Arbeidsnotater

STATISTISK SENTRALBYRÅ

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

IO 77/45

29. november 1977

ESTIMERING AV ENGELDERIVERTE PÅ DATA MED MÅLEFEIL

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. DATA FRA FORBRUKSUNDERSØKELSER

Datamaterialer i frobruksundersøkelser 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 anskaffelser over en lengre periode av varer som kjøpes sjelden (f.eks. privat bil).

Av praktiske grunner er det neppe mulig å foreta noen særlig mer fullstendig kartlegging av forbruket i den enkelte husholdning som deltar, 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, sammenliknet med 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.)

Som følge av at observasjonsperioden er kort vil det være store tilfeldige variasjoner mellom husholdningene i forbruksutgifter, variasjoner som kan tenkes utelukkende å ha sammenheng med tilfeldige variasjoner fra periode til periode for den enkelte husholdning. Av denne grunn er det store problemer med å nytte total forbruksutgift i bokføringsperioden som indikator på årlig faktisk forbruk ved analyse av forbruksdataene (ved tabeller eller ved regresjonsanalyse). Dette er nærmere behandlet i Arbeidsnotat IO 77/44: Estimering av engelderiverte ved manglende inntektsdata.

I Statistisk Sentralbyrå's forbruksundersøkelser er det ofte et problem at man har inntektsoppgaver med målefeil. Dette vil i høy grad vanskeliggjøre estimering av engelderiverte. I dette notatet har vi sett nærmere på estimeringsmulighetene ved forskjellige typer målefeil.

Framstillingen er knyttet til datamateriale fra norske forbruksundersøkelser. Men dét betyr neppe at synspunkter og resultater 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.

II. ESTIMERING VED MÅLEFEIL

1. Innledning

Det er et meget stort behov for å arbeide innenfor modeller som tar hensyn til mulige målefeil i variablene. Vår erfaring med data fra forbruksundersøkelser er at man "alltid" må regne med målefeil i inntekts- og forbruksoppgavene. Og ofte er antakelig målefeilene temmelig store.

Det viser seg i praksis meget vanskelig å få pålitelige opplysninger om den forbruksdisponible inntekten for private husholdninger. Det er mange årsaker til dette. Nettoinntekten ved skattelikningen er f.eks. av flere grunner et dårlig mål for forbruksdisponibel inntekt, bl.a. fordi en del forbruksutgifter er trukket fra (reiser til og fra arbeidssted, renter på boliglån, visse spareformer m.v.). Dessuten er selvsagt ikke skattefrie inntekter med (mange former for sosiale stønader). I tillegg kommer mulige inntektsunndragelser i forbindelse med beskatningen.

Det kan tenkes mange typer målefeil, og mange forskjellige måter å karakterisere målefeil på. Vi vil forutsette en todeling av feilene, henholdsvis systematiske og tilfeldige målefeil[1)].

1) Denne todelingen er nyttet av bl.a. E. Malinvaud i Statistical Methods of Econometrics, Ch. 10 (1968).

For å forklare begrepene innføres følgende symboler:

z betegner observert variabelverdi
z' " sann "
z'' " stokastisk feilledd

Vi forutsetter at den observerte verdien kan uttrykkes som en eller annen eksakt funksjon (her symbolisert ved g (·)) av den sanne verdien og evt. andre variable, og av det stokastiske leddet, altså:

$$z = g(z', \text{evt. andre variable}) + z''$$

Her representerer g (·)-funksjonen den systematiske målefeilen, mens tilfeldige feil er representert ved z'' . Vi mener at denne modellen kan være hensiktsmessig. Den gir en viss presisering av begrepet målefeil, samtidig som kanskje de fleste typer av målefeil i praksis kan omfattes av modellen.

I avsnitt 2 vil vi se på det spesielle tilfellet at z'' = 0 for alle observasjoner, altså at vi bare har systematiske målefeil. I avsnitt 3 behandles tilfeldige målefeil. En antar da at g (·) = z', altså at z = z' + z'' . Fra avsnitt 4 og utover er forutsetningen at systematiske og tilfeldige målefeil kan opptre samtidig i materialet. Vi må vente at dette sistnevnte er det mest vanlige i forbruksundersøkelsene.

2. Systematiske målefeil

Byråets forbruksundersøkelser gjennomføres blant annet ved at husholdningene som deltar, fører alle sine forbruksutgifter i regnskapshefter i løpet av to uker. Hver husholdning får tildelt et såkalt hovedhefte. Dette føres vanligvis av det husholdningsmedlemmet som foretar størstedelen av de daglige innkjøpene. Dessuten får hver av de øvrige personene (som er 15 år eller eldre) tildelt hvert sitt hefte, hvor særlig deres personlige utgifter føres. Det kreves stor påpasselighet fra husholdningenes side om alle utgifter skal bli registrert. Vi regner med at det oppgitte forbruket stort sett er noe for lavt i forbruksundersøkelsene, og at dette særlig skyldes at en del utgifter blir glemt. Det finnes også andre grunner til at forbruket kan være feil registrert (husholdningene ønsker f.eks. å vise at de har god råd, eller at de "ikke sløser", at de gjør "fornuftige" innkjøp m.v.). Målefeilene i forbruket gir erfaringsvis ulikt utslag for de forskjellige vare- og tjenestegruppene. Feilene er større for f.eks. varegruppen drikkevarer og tobakk enn for matvarer.

Visse større varige forbruksvarer, som f.eks. privat bil, kjøpes sjelden, og vi får derfor få observasjoner i regnskapsheftene. Slike utgifter blir registrert ved intervju, hvor vi spør etter anskaffelser i løpet av 12 måneder bakover i tiden. Også her kan det forekomme målefeil. Folk husker ikke alt de har kjøpt, eller de husker kanskje ikke når et kjøp ble foretatt (om det er mer eller mindre enn ett år siden).

Som nevnt i avsnitt 1 er det mange muligheter for målefeil i inntektsoppgavene. Det gjøres av og til forsøk - ved å nytte likningsdata og andre inntektsdata - på å finne en slags forbruksdisponibel inntekt. Det må imidlertid antas at oppgavene ofte inneholder vesentlige målefeil.

Systematiske målefeil på tilgjengelige inntektsdata vil vel ofte variere med hensyn på visse bakgrunnskjennetegn for husholdninger. Målefeilene kan f.eks. være annerledes for husholdninger hvor hovedinntektstakeren er selvstendig næringsdrivende enn hvor han er lønnstaker, og kanskje atter annerledes igjen fra pensjonister. Det kan derfor tenkes at yrkesstatus bør være med som fordelingsvariabel til målefeilen, z = g (z', yrkesstatus til hovedinntektstaker m.v.). Som et alternativ til dette kan det imidlertid hende en bør ha én funksjon for hver yrkesstatusgruppe.

Det sier seg selv at det ikke er mulig å si noe generelt om hva som er en sannsynlig form på g (·)-funksjonen. Det er i praksis nødvendig å ha faktiske kunnskaper om svakheter ved det statistiske materialet. Som et eksempel betrakter vi her et tilfelle med lineær utforming av g (·)-funksjonen[1].

1) Se f.eks. E. Malinvaud, Statistical Methods of Econometrics, Ch. 10 (1968).

Vi nytter følgende symboler:

X betegner observert forbruk (av en vare-/tjenestegruppe)

X' " faktisk "

R " observert inntekt

R' " sann " (forbruksdisponibel)

Vi antar at X og R kan uttrykkes som lineære funksjoner av henholdsvis X' og R':

$$X = c + d\,X'$$

$$R = e + k\,R'$$

hvor c, d, e og k betegner (ukjente) konstanter.
Det forutsettes at R' (og dermed R) er ikke stokastisk.

Videre antas følgende enkle konsumfunksjon:

$$X' = a + b\,R' + E \quad \text{(II. 2.1.)}$$

hvor a og b betegner konstanter og E et restledd med forventning null og konstant varians. Vi vil senere innføre visse modifikasjoner når det gjelder fordelingen av restleddet (avsnitt 3 i dette kapitlet).

