og restleddsvarianser dersom man kjenner sentralmoment og gjennomsnitt av husholdningsinntekten i populasjonen på forhånd.

I tilfelle vi skulle ønske å nytte en konsumfunksjon med flere forklaringsvariable enn inntekten, vil beregningene i høy grad bli vanskeliggjort. Betingelsene for å kunne gjennomføre beregningene er altså nokså strenge. Hovedkonkulsjonen må være at det er problematisk å estimere inntektsderiverte og -elastisiteter når vi ikke har inntektsoppgaver for de aktuelle husholdninger.

Det kan neppe overraske noen at det er vanskelig å estimere engelderiverte når inntektsdata mangler. Det er her heller ikke antatt at det foreligger andre variable som kan tenkes nyttet som erstatning for inntektsdata under estimeringen (utenom observerte utgiftsdata). Situasjonen kan altså være at man bare har utgiftstall, men ingen andre (for tilfellet) relevante opplysninger. Det er nokså rimelig at dette må gi begrensede muligheter for analyse.

Det er vel heldigvis sjelden i praksis at man bare har forbruksoppgaver for husholdningene. Ved forbruksundersøkelsen 1973 ble f.eks. en rekke opplysninger om husholdningene hentet inn, f.eks. størrelse og sammensetning av husholdningen, yrkesaktivitet for husholdningsmedlemmene (blant annet samlet timetall inntektsgivende arbeid pr. måned). Det må være et forsøk verd å dra nytte av slike opplysninger når man vil estimere inntekts-/utgiftsderiverte. Denne situasjonen drøftes imidlertid ikke i dette notatet, men er behandlet i ANO IO 77/45: Estimering av engelkurver ved data med målefeil.

Vedlegg 1

Anslag på innteksfordelingen

I dette vedlegget gjør vi rede for hvordan vi har gått fram for å beregne fordelingen av husholdningene etter inntekt anvendt til privat forbruk i forbruksundersøkelsesmaterialet 1973.

Via Skattedirektøren er det innhentet opplysninger om nettoinntekten ved skatteligningen og sum direkte skatter og trygdepremie for husholdningene som deltok i forbruksundersøkelsen 1973. Vi har tatt utgangspunkt i nettoinntekten[1] fratrukket skatter og trygdepremie (differansen er i det følgende symbolisert ved R^*). Dette inntektsbegrepet er et åpenbart dårlig mål for inntekt anvendt til privat forbruk for husholdningene. Beløpet utgjør gjennomsnittlig bare ca. 77 prosent av total forbruksutgift for alle husholdningene, sett under ett. Det er flere grunner til at beløpet ligger så lavt, f.eks. inngår ikke skattefrie inntekter i beløpet. Videre er det en del forbruksutgifter som er trukket fra ved beregning av nettoinntekten (f.eks. utgifter til arbeidsreiser, renter på boliglån m.v.). Det kan også nevnes at en del husholdninger er oppført med nettoinntekt lik null selv om inntekten er positiv.

Ved å ta utgangspunkt i R^* estimeres fordelingen av den disponible inntekten for husholdningene. Vi danner først regresjonen

$$X = c + d\,R^* + e_1 Z_1 + e_2 Z_2 + F$$

hvor X betegner (observert) total forbruksutgift, mens Z og Z_2 symboliserer binære variable som "samler opp" virkningen på forbruket som følge av forskjellige husholdningsstørrelser, ved at Z_1 er satt lik 1 hvis det er 3 eller 4 personer i husholdningen og null ellers, Z_2 er lik 1 hvis det er 5 eller flere personer og null ellers. (1-2 personer i husholdningen er referansegruppe.) c, d, e_1 og e_2 betegner konstanter og F et restledd. Konstantene er estimert ved minste kvadraters metode (estimater $\hat{c}$, $\hat{d}$, $\hat{e}_1$ og $\hat{e}_2$). Deretter er "beregnet faktisk årlig forbruksutgift" (R^e) for den enkelte husholdning estimert ved

$$R^e = \hat{c} + \hat{d}\,R^* + \hat{e}_1 Z_1 + \hat{e}_2 Z_2$$

R^e er således beregnet slik at gjennomsnittet for husholdningene blir lik total observert forbruksutgift. Vi har så tatt fordelingen av R^e som anslag på fordelingen av R i populasjonen.

Formålet med denne estimeringen er bare å anslå sentralmomentet M_R^2 og gjennomsnittet $\bar{R}$ i populasjonen. Vi benytter derimot ikke R^e på noen direkte måte for å estimere utgiftsderiverte.

I kapittel III avsnitt 11 har vi estimert M_R^2 og $\bar{R}$ for ektepar med to barn. Vi har der nyttet funksjonsformen

$$X = c + d\,R^* + F$$

som grunnlag ved estimeringen.