Estimering av parameteren b ved å nytte observerte i stedet for sanne verdier gir minste-kvadratersestimatoren

$$\hat{b} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})\,(R_i - \bar{R})}{\sum_{i=1}^{n} (R_i - \bar{R})^2}$$

hvor $\bar{X}$ og $\bar{R}$ betegner gjennomsnitt for X og R og hvor antall husholdninger er n. Ved å sette in for X og R gir dette

$$\hat{b} = \frac{d}{k} \cdot \frac{\sum_{i=1}^{n} (X'_i - \bar{X}')\,(R'_i - \bar{R}')}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2}$$

b vil her ikke være identifiserbar medmindre d og k er kjent. Forventningen til $\hat{b}$ blir:

$$\Sigma\hat{b} = b \cdot \frac{d}{k}$$

Vi ser at $\hat{b}$ er forventningsrett bare hvis d = k. Derimot spiller ikke konstantleddene noen rolle for forventningsrettheten.

På tilsvarende måte finner vi forventningen på estimatoren til (den gjennomsnittlige) elastisiteten (symbolisert ved $\hat{E}_R$):

$$E\,\hat{E}_R = b \cdot \frac{d}{k} \cdot \frac{e + k\,\bar{R}'}{c + d\,\bar{X}'}$$

Estimatoren til elastisiteten er generelt påvirket både av konstantleddene og vinkelkoeffisientene. Dersom konstantleddene er lik null, vil imidlertid estimatoren til elastisiteten være forventningsrett. Og likeledes, hvis e er "liten" i forhold til k $\bar{R}'$ og samtidig c er "liten" i forhold til d $\bar{X}'$, er $\hat{E}_R$ tilnærmet forventningsrett som estimator på elastisiteten av X' m.h.p. R'. Man må i praksis vurdere estimatorene ut fra kjennskapen til data. Det kan f.eks. tenkes at et datamateriale kan gi grunnlag for å estimere elastisiteter, men ikke deriverte, eller eventuelt omvendt.

Det kan ofte være rimelig å nytte en konsumfunksjon med flere variabler, f.eks.

$$X' = a + b\,R' + c_1\,y_1 + c_2\,y_2 + \ldots + c_r\,y_r + E$$

hvor c-ene betegner konstanter og y-ene variabelverdier. y-ene antas observerbare uten målefeil. Ved å studere momentmatrisene[1] er det lett å se at også en modell med flere variable gir en estimator for b med tilsvarende forventningsverdi som foran ($E\,\hat{b} = b \cdot \frac{d}{k}$). En finner videre at estimatorene til $c_1, c_2, \ldots, c_r$ er konsistente bare hvis d = 1 ($E\,\hat{c}_j = d\,c_j$, j = 1, 2, ..., r).

En annen utvidelse av modellen er å nytte et polynom av høyere grad enn én, f.eks.

$$X' = a + b_1\,R' + b_2\,R'^2 + E$$

I dette tilfelle vil også konstantleddene påvirke resultatene og bidra til at en ikke får identifikasjon (I de tilfelle konstantleddene kan settes lik null, finner vi ved å rette inn de observerbare størrelsene X og R i relasjonen estimatorene $\hat{b}_1$ og $\hat{b}_2$ med forventning

$$E\,\hat{b}_1 = \frac{d}{k}\,b_1 \text{ og } E\,\hat{b}_2 = \frac{d}{k^2}\,b_2)$$

Det virker her som at det ikke er mulig å si særlig mye generelt om hvordan systematiske målefeil bør behandles. Vi vil imidlertid komme tilbake til dette emnet i avsnitt 6 i dette kapitlet og behandle en metode som ser ut til å være nyttig.

3. Tilfeldige målefeil

I kapittel I er det påpekt at en må regne med betydelige tilfeldige målefeil i forbruksoppgavene som følge av at registreringsperioden er kort. Også i de tilgjengelige inntektsoppgavene er det rimelig å anta store variasjoner i målefeilen (mellom husholdninger med f.eks. samme forbruksdisponible inntekt). De observerte forbruks- og inntektsoppgavene betegnes henholdsvis X og R og tilsvarende feilledd X'' og R'' :

$$X = X' + X''$$
$$R = R' + R''$$

hvor X'' og R'' antas å ha forventning lik null for alle verdier av X' og R'. R' forutsettes å være ikke-stokastisk. Denne modellen er nokså vanlig forekommende i litteraturen om feil i variablene. Vi vil her også gjøre en vanlig forutsetning om at X'' og R'' er ukorrelerte med hverandre, og at de har konstant varians. (Denne siste forutsetningen vil bli modifisert i neste avsnitt.)

Ved å sette inn de observerte i stedet for de sanne verdiene får vi relasjonen:

$$X = a + b\,R + U \quad \text{(II. 3.1.)}$$

hvor

$$U = E + X'' - b\,R'' \quad \text{(II. 3.2.)}$$

Estimering ved minste kvadraters metode gir

$$\hat{b} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})\,(R_i - \bar{R})}{\sum_{i=1}^{n} (R_i - \bar{R})^2}$$

1) Se f.eks. H. T. Amundsen: Innføring i teoretisk statistikk, kap. 9.3. (1962).

Siden det ikke er så enkelt å finne noe eksakt uttrykk for forventningen til $\hat{b}$, vil vi se på de asymtotiske egenskapene. Vi forutsetter at sentralmomentet for inntekten (M_R^2) konvergerer mot en konstant m^2_R ved økende antall observasjoner. $\hat{b}$ vil da gå mot grenseverdien[1].

$$P\lim \hat{b} = b\left(1 - \frac{\sigma^2_{R''}}{m^2_R + \sigma^2_{R''}}\right)$$

hvor $\sigma^2_{R''}$ betegner den teoretiske variansen til det stokastiske leddet R" . b er ikke identifiserbar og $\hat{b}$ vil konvergere mot en størrelse som (i tallverdi) er mindre enn b (dersom det er målefeil i inntekten ($\sigma^2_{R''} > 0$)). Vi ser også at jo større variansen på målefeilen ($\sigma^2_{R''}$) er i forhold til spredningen på inntekten ($m^2_{R'}$), jo større fare er det for at b blir underestimert når $\hat{b}$ nyttes.

Vi ser videre at dersom $m^2_{R'}/\sigma^2_{R''}$ er kjent (eller kan anslås) vil uttrykket

$$\hat{\hat{b}} = \frac{\hat{b}}{\frac{m^2_{R'}}{m^2_{R'} + \sigma^2_{R''}}}$$

kunne nyttes som (konsistent) estimator for b. Det kan imidlertid i praksis være vanskelig å finne et godt anslag på $m^2_{R'}/\sigma^2_{R''}$.

Vi vil her se nærmere på $\hat{b}$. For enkelhets skyld innføres symbolet

$$K = \frac{\sigma^2_{R''}}{m^2_{R'} + \sigma^2_{R''}}$$

for den asymtotiske ("teoretiske") brøken, mens vi symboliserer tilsvarende størrelse ved et endelig antall observasjoner med $\hat{K}$. Siden R" er stokastisk, vil også $\hat{K}$ være det.

$\hat{K}$ (og K) vil selvsagt være den samme for alle vare- og tjenestegrupper (avhenger bare av målefeilen til inntekten), altså

$$\hat{b}_j = \hat{K}\, b_j \quad (j = 1, 2, \ldots, m)$$

Størrelsen

$$\hat{\hat{b}}_j = \frac{\hat{b}_j}{\hat{K}} \quad (j = 1, 2, \ldots, m) \quad \text{(II. 3.3.)}$$

er et asymtotisk forventningsrett og konsistent uttrykk for b_j og vil kunne nyttes som estimator dersom $\hat{K}$ er kjent. Når man setter inn estimatene ($\hat{b}$-ene), blir det et system med m ligninger mellom m + 1 ukjente ($\hat{\hat{b}}$-ene og $\hat{K}$). Systemet er altså determinert på en konstant nær. Hvis vi imidlertid lar en av "varegruppene" representere sparingen (f.eks. nr. m), vil summen av utgiftene være lik inntekten, og dermed

$$\sum_{j=1}^{m} b_j = 1$$

Ved å pålegge restriksjonen

$$\sum_{j=1}^{m} \hat{\hat{b}}_j = 1$$

blir systemet determinert og asymtotisk forventningsrette og konsistente estimatorer for b-ene og K finnes.[1] Restriksjonen innebærer at man setter

$$\hat{K} = \sum_{j=1}^{m} \hat{b}_j$$

---

1) Metoden er identisk med en spesiell form for bruk av instrumentvariable. Se vedlegg 3.

I stedet for forbruksdisponibel inntekt kan selvsagt faktisk total forbruksutgift ($\sum_{j=1}^{m} X'_j$) tenkes brukt som forklaringsvariabel. Vi holder altså sparingen utenfor og estimerer utgiftsderiverte.

Det vil ofte være rimelig å anvende modeller med flere eksogene variable, f.eks. følgende konsumfunksjon

$$X' = a + b\,R' + c_1 y_1 + c_2 y_2 + \ldots + c_r y_r + E \quad \text{(II.3.4.)}$$

hvor c-ene betegner konstanter. y-ene kan f.eks. betegne antall personer, binære variable for husholdningstype, geografisk område e.l. Vi antar her at y-ene ikke er beheftet med målefeil, og at EU = 0 for alle verdier på alle de eksogene variablene. Under disse forutsetninger viser det seg at estimatoren til b får tilsvarende (asymtotiske) egenskaper som ovenfor angitt når vi nytter observerte variable (X og R) under estimeringer. Dette går fram av momentmatrisene når man setter inn sentralmomentene for de observerte verdiene og lar antall observasjoner gå mot uendelig[1]. Av momentmatrisene går det ellers fram at heller ikke estimatorene for c-ene vil være konsistente ved målefeil i inntekten. Ved å legge på restriksjonen $\sum_{j=1}^{m} \hat{\hat{b}}_j = 1$ kan det på samme måte som foran finnes konsistente estimatorer for b-ene. Videre kan (om man ønsker det) a og c-ene etterpå estimeres direkte via ligningen

$$X - \hat{\hat{b}}\,R = a + c_1 y_1 + c_2 y_2 + \ldots + c_r y_{r4} + U.$$

4. Varians på estimatoren

Vi tar utgangspunkt i formel (II. 3.3.) og sammenhengen

$$\hat{K} = \sum_{j=1}^{m} \hat{b}_j$$

i avsnittet foran. Dette gir

$$\hat{\hat{b}}_j = \frac{\hat{b}_j}{\sum_{j=1}^{m} \hat{b}_j} \quad (j = 1, 2, \ldots, m)$$

Innsetting av $\hat{b}$-ene gir

$$\hat{\hat{b}}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R_i - \bar{R})}{\sum_{i=1}^{n} (X_i - \bar{X})(R_i - \bar{R})} \quad (j = 1, 2, \ldots, m)$$

Metoden innebærer altså at man nytter brøken

$$\frac{\text{kovar } X_j\ R}{\text{kovar } X\ R}$$

som estimator for b_j[2].

1) Se f.eks. H. T. Amundsen: Innføring i teoretisk statistikk, kap. 9.3. (1962). 2) Estimatoren er identisk med en spesiell form for bruk av instrumentvariabel. Se vedlegg 3.

Det er en selvfølge at metoden kan brukes også ved en modell uten målefeil i inntekten. Det er en nærliggende oppgave å undersøke variansegenskapene ved estimatoren. Vi vil nå først anta at det ikke er målefeil i inntekten og vil sammenlikne $\hat{b}$ med den vanlige minste kvadratersestimatoren (i det følgende symbolisert ved $\tilde{b}$). Funksjonsformen er som foran

$$X_j = a_j + b_j R' + U_j \quad (j = 1, 2, \ldots, m)$$

hvor

$$U_j = E_j + X''_j$$

siden $R'' = 0$.

Vi skal altså sammenlikne estimatorene:

$$\tilde{b}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R'_i - \bar{R}')}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2}$$

og

$$\hat{b}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R'_i - \bar{R}')}{\sum_{i=1}^{n} (X_i - \bar{X})(R'_i - \bar{R}')} \quad \text{(II. 4.2.)}$$

hvor $X = \sum_{j=1}^{m} X_j$. X-ene er stokastiske, mens R' antas ikke stokastisk. Ved å sette inn for X-ene kan estimatorene skrives

$$\tilde{b}_j = \frac{b_j \sum_{i=1}^{n} (R'_i - \bar{R}')^2 + \sum_{i=1}^{n} (R'_i - \bar{R}')(U_{ji} - \bar{U}_j)}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2}$$

og

$$\hat{b}_j = \frac{b_j \sum_{i=1}^{n} (R'_i - \bar{R}')^2 + \sum_{i=1}^{n} (R'_i - \bar{R}')(U_{ji} - \bar{U}_j)}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2 + \sum_{i=1}^{n} (R'_i - \bar{R}')(U_i - \bar{U})}$$

hvor $U = \sum_{j=1}^{m} U_j$

Under forutsetningene i avsnitt 2 i dette kapitlet om restleddsegenskapene er som kjent $\tilde{b}$ forventningsrett

$$E\,\tilde{b}_j = b_j$$

med varians

$$\operatorname{var} \tilde{b}_j = \frac{\sigma^2 U_j}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2}$$

Når det gjelder estimatoren $\hat{\hat{b}}_j$ er det mer komplisert å finne eksakte uttrykk for forventning og varians, fordi det er stokastikk både i telleren og nevneren. Det er imidlertid vist i avsnittet foran at estimatoren er asymtotisk forventningsrett og konsistent, i det grenseverdiene

$$P \lim \frac{1}{n} \sum_{i=1}^{n} (R'_i - \bar{R}') (U_i - \bar{U})$$

og

$$P \lim \frac{1}{n} \sum_{i=1}^{n} (R'_i - \bar{R}') (U_{ji} - \bar{U}_j)$$

begge går mot null ved økende antall observasjoner, altså er

$$P \lim \hat{\hat{b}}_j = b_j \quad (j = 1, 2, \ldots, m)$$

For beregning av variansen skal vi nytte en tilnærmelsesformel[1]. Dersom vi lar T betegne telleren og N betegne nevneren, kan formelen skrives[1]:

$$\text{var } \hat{\hat{b}}_j = \text{var } \frac{T}{N} \approx \left(\frac{1}{EN}\right)^2 \left[\text{var } T + b^2_j \text{ var } N - 2 b_j \text{ kovar } (T, N)\right]$$

Til sammenlikning vil variansen til $\tilde{b}_j$ bestå av det første leddet i uttrykket:

$$\text{var } \tilde{b}_j = \left(\frac{1}{EN}\right)^2 \text{var } T$$

(nevneren til $\tilde{b}_j$ er ikke-stokastisk). Det vil således avhenge av fortegnet på

$$b^2_j \text{ var } N - 2 b_j \text{ kovar } (T, N)$$

om var $\hat{\hat{b}}_j$ skal være større eller mindre enn var $\tilde{b}_j$. For å gjøre de følgende beregninger enklest mulig, lar vi forbruket være delt i to vare- og tjenestegrupper (gruppe 1 er den som analyseres, mens gruppe 2 er resten). Leddet ovenfor blir[1]

$$b_1^2 \text{ var } N - 2 b_1 \text{ kovar } (T, N) =$$

$$b_1^2 M_{R'}^2 \sigma_u^2 - 2 b_1 M_{R'}^2 (\sigma_{u_1}^2 + \sigma_{u_1 u_2}) =$$

$$b_1 M_{R'}^2 \left[b_1 \sigma_u^2 - 2 (\sigma_{u_1}^2 + \sigma_{u_1 u_2})\right],$$

hvor $\sigma^2_u = \sigma^2_{u_1} + 2 \sigma_{u_1 u_2} + \sigma^2_{u_2}$

Uttrykket består av et positivt og et negativt ledd. Det er altså ikke mulig å trekke generelle slutninger om fortegnet, og dermed heller ikke om hvilken av estimatorene $\tilde{b}_1$ og $\hat{\hat{b}}_1$ som gir minst varians.

For å kunne sammenlikne variansene på $\tilde{b}_1$ og $\hat{\hat{b}}$ må en ha anslag på varianser/kovarianser og b_1 i et materiale. Dette gjør det selvsagt vanskelig i praksis å sammenlikne variansene på $\tilde{b}_1$ og $\hat{\hat{b}}_1$.

I vedlegg 2 er det gjort et forsøk på å anslå disse størrelsene i forbruksmaterialet fra 1973, for å kunne foreta en sammenlikning. Dette nyttes i det følgende. Vi lar varegruppe 1 være henholdsvis

- matvarer (gruppe 2 er "resten")
- drikkevarer og tobakk (gruppe 2 er "resten")

osv.

---

1) Se vedlegg 1.

I tabell 1 er det beregnet standardavvik på $\hat{\tilde{b}}_1$ i forhold til på $\tilde{b}_1$ (standardavvik på $\tilde{b}_1$ er satt lik 100,0).