1) Nettoinntekt ved statsskatteligningen + nettoinntekt ved sjømannsskatteordningen + særfradrag.

Anslag på restleddsvarianser

For å kunne simulere prøvematerialet må vi gjøre forutsetninger om størrelsen på standardavvikene på restleddene. Det er selvsagt ikke lett å finne noe nøyaktig anslag på standardavvikene.

Vi har tatt utgangspunkt i relasjonen fra vedlegg 1

$$X = c + d\,R^{*} + e_1\,Z_1 + e_2\,Z_2 + F$$

hvor X i dette tilfelle symboliserer total forbruksutgift. Dessuten har vi nyttet den estimerte utgiften[1)]

$$\hat{X} = \hat{c} + \hat{d}\,R^{*} + \hat{e}_1\,Z_1 + \hat{e}_2\,Z_2.$$

Spredningen på residualene $(X - \hat{X})$ er nyttet som anslag på restleddsvariansene:

$$\hat{\sigma}_U^2 = \frac{1}{n-T}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2$$

hvor fotskrift i betegner husholdning nr. i. $\hat{\sigma}_U^2$ er i forbruksmaterialet 1973 funnet å være

$$\hat{\sigma}_U^2 = (21\,079)^2$$

På samme måte har vi estimert restleddsvarianser for forskjellige vare- og tjenestegrupper. Kovarianser mellom restleddene for forskjellige vare- og tjenestegrupper, f.eks. gruppe j og k er anslått ved kryssmomentene

$$\frac{1}{n-T}\sum_{i=1}^{n}(X_{ji} - \hat{X}_{ji})\,(X_{ki} - \hat{X}_{ki})$$

Det kan nevnes at tilsvarende beregninger er utført med en noe mer "omfattende" modell. Vi trakk inn antall voksne og antall barn og fordelingen antall menn/antall kvinner som forklaringsvariable til forbruket (ved siden av inntekten R^{*}). Vi nyttet i tillegg også samlet ukentlig arbeidstid for husholdningen og binærvariable for hovedinntektstakerens yrkesstatus. I alt var det 7 forklaringsvariable. Dette syntes imidlertid ikke å føre til andre resultater enn det vi fant ved å nytte den enklere modellen.

I avsnitt 11 har vi beregnet restleddsvarianser for gruppen ektepar med to barn. Vi har der benyttet funksjonsformen

$$X = c + d\,R^{*} + F$$

som grunnlag ved estimering av restleddsvariansene.

1) Den estimerte inntekten er definisjonsmessig lik estimert total forbruksutgift, altså $R^{e} = \hat{X}$ (jfr. vedlegg 1).

Standardavvik på estimatoren

Vi ønsker å finne variansen for uttrykket

$$\hat{b}_1 = \frac{(m_R^2 + \sigma_U^2)\ \hat{\beta}_1 - \sigma_{U_1}^2}{m_R^2}$$

For å kunne nytte $\hat{b}_1$ som estimator for b_1 må selvsagt størrelsene m_R^2, σ_U^2 og $\sigma_{U_1}^2$ forutsettes på forhånd å være kjent. (Symboler: $\hat{m}_R$, $\hat{\sigma}_U^2$ og $\hat{\sigma}_{U_1}^2$). Variansen blir:

$$\text{var } \hat{b}_1 = \left[\frac{\hat{m}_R^2 + \hat{\sigma}_U^2}{\hat{m}_R^2} \right] . \text{ var } \hat{\beta}$$

Av ligningen $X_1 = \alpha_1 + \beta_1 X + V_1$ følger det at variansen til $\hat{\beta}_1$ er lik

$$\text{var } \hat{\beta}_1 = \frac{\text{var } V_1}{\sum_{i=1}^{n} (X - \bar{X})^2} .$$

Samme ligning gir at

$$V_1 = X_1 - \alpha_1 - \beta_1 X$$

Variansen til V_1 blir

$$\text{var } V_1 = \text{var } X_1 + \beta_1^2 \text{ var } X - 2\ \beta_1 \text{ kovar } X_1 X.$$