Tabell 1. Standardavvik på $\hat{\tilde{b}}_1$ i prosent av standardavvik på $\tilde{b}_1$. Forbruksundersøkelsen 1973. Hele befolkningen

Vare- og tjenestegruppe	$100 \cdot (\frac{\text{var } \hat{\tilde{b}}_1}{\text{var } \tilde{b}_1})^{\frac{1}{2}}$
Matvarer	101,7
Drikkevarer og tobakk	95,2
Klær og skotøy	90,7
Bolig, lys og brensel	92,6
Møbler og husholdningsartikler	90,4
Helsepleie	97,4
Reiser og transport	93,1
Fritidssysler og utdanning	90,7
Andre varer og tjenester	90,9

Tabellen viser at estimatoren $\hat{\tilde{b}}_1$ har et mindre standardavvik enn estimatoren $\tilde{b}_1$ unntatt for matvarer. (En kan som sagt ikke si noe generelt om hvordan forholdet mellom variansene er.)

Det finnes mange forskjellige måter å dele inn det private frobruket i undergrupper av varer og tjenester på. Av dette følger det også at det vil kunne avhenge av hvordan denne inndelingen gjøres, hvilken av de to estimatorene som er den beste. Det kan nevnes at prøveregninger som vi har foretatt viser at $\hat{\tilde{b}}_1$ synes å være særlig gunstig ved en inndeling av forbruket i grove vare- og tjenestegrupper. (En ekstrem grovdeling er selvsagt at man har bare én vare- og tjenestegruppe, nemlig total forbruksutgift. I dette spesialtilfellet blir selvsagt $\hat{\tilde{b}}$ alltid lik 1, med varians lik null.) Grunnen til at man i mange tilfelle reduserer variansen ved å nytte estimatoren $\hat{\tilde{b}}$ framfor $\tilde{b}$ er selvsagt at $\hat{\tilde{b}}$ impliserer at det er lagt en tilleggsrestriksjon (innebygget en betingelse) ved estimeringen, nemlig $\sum_{j=1}^{m} \hat{\tilde{b}}_j = 1$.

Ellers må man regne med at opplegg og gjennomføring av en forbruksundersøkelse generelt vil kunne influere på forholdet var $\hat{\tilde{b}}$/var $\tilde{b}$.

I dette regneeksemplet har det vært forutsatt at det ikke er målefeil på inntekten. Dersom det derimot er målefeil i inntekten ($R'' \neq 0$), vil ikke $\tilde{b}$-estimatoren være konsistent (b blir ikke identifiserbar). I vedlegg 1 er et tilnærmet uttrykk for variansen på $\hat{\tilde{b}}$-estimatoren beregnet for tilfellet ved bl.a. målefeil i inntekten. Variansen avhenger bl.a. av b og momentene for de stokastiske leddene. Vi har ikke arbeidet noe med å finne fram til en estimator for variansen på $\hat{\tilde{b}}$.

5. Målefeil i forbruksdata 1973

Ved analyser av forbruksundersøkelsen nyttes ofte inntektsdata (foruten forbruksdata). I tillegg er det gjerne rimelig å trekke inn i modellen variable som husholdningstype (evt. persontall i husholdningen), alder på hovedinntektstakeren, geografisk område m.v. Mens forbruks- og inntektsoppgavene antakelig som regel er beheftet med vesentlige målefeil, vil det være mer rimelig å tro at "enklere" variable, som f.eks. husholdningstype e.l. er registrert uten at det i større omfang forekommer målefeil. Det kan derfor være tillatelig å arbeide innenfor modellen med målefeil i forbruks- og inntektsoppgavene, men uten feil i de andre variablene. I det følgende vil dette bli gjort.

Vi skal først vurdere feil i forbruksdata. Vi regner med at det gjennomsnittlig vil være en viss underregistrering i forbruket, dvs. at ikke alle husholdningens utgifter kommer med under regnskapsføringen og intervjuingen. Det er ikke lett å si noe sikkert om hvilken karakter underestimeringen har. Det er vel imidlertid nærliggende å tro at underestimeringen vil være avhengig av selve nivået på forbruket. Hvis f.eks. en familie har relativt høye utgifter og mange utgiftsposter,

vil en vel også tro at det er lett å glemme en del av postene under regnskapsføringen. Vi vil her forutsette at den systematiske målefeilen er proporsjonal med utgiftsbeløpet. Vi får altså sammenhengen

$$X = d\,X' + X'',$$

hvor d betegner en konstant. Størrelsen på konstanten vil kunne variere fra varegruppe til varegruppe. Sammenlikninger med annen statistikk (f.eks. varehandelsstatistikk) tyder f.eks. på at underregistreringen er større for varegruppen drikkevarer enn for matvarer.

Ved forbruksundersøkelsen 1973 har vi innhentet likningsdata fra Skattedirektørens magnetbånd. På dette båndet finnes det bl.a. opplysninger om nettoinntekter ved kommune- og statsskattelikningen og direkte skatter og avgifter. Ut fra opplysninger på båndet kan følgende inntektsbegrep dannes:

nettoinntekt ved statsskattelikningen
\+ nettoinntekt ved sjømannsskatteordningen
\+ særfradrag
\- direkte skatter og avgifter
= inntekt

Det viser seg at dette inntektsbegrepet ikke gir noe godt uttrykk for den forbruksdisponible inntekten. For alle husholdninger under ett utgjør inntekten ca. 3/4 av forbruksutgiftene. Skattefrie inntekter er ikke med i dette inntektsbegrepet. Videre er ikke fradragsberettigede utgifter ved skattelikningen med (en del av disse utgiftene er med i forbruksutgiftsbegrepet). Videre er på skattebåndet inntekten satt lik null dersom vedkommende person ikke har så høy inntekt at han/hun betaler inntektsskatt. En god del personer står av denne grunn oppført med null inntekt på skattebåndet. (Husholdningsinntekten beregnes ved å summere husholdningsmedlemmenes inntektsbeløp.) I tillegg vil selvsagt skatteunndragelser kunne influere på målefeilene.

Det synes rimelig å anta at målefeilen på inntekten delvis vil variere på en mer eller mindre systematisk måte med en persons yrkesstatus. En kan f.eks. lage følgende inndeling av personer etter yrkesstatus:

1. Lønnstakere
2. Selvstendig næringsdrivende i jordbruk, skogbruk og fiske
3. Selvstendig næringsdrivende ellers
4. Ikke yrkesaktive

Det er nærliggende å tro at målefeilen gjennomsnittlig er annerledes for f.eks. lønnstakere enn for selvstendig næringsdrivende. Sammenhengen mellom sann og observert inntekt kan således være noe forskjellig for forskjellige sosialgrupper. I avsnitt 1 i dette kapitlet ble det skilt mellom tilfeldige og systematiske målefeil, for inntektens vedkommende symbolisert ved

$$R = g(R', \text{evt. andre variable}) + R''$$

hvor $g(\cdot)$ antas å være en eksakt funksjon av R'.

I neste avsnitt studeres tilfellet med både systematiske og tilfeldige målefeil i inntekten og tilfeldige målefeil i forbruket.

6. Tilfeldige målefeil i forbruksutgiftsdata, tilfeldige og systematiske målefeil i inntektsoppgavene

Vi betrakter tilfellet

$$X_j = X_j' + X_j'' \quad (j = 1, 2, \ldots, m)$$

$$R = g(R') + R'' \quad \text{(II. 6.1.)}$$

Innsetting av observerte verdier i konsumfunksjonen

$$X'_j = a_j + b_j R' + E_j$$

gir

$$X_j = a_j + b_j R + U^*_j$$

hvor $U^*_j = X''_j + E_j + b_j (R' - R) = U_j + b_j (R' - R)$

I estimatoren

$$\hat{b}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R_i - \bar{R})}{\sum_{i=1}^{n} (X_i - \bar{X})(R_i - \bar{R})}$$

setter vi inn $R = g(R') + R''$ og får

$$\hat{b}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(g(R'_i) - \overline{g(\cdot)}) + \frac{1}{n}\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R''_i - \bar{R}'')}{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(g(R'_i) - \overline{g(\cdot)}) + \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})(R''_i - \bar{R}'')}$$

Ved dessuten å sette inn bl.a.

$$X_j = X'_j + X''_j,\ X = X' + X'' \text{ og } X' = R'$$

får vi (uttrykt ved sentralmomentene M):

$$\hat{b}_j = \frac{b_j M_{R'\, g(R')} + M_{XJ'\, R''} + M_{Uj\, g(R')} + M_{Uj\, R''}}{M_{R'\, g(R')} + M_{R'\, R''} + M_{X''\, g(R')} + M_{X''\, R''}}$$

Under forutsetning av at $ER'' = 0$ for alle verdier på X_j, og R' og $E\, U_j = E_{X''} = 0$ for alle verdier på R' og U_j og R'' og X'' og R'' er ukorrelerte, vil $\hat{b}_j$ være asymtotisk forventningsrett og konsistent, i det

$$\operatorname{Lim} \hat{b}_j = \frac{b_j M_{R'\, g(R')}}{M_{R'\, g(R')}} = b_j$$

ved økende antall observasjoner. Dette gjelder uansett formen på den systematiske målefeilen på inntekten, dvs. uansett formen på g (·)-funksjonen - bare g (R') er korrelert med R'. Dette må (isolert sett) sies å være en nokså lempelig forutsetning. Det er et viktig poeng at man ikke behøver å kjenne formen på g (·)-funksjonen for å få en konsistent estimator for b_j.

7. Systematiske og tilfeldige målefeil i forbruks- og inntektsoppgaver. Modifisert modell

Sammenhengen mellom faktiske og observerte forbruks- og inntektstall antas å være

$$X_j = d_j X'_j + X''_j \quad (j = 1, 2, \ldots, m)$$

$$R = g(R') + R''$$

Innsetting av observerte i stedet for faktiske verdier i konsumfunksjonen gir

$$X_j = a_j + b_j R + U_j^* \quad (j = 1, 2, \ldots, m)$$

hvor

$$U_j^* = d_j E_j + X''_j + b_j (d_j R' - R) + d_j (a_j - 1)$$

I dette avsnittet nytter vi for enkelhets skyld symbolet

$$U_j = d_j E_j + X''_j \quad (j = 1, 2, \ldots, m)$$

Det er hittil forutsatt at variansene på de stokastiske leddene er konstante (homoscedastisitet) og at leddene er stokastisk uavhengig av hverandre. 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. Vi skal her forutsette at variansen på de stokastiske leddene er proporsjonale med den faktiske inntekten,

$$\text{var } U_j = R'^2 \tau^2_{u_j} \quad (j = 1, 2, \ldots, m)$$

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

hvor $\tau^2_{u_j}$ og $\tau^2_{R''}$ betegner konstanter. Videre forutsettes det at kovariansen mellom U-ene kan skrives:

$$\text{kovar } U_j U_k = R'^2 \tau_{u_j u_k}$$

hvor $\tau^2_{u_j u_k}$ betegner en konstant.

For en gruppe husholdninger blir den gjennomsnittlige variansen:

$$\frac{1}{n} \sum_{i=1}^{n} \text{var } U_j = \frac{1}{n} \Sigma R'^2_i \tau^2_{u_j} = (M^2_{R'} + \bar{R}'^2) \tau^2_{u_j} \quad (j = 1, 2, \ldots, m)$$

(og tilsvarende for R'') når $M^2_{R'}$ og $\bar{R}'$ betegner henholdsvis sentralmoment og gjennomsnitt for $\bar{R}$. U_j og R'' antas å være ukorrelert (j = 1, 2, ..., m).

Vi tar (den uveide) minste kvadratersestimatoren av de observerte størrelsene

$$\hat{b}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(R_i - \bar{R})}{\sum_{i=1}^{n} (R_i - \bar{R})^2}$$

Innsetting for X_j og R gir

$$\hat{b}_j = \frac{\sum_{i=1}^{n} \left[d_j b_j (R'_i - \bar{R}') + U_{ji} - \bar{U}_j\right] \left[(g(R'_i) - \overline{g(\cdot)})(R''_i - \bar{R}'')\right]}{\sum_{i=1}^{n} \left[(g(R') - \overline{g(\cdot)}) + (R''_i - \bar{R})\right]^2}$$

Uttrykt ved sentralmomentene vil $\hat{b}_j$ asymptotisk gå mot

$$\text{Lim } \hat{b}_j = \frac{d_j\, b_j\, M_{R'}\, g\,(R')}{m^2_{g\,(\cdot)} + (m^2_{R'} + \bar{r}'^2)\,\tau^2_{R''}}$$

Vi innfører symbolet

$$K = \frac{M_{R'}\, g\,(R')}{M^2_{g\,(R')} + (M^2_{R'} + \bar{r}'^2)\,\tau^2_{R''}}$$

Innsetting av dette gir:

$$\lim \hat{b}_j = d_j \cdot b_j \cdot K \quad (j = 1, 2, \ldots, m)$$

I tilfellet med systematiske målefeil i forbruket ($d_j \neq 1$) vil ikke b_j være identifiserbar. En kan heller ikke finne forholdet mellom f.eks. b_1 og b_2 i det

$$\frac{\hat{b}_1}{\hat{b}_2} = \frac{d_1\, b_1}{d_2\, b_2}$$

med mindre d_1/d_2 er kjent eller kan anslås (Man "vet" f.eks. at den systematiske målefeilen er lik for to varegrupper).

Vi vil i stedet betrakte (den gjennomsnittlige) inntektselastisiteten. Estimatoren for varegruppe j blir, ved å nytte observerte størrelser

$$\hat{E}_{jR} = \hat{b}_j \frac{\bar{R}}{\bar{X}_j} = d_j\, b_j\, K \cdot \frac{\bar{R}}{\bar{X}_j} = d_j\, b_j\, K \cdot \frac{\bar{g}(\cdot) + R''}{d_j \bar{X}'_j + \bar{X}''_j}$$

Ved økende antall observasjoner vil $\bar{R}''$ og $\bar{X}''$ gå mot null, altså

$$\lim \hat{E}_{jR} = b_j \cdot \frac{\bar{R}'}{\bar{X}_j} \cdot K^*$$

hvor $K^* = K \cdot \bar{g}\,(\cdot)$

Vi symboliserer den sanne gjennomsnittlige elastisiteten med

$$E_{jR'} = b_j \cdot \frac{\bar{R}'}{\bar{X}'_j}$$

og får

$$\lim \hat{E}_{jR} = E_{jR'} \cdot K^* \quad (j = 1, 2, \ldots, m).$$

Inntektselastisitetene er altså identifiserbare på en konstant nær. Dersom den ene av utgiftsgruppene (f.eks. m. m) representerer sparingen, kan også de absolutte størrelsene på elastisitetene finnes. Man tar regresjonen av total (observert) anvendelse (forbruk + sparing) med hensyn på (observert) disponibel inntekt, og får

$$\hat{E}_R = E_{R'} \cdot K^*$$

Her er "pr. definisjon" den sanne elastisiteten ($E_{R'}$) lik 1,0, slik at K^* vil være lik den observerte elastisiteten ($\hat{E}_R$):

$$K^* = \hat{E}_R$$

Konsistente estimatorer finnes altså ved

$$\hat{\hat{E}}_{jR'} = \hat{E}_{jR}/\hat{E}_R \quad (j = 1, 2, \ldots, m)$$

Dersom vi ikke kjenner sparingen, kan ikke det absolutte nivå på inntektselastisitetene finnes. Man kan imidlertid finne utgiftselastisitetene på analog måte (jfr. avsnitt 3 i dette kapitlet).

Det vil ofte være ønskelig å nytte en konsumfunksjon hvor det inngår flere forklaringsvariable enn inntekt. Vi betrakter her funksjonen (II. 3.3.):

$$X_j' = a_j + b_j R' + C_{1j}Y_1 + C_{2j}Y_2 + \cdots + C_{rj}Y_r + E_j$$

Vi setter inn observerte tall for inntekt og forbruksutgifter og estimerer ved minste kvadraters metode. Sentralmomentene påvirkes av målefeilene. Vi finner følgende (asymptotiske) sammenhenger mellom sentralmomentene for observerte variable (M^*) og sanne variable (M):

$$\text{Lim } M^{2*}_R = m^2_{R'} + (m^2_{R'} + \bar{r}^2)\,\tau^2_{R''}$$

$$\text{Lim } M^{2*}_{x_jR} = d_j\, m_{xj}\, g\,(\cdot) \quad (j = 1, 2, \ldots, m)$$

$$\text{Lim } M^*_{x_jY_k} = d_j\, m_{x'j}\, Y_k \quad \begin{matrix}(j = 1, 2, \ldots, m)\\ (k = 1, 2, \ldots, 4)\end{matrix}$$

Ved å betrakte sentralmomentmatrisen[1] er det lett å se at estimatorene til b-ene blir analoge med tilfellet med bare inntekten som forklaringsvariabel.