Ved å sette inn $X_1 = a_1 + b_1 R + U_1$,

$$X = R + U_1 + U_2$$

og sette variansen til R lik $\hat{m}_R^2$

får vi:

$$\text{var } V_1 = \text{var } (a_1 + b_1 R + U_1) + \beta_1^2 \text{ var } (R + U_1 + U_2)$$

$$- 2\ \beta_1 \text{ kovar } \left[(a_1 + b_1 R + U_1)\ (R + U_1 + U_2) \right] =$$

$$b_1^2 \cdot \hat{m}_R^2 + \hat{\sigma}_{U_1}^2 + \beta_1^2\ (m_R^2 + \hat{\sigma}_U^2) -$$

$$2\ \beta_1\ (b_1 \hat{m}_R^2 + \hat{\sigma}_{U_1}^2)$$

I stedet for β_1 setter vi inn den anslåtte asymtotiske verdien for $\hat{\beta}_1$:

$$\text{P lim } \hat{\beta}_1 = \frac{b_1\ \hat{m}_R^2 + \hat{\sigma}_{U_1}^2}{\hat{m}_R^2 + \hat{\sigma}_U^2}$$

og får:

$$\text{var } V_1 = b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2 + \frac{(b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{(\hat{m}_R^2 + \hat{\sigma}_U^2)^2}(\hat{m}_R^2 + \hat{\sigma}_U^2)$$

$$- 2\frac{b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2}{\hat{m}_R^2 + \hat{\sigma}_U^2}(b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2) =$$

$$b_1^2\hat{m}_R^2 + \hat{\sigma}_1^2 + \frac{(b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2 - 2(b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{\hat{m}_R^2 + \hat{\sigma}_U^2} =$$

$$\frac{(b_1^2\hat{m}_R^2 + \hat{\sigma}_{U_1}^2) \cdot (\hat{m}_R^2 + \hat{\sigma}_U^2) - (b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{\hat{m}_R^2 + \hat{\sigma}_U^2}$$

Vi setter også inn $\sum_{i=1}^{n}(X_i - \bar{X}) = n \cdot (\hat{m}_R^2 + \hat{\sigma}_U^2)$

og får:

$$\text{var } b_1 = \left[\frac{m_R^2 + \hat{\sigma}_U^2}{\hat{m}_R^2}\right]^2 \cdot \frac{(b_1^2\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)(\hat{m}_R^2 + \hat{\sigma}_U^2) - (b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{(\hat{m}_R^2 + \hat{\sigma}_U^2) \cdot n(\hat{m}_R^2 + \hat{\sigma}_U^2)} =$$

$$\frac{(b_1^2\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)(\hat{m}_R^2 + \hat{\sigma}_U^2) - (b_1\hat{m}_R^2 + \hat{\sigma}_{U_1}^2)^2}{n\,\hat{m}_R^4}$$

Simultan estimering

Ved ikke konstant varians og avhengighet mellom restleddene får en 6 ligninger til estimering av 6 konstanter (b_1, b_2, $\tau^2_{U_2}$, $\tau_{U_1U_2}$, τ^2_U):

1. $$\hat{\beta}_1 = \frac{\left[m_R^2 + (m_R^2 + \bar{r}^2)\,\hat{\tau}_U^2 \right] \hat{b}_1 - (m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2})}{m_R^2}$$

2. $$\hat{\beta}_2 = \frac{\left[m_R^2 + (m_R^2 + \bar{r}^2)\,\hat{\tau}_U^2 \right] \hat{b}_2 - (m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_2}^2 + \hat{\tau}_{U_1U_2}}{m_R^2}$$

3. $$\hat{\tau}_U^2 = \hat{\tau}_{U_1}^2 + 2\,\hat{\tau}_{U_1U_2} + \hat{\tau}_{U_2}^2$$

4. $$S_{V_1}^2 = m_R^2\,(\hat{b}_1 - \hat{\beta}_1)^2 + \hat{\beta}_1^2\,(m_R^2 + \bar{r}^2)\,\hat{\tau}_U^2 + (m_R^2 + \bar{r}^2)\,\hat{\tau}_{U_1}^2$$
$$- 2\,\hat{\beta}_1\,(m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2}$$

5. $$S_{V_2}^2 = m_R^2\,(\hat{b}_2 - \hat{\beta}_2)^2 + \hat{\beta}_2^2\,(m_R^2 + \bar{r}^2)\,\hat{\tau}_U^2 + (m_R^2 + \bar{r}^2)\,\hat{\tau}_{U_2}^2$$
$$- 2\,\hat{\beta}_2\,(m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2})$$

6. $$S_{V_1V_2} = m_R^2\,(\hat{b}_1\hat{b}_2 - \hat{b}_1\,\hat{\beta}_2 - \hat{b}_2\,\hat{\beta}_1 + \hat{\beta}_1\hat{\beta}_2) + (m_R^2 + \bar{r}^2)\,\hat{\tau}_{U_1U_2}$$
$$+ \hat{\beta}_1\hat{\beta}_2\,(m_R^2 + \bar{r}^2)\,\hat{\tau}_U^2 - \hat{\beta}_2\,(m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_1}^2 + \hat{\tau}_{U_1U_2})$$
$$- \hat{\beta}_1\,(m_R^2 + \bar{r}^2)\,(\hat{\tau}_{U_1} + \hat{\tau}_{U_1U_2})$$