Det kan tenkes at det vil være interesse for også å estimere konstantene $c_1, c_2, \ldots, c_r$ i ligningen. Det viser seg at c-ene ikke blir identifiserbare ved de spesifiserte typer målefeil i forbruk og inntekt. Dette er ikke overraskende. Vi fant i avsnitt 2 at en modell med bare systematiske målefeil i forbruket heller ikke gir identifikasjon. Dersom det imidlertid bare er tilfeldige målefeil, kan man finne konsistente estimatorer til c-ene (se avsnitt 3). Bare dersom det på en eller annen måte er gjørlig å korrigere for de systematiske målefeilene i forbruket, er det mulig å identifisere c-ene.

Vi har i dette avsnittet betraktet en modell med tilfeldige og systematiske målefeil i inntekt og forbruksutgifter. Det er vist at de tilfeldige feilene på sett og vis lar seg mestre, forutsatt visse krav om fordelingsegenskapene. Det samme gjelder nokså generelt for systematiske målefeil i inntekten. Det kreves f.eks. ikke at den systematiske feilen skal ha noen bestemt form. Den må imidlertid ikke være korrelert med den tilfeldige målefeilen i forbruksutgiftene eller restleddet i konsumfunksjonen.

Største problemet synes å være evt. systematiske målefeil i forbruksutgiftene. Bare dersom målefeilen har en meget enkel (og kjent) form, kan man eventuelt gardere seg mot den.

I dette avsnittet er det forutsatt et spesialtilfelle hvor den systematiske målefeilen i forbruket er proporsjonal med utgiftsnivået. Dette gir åpenbart ikke mulighet for å identifisere b-ene, men derimot (de gjennomsnittlige) elastisitetene. Det kan imidlertid tenkes at man fra andre statistiske kilder kjenner den totale omsetning innenlands til privat forbruk av de forskjellige vare- og tjenestegrupper (f.eks. via varehandelsstatistikken). Budsjettandelen kan derved beregnes, og estimater for b-ene finnes ved

$$\hat{b}_j = \hat{\hat{E}}_{jR'} \cdot \hat{\alpha}_j \quad (j = 1, 2, \ldots, m)$$

1) Se f.eks. H. T. Amundsen: Innføring i teoretisk statistikk, kap. 9.3. (1962).

hvor $\hat{\hat{\alpha}}$-ene betegner beregnede budsjettandeler. (Dersom vi utfører analysen på deler av befolkningen - f.eks. lønnstakerne - er det trolig vanskelig å anslå budsjettandeler alene ut fra totaltall for hele landet. Dersom imidlertid den systematiske målefeilen i forbruket ved forbruksundersøkelsen med rimelighet kan antas å være den samme for alle befolkningsgrupper, vil budsjettandeler kunne anslås ved at forbruksundersøkelsesdata og totaltall for landet "kombineres", hvoretter b-ene beregnes).

8. Noen empiriske resultater

I dette kapitlet er det gitt et eksempel på numeriske beregninger på datamaterialet fra forbruksundersøkelsen 1973. Samme modell som i avsnittet foran er nyttet. Konsumfunksjonen er

$$X_j' = a + b\,R' + c_1 y_1 + c_2 y_2 + c_3 y_3 + E_j$$

for vare- og tjenestegruppe nr. j (j = 1, 2, ..., m), hvor y-ene er symboler for husholdningsstørrelse og -sammensetning:

Y_1 = antall personer under 16 år i husholdningen ("barn")

Y_2 = " " 16-66 år

Y_3 = " " 67 år og over ("pensjonister")

Ved å nytte nasjonalregnskapets tall for privat forbruk og dividere med antall innbyggere har vi gjort anslag på den systematiske målefeilen på forbruket i forbruksundersøkelsen (representert ved d-ene). Når den systematiske delen av målefeilen i forbruket er eliminert, kan konsistente estimater for b-ene finnes.

Tabell 2 viser estimatene $\hat{\hat{b}}_j$ (j = 1, 2, ..., m). Som nevnt i avsnitt 4 har vi ikke funnet noen god estimator for variansen til $\hat{\hat{b}}$-ene. I tabellen er størrelsesordenen på variansene indikert ved $\hat{\text{var}}\,\hat{\hat{b}}_j = (\frac{1}{\hat{b}})^2 \text{ var } \hat{b}_j$ (hvor $\hat{b} = \sum_{j=1}^{m} \hat{b}_j$).

Inntektsbegrepet som er nyttet er definert i avsnitt 5 i dette kapitlet og bygger på data fra Skattedirektøren.

Tabell 2. Utgiftsderiverte og gjennomsnittlige elastisiteter for forskjellige vare- og tjenestegrupper. Standardavvik i parentes[1])

Vare- og tjenestegruppe	Deriverte $\hat{\hat{b}}_j$	Elastisiteter $\hat{E}_{X'_j R'}$
Matvarer	0,093 (0,011)	0,38
Drikkevarer og tobakk	0,099 (0,009)	1,23
Klær og skotøy	0,078 (0,010)	0,78
Bolig, lys og brensel	0,162 (0,016)	1,22
Møbler og husholdningsartikler	0,109 (0,010)	1,17
Helsepleie	0,035 (0,010)	1,06
Reiser og transport	0,194 (0,017)	1,48
Fritidssysler og utdanning	0,067 (0,009)	0,79
Andre varer og tjenester	0,162 (0,016)	1,61
Total forbruksutgift	1,000 (0,000)	1,00

1) Variansene er beregnet ved $\hat{\text{var}}\,\hat{\hat{b}}_j = (\frac{1}{\hat{b}})^2 \text{ var } \hat{b}_j$.

Det er også utført beregninger hvor det i tillegg til de spesifiserte variablene er innført binære variable for sosial status (selvstendig næringsdrivende, lønnstaker m.v.) for hovedinntektstakeren. Dette førte imidlertid ikke til vesentlig annerledes resultater.

Gruppen matvarer har lavest elastisitet (0,38). Dette stemmer bra overens med de fleste analyser som er utført på forbruksdata. Videre ser en at gruppen reiser og transport har høy elastisitet. Dette er heller ikke overraskende; gruppen inneholder bl.a. utgifter til anskaffelse, drift og vedlikehold av privat bil. Imidlertid er det en annen gruppe, kalt "andre varer og tjenester" som har høyest elastisitet. Denne omfatter bl.a. hotellopphold, restaurantbesøk og visse feriereiser (såkalte "pakketurer").

I forbindelse med framstillingen i Kap. III er det gitt noen flere resultater fra 1973-undersøkelsen.

III. ESTIMERING VED MANGLENDE INNTEKTSDATA

1. Innledning

Vi har hittil betraktet tilfellet med målefeil i inntektsdata, og estimeringsmuligheter som da foreligger. En litt annen situasjon er at inntektsopplysninger fullstendig mangler. Dette har imidlertid ikke vært uvanlig ved våre forbruksundersøkelser og vil være en realistisk situasjon også i den nærmeste framtid. Emnet for dette kapitlet er en måte til estimering av deriverte/elastisiteter også i dette tilfelle.

I våre forbruksundersøkelser blir en rekke opplysninger innhentet (utenom forbruksoppgaver), bl.a. yrkesdeltaking for husholdningsmedlemmene (f.eks. total ukentlig arbeidstid) og yrkesstatus (selvstendig næringsdrivende, ansatt m.v.). Vi vil her drøfte mulighetene for å estimere engel-/utgiftsderiverte på "indirekte" måte ved å nytte nevnte (eller lignende) opplysninger.

2. Estimering ved genererte inntektsoppgaver

Vi ønsker å estimere parametre i en sammenheng mellom inntekt og forbruksutgifter, men mangler inntektsoppgaver.

Det er en nærliggende løsning å prøve å finne fram til de "mekanismer" som ligger bak husholdningenes inntekter, for så å nytte dette under estimeringen.

I vår undersøkelse har vi oppgaver over husholdningens arbeidstid. En må regne med at ervervsinntekter for husholdningene har sammenheng med hvor mange av husholdningsmedlemmene som er i inntektsgivende arbeid og med hvor lang arbeidstid de har. Den samlede arbeidstiden for medlemmene kan derfor være en viss indikator på ervervsinntekten.

Det er imidlertid neppe noen "streng" sammenheng mellom arbeidstid og yrkesinntekt. Det vil dessuten avhenge av bl.a. yrke og næring hvor høy inntekten vil være. Vi vil her nytte følgende grovinndeling av yrkesstatus for hovedinntektstakeren som indikator:

- ansatte
- selvstendig næringsdrivende i jordbruk, skogbruk og fiske
- selvstendig næringsdrivende ellers
- ikke yrkesaktive

For ikke yrkesaktive vil selvsagt ikke arbeidstiden indikere noe om inntekten, i dét det er helt andre forhold som bestemmer inntekten for disse gruppene (nivå på pensjoner og trygder m.v.).

Vi antar her at inntekten kan skrives som en funksjon av bl.a. arbeidstid og yrkesstatus for hovedinntektstakeren.

Inntekt = f (arbeidstid for husholdningen, yrkesstatus for hovedinntektstakeren, andre variable)
Vi nytter følgende symboler

R' = disponibel inntekt
T = samlet arbeidstid for husholdninger
Z_1 = 1 hvis hovedinntektstaker er selvstendig næringsdrivende i jordbruk, skogbruk, fiske, 0 ellers
Z_2 = 1 hvis hovedinntektstaker er annen selvstendig næringsdrivende, 0 ellers
Z_3 = 1 hvis hovedinntektstaker er ikke yrkesaktiv (ansatte er ref.-gruppe)
V = et restledd

Vi forutsetter en lineær relasjon mellom disponibel inntekt og de andre variablene:

$$R' = \alpha + \beta T + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + V$$

Restleddet V inneholder altså alle andre variable enn de spesifiserte og som "forklarer" inntekten.

Vi forutsetter samme konsumfunksjon som i kap. III. avsn. 8:

$$X'_j = a + b R' + c_1 y_1 + c_2 y_2 + c_3 y_3 + E_j \quad (j = 1, 2, \ldots, m)$$

med tilfeldige målefeil i forbruksoppgavene

$$X_j = X'_j + X''_j$$

(Det er korrigert for systematiske feil slik som i kap. III. avsn. 8)
Symbolet U_j betegner

$$U_j = E_j + X''_j \quad (j = 1, 2, \ldots, m)$$

Vi får altså følgende relasjoner:

1. $R' = \alpha + \beta T + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + V$

2. $X_j = a_j + b_j R' + c_j Y_1 + c_{j2} Y_2 + c_{j3} Y_3 + U_j \quad (j = 1, 2, \ldots, m)$

Relasjon 1 "forklarer" hvordan inntekten "blir til", relasjon 2 hvordan den anvendes. I dette systemet er (foruten parameter- og restleddsverdiene) selve "mellomleddet" - inntekten (R') - ikke observerbar, mens altså bl.a. arbeidstidens lengde og forbruksutgifter er observerbare (den sistnevnte riktignok med målefeil).

Ved estimering av parametrene i relasjon 2 er det nærliggende å nytte relasjon 1 til å "generere" inntektsdata til innsetting under estimeringen. Vi vil nytte (observert) total forbruksutgift under genereringen (etter korreksjon for systematiske feil). Vi setter inn observert total forbruksutgift ($X = \sum_{j=1}^{m} X_j$) i relasjon 1 i stedet for inntekt:

$$X = \alpha + \beta T + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + V$$

og estimerer parametrene ved minste kvadraters metode. Deretter genereres "inntektstall" ($\hat{R}'$) for de enkelte husholdningene ved estimatene (merket ^):

$$\hat{R}'_i = \hat{\alpha} + \hat{\beta} T_i + \hat{\gamma}_1 Z_{1i} + \hat{\gamma}_2 Z_{2i} + \hat{\gamma}_3 Z_{3i} + V_i \quad (i = 1, 2, \ldots, n)$$

Under estimeringen av parametrene i relasjon 2 erstattes altså R_i' med $\hat{R}'$, og estimeringen foretas som i kap. III, avsn. 8.

Slik som inntekten her er generert, vil gjennomsnittsinntekten for husholdningene alltid (definisjonsmessig) være lik total forbruksutgift:

$$\sum_{i=1}^{n} \hat{R}_i' = \sum_{i=1}^{n} X_i$$

Dette innebærer at sparingen holdes utenfor inntektsbegrepet. $\hat{R}'$ må tolkes som faktisk total forbruksutgift (målefeilene er "generert vekk"). Denne måten å generere inntektstall på medfører selvsagt at sparetilbøyeligheten ikke kan estimeres, og dermed heller ikke engelderiverte/-elastisiteter (bare utgiftsderiverte/-elastisiteter).

3. Noen empiriske resultater

I forbindelse med forbruksundersøkelsen 1973 ble det gjennomført en inntekts- og formuesundersøkelse for de samme husholdningene. Materialet fra denne undersøkelsen viser et mer fullstendig bilde av husholdningenes disponible inntekt til privat forbruk enn hva materialet fra Skattedirektøren kan gjøre. De fleste ikke skattepliktige inntektstypene er f.eks. tatt med i begrepet.

Vi har nyttet samme konsumrelasjon som i forrige avsnitt. Den utgiftsderiverte er estimert ved estimatoren $\hat{\hat{b}}$ fra kapittel III.[1)]

Følgende alternative indikatorer på inntekt er nyttet:

- inntektsbegrepet bygd på Skattedirektørens materiale
- inntekt beregnet ut fra inntekts- og formuesundersøkelsens inntektsoppgaver
- generert inntekt ($\hat{R}'$) fra avsnittet foran

Tabell 3 viser estimater på utgiftsderiverte ($\hat{\hat{b}}$) for forskjellige vare- og tjenestegrupper

Tabell 3. Estimerte utgiftsderiverte for forskjellige vare- og tjenestegrupper ved forskjellige inntektsindikatorer. Standardavvik i parentes[1)]

Vare- og tjenestegruppe	Inntektsindikatorer					
	Skattedirektørens materiale		Inntektsundersøkelsens inntektsbegrep		Antall timer inntektsgivende arbeid og yrkesstatus/næring	
Matvarer	0,093	(0,011)	0,091	(0,011)	0,092	(0,020)
Drikkevarer og tobakk	0,098	(0,009)	0,096	(0,009)	0,122	(0,016)
Klær og skotøy	0,078	(0,010)	0,087	(0,010)	0,102	(0,018)
Bolig, lys og brensel	0,162	(0,016)	0,162	(0,017)	0,116	(0,030)
Møbler og husholdningsartikler .	0,108	(0,010)	0,124	(0,010)	0,092	(0,018)
Helsepleie	0,035	(0,010)	0,027	(0,010)	0,022	(0,018)
Reiser og transport	0,195	(0,017)	0,186	(0,017)	0,193	(0,030)
Fritidssysler og utdanning	0,067	(0,009)	0,072	(0,010)	0,075	(0,017)
Andre varer og tjenester	0,162	(0,016)	0,156	(0,017)	0,187	(0,029)
Total forbruksutgift	1,000		1,000		1,000	

1) Variansene er beregnet ved $\text{var}\,\hat{\hat{b}}_j = \frac{1}{\hat{k}}\,\text{var}\,\hat{b}_j$ (jfr. tabell 14).

Tabellen viser at det er tildels vesentlige forskjeller mellom estimatene ved forskjellige inntektsindikatorer. Videre ser vi at de estimerte standardavvikene overalt er klart høyere ved bruk av antall timer inntektsgivende arbeid og yrkesstatus/næring som inntektsindikator enn ved inntektsbegrepene.

1) Utgiftsdataene er korrigert for systematiske målefeil.

Det synes ikke på noen måte å være problemfritt å finne estimater på utgiftsderiverte ved manglende inntektsdata. Det kan tenkes at det finnes andre og bedre inntektsindikatorer enn dem som er forsøkt her. Det er imidlertid selvsagt vanskelig i praksis å avgjøre om en indikator er god eller dårlig. En må antakelig i hvert enkelt tilfelle prøve å vurdere indikatorens egenskaper og om kravene til konsistens kan ventes å være oppfylt med rimelig tilnærmelse.

Det faktum at det kan være problemer forbundet ved å utføre visse statistiske beregninger bør imidlertid ikke i seg selv avskrekke noen fra å prøve. Spørsmålet kan heller være om de beregninger som kan gjennomføres ved hjelp av ett eller annet tilgjengelig materiale gir informasjon som kan antas å være av interesse eller av verdi i en gitt situasjon i praksis.

Vedlegg 1

Estimering av varianser

Estimatoren til b_1 i kap. II, avsnitt 4 er

$$\hat{b}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(R_i - \bar{R})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(R_i - \bar{R})} = \frac{T}{N}$$

hvor $X_i = \sum_{j=1}^{m} X_{ji}$ $(i = 1, 2, \ldots, n)$

Dersom vi har en vilkårlig funksjon

$$Z = g(T, N)$$

så kan variansen til Z (under visse forutsetninger) uttrykkes ved

$$\text{var } Z \approx \left[\frac{\partial g(ET, EN)}{\partial (ET)}\right]^{-2} \cdot \text{var } T + \left[\frac{\partial g(ET, EN)}{\partial (EN)}\right]^{-2} \cdot \text{var } N$$

$$+ 2\,\frac{\partial g(ET, EN)}{\partial (ET)} \cdot \frac{\partial g(ET, EN)}{\partial (EN)} \cdot \text{kovar } (T, N)^{1)}$$

Ved funksjonen $\hat{b}_1 = \frac{T}{N}$ blir dette

$$\text{var } \hat{b}_1 \approx \left[\frac{1}{EN}\right]^{-2} \text{var } T + \left[\frac{ET}{(EN)^2}\right]^{-2} \text{var } N - 2\,\frac{1}{EN}\,\frac{ET}{(EN)^2}\,\text{kovar } (T, N).$$

Vi forutsetter at det ikke er målefeil i inntekten ($R'' = 0$).

Forventningen til nevneren blir

$$EN = E\,\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(R'_i - \bar{R}')$$

Siden $X_i - \bar{X} = (R_i' - \bar{R}') + (U_i - \bar{U})$

hvor $U_i = \sum_{i=1}^{n} U_{ji}$ $(i = 1, 2, \ldots, n)$

blir

$$EN = E\,\frac{1}{n}\sum_{i=1}^{n}\left[(R_i' - \bar{R}') + (U_i - \bar{U})\right](R_i' - \bar{R}') = M^2_{R'}$$

under de spesifiserte forutsetninger.

Forventningen til telleren blir

$$ET = E\,\frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(R'_i - \bar{R}')$$

Siden $X_{1i} - \bar{X}_1 = b_1(R'_i - \bar{R}') + (U_{1i} - \bar{U}_1)$

blir forventningen

$$ET = E\,\frac{1}{n}\sum_{i=1}^{n}\left[b_1(R'_i - \bar{R}') + U_{1i} - \bar{U}_1)\right](R'_i - \bar{R}') = b_1 M^2_{R'}$$

1) Se f.eks. E. Sverdrup: Lov og tilfeldighet I, kap. V.6.

Variansen til $\hat{\hat{b}}_1$ kan derfor skrives

$$\text{var } \hat{\hat{b}}_1 \approx \frac{1}{M^4_{R'}} \left[\text{var } T + b_1^2 \text{ var } N - 2 b_1 \text{ kovar } (T, N) \right]$$

Av dette følger det at

$$\text{var } \hat{\hat{b}}_1 \approx \text{var } \tilde{b}_1 + \frac{1}{M^4_{R'}} \left[b_1^2 \text{ var } N - 2 b_1 \text{ kovar } (T, N) \right]$$

Variansen til telleren blir

$$\text{var } T = \text{var } \frac{1}{n} \sum_{i=1}^{n} \left[b_1 (R'_i - \bar{R}') + (U_{1i} - \bar{U}_1) \right] (R_i - \bar{R}') =$$

$$\text{var } \frac{1}{n} \sum_{i=1}^{n} \left[b_1 (R'_i - \bar{R}')^2 + (U_{1i} - \bar{U}_1) (R'_i - \bar{R}') \right] =$$

$$\frac{1}{n^2} \sum_{i=1}^{n} (R'_i - \bar{R}')^2 \text{ var } U_{1i} = \frac{1}{n} M_{R'}^2 \sigma^2_{u_1}$$

Variansen til nevneren blir

$$\text{var } N = \text{var } \frac{1}{n} \sum_{i=1}^{n} \left[(R'_i - \bar{R}') + (U_i - \bar{U}) \right] (R'_i - \bar{R}') =$$

$$\text{var } \frac{1}{n} \sum_{i=1}^{n} \left[(R'_i - \bar{R}')^2 + (U_i - \bar{U}) (R'_i - \bar{R}') \right] =$$

$$\frac{1}{n^2} \sum_{i=1}^{n} (R'_i - \bar{R}')^2 \text{ var } U_i = \frac{1}{n} M^2_{R'} \sigma^2_u$$

hvor $\sigma^2_u = \sigma^2_{u_1} + \sigma^2_{u_2} + 2 \sigma_{u_1 u_2}$

Kovariansen mellom telleren og nevneren kan skrives:

$$\text{Kovar} \left[\frac{1}{n} \sum_{i=1}^{n} (X_{1i} - \bar{X}_1) (R'_i - \bar{R}') \right] \left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (R'_i - \bar{R}') \right] =$$

$$\frac{1}{n^2} \sum_{i=1}^{n} (R'_i - \bar{R}')^2 \text{ kovar } (X_{1i} - \bar{X}_1) (X_i - \bar{X}) =$$

$$\frac{1}{n^2} \sum_{i=1}^{n} (R_i - \bar{R}')^2 \text{ kovar } (U_{1i} - \bar{U}_1) (U_i - \bar{U}) =$$

$$\frac{1}{n} U^2_{R'} (\sigma^2_{u_1} + \sigma_{u_1 u_2}).$$

Variansen til $\hat{\hat{b}}_1$ blir følgelig:

$$\text{var } \hat{\hat{b}}_1 \approx \frac{M^2_{R'} \sigma^2_{u_1} + b_1^2 M^2_{R'} \sigma^2_u - 2 b_1 M^2_{R'} (\sigma^2_{u_1} + \sigma_{u_1 n_2})}{n M^4_{R'}}$$

$$= \frac{\sigma^2_{u_1} + b_1^2 \sigma^2_u - 2 b_1 (\sigma^2_{u_1} + \sigma_{u_1 u_2})}{n M^2_{R'}}$$

Tilsvarende formel kan finnes ved målefeil i inntekten,

$$R = R' + R''$$

og ikke konstant varians på leddene R'' og U-ene. (var $R'' = R'^2 \tau^2_{R''}$ og var $U = R'^2 \tau^2_u$).

Variansen på $\hat{\hat{b}}_1$ blir her:

$$\text{var } \hat{\hat{b}}_1 \approx \frac{b_1 M^2_{R'} (M^2_{R'} + \bar{R}'^2) \tau^2_{R''} + M^2_{R'} (M^2_{R'} + \bar{R}'^2) \tau^2_{u_1}}{n M^4_{R'}}$$

$$+ \frac{(\frac{n-1}{n})^2 (M^2_{R'} + \bar{R}'^2) \tau^2_{u_1} \tau^2_{R''}}{n M^4_{R'}}$$

$$+ b_1^2 \left[\frac{M^2_{R'} (M^2_{R'} + \bar{R}'^2) \tau^2_{R''} + M^2_{R'} (M^2_{R'} + \bar{R}'^2) \tau^2_u + (\frac{n-1}{n})^2 (M^2_{R'} + \bar{R}'^2) \tau^2_u \tau^2_{R''}}{n M^4_{R'}} \right]$$

$$- 2b_1 \left[\frac{b_1 M^2_{R'} (M^2_{R'} + \bar{R}'^2) \tau^2_{R''} + M^2_{R'} (M^2_{R'} + \bar{R}'^2) (\tau^2_{u_1} + \tau_{u_1} + \tau_{u_1 u_2}) + (\frac{n-1}{n})^2 (M^2_{R'} + \bar{R}'^2) (\tau^2_{u_1} + \tau_{u_1 u_2}) \tau^2_{R''}}{n M^4_{R'}} \right]$$

Vedlegg 2

Anslag på restleddsvarianser m.v.

I dette vedlegget gjør vi rede for hvordan vi har gått fram for å beregne anslag på restleddsvarianser/-kovarianser for de forskjellige vare- og tjenestegruppene, samt anslag på b_1.

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.

Vi danner regresjonen

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

hvor X betegner (observert) total forbruksutgift, mens Z_1 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

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

($R^e = \hat{X}$ er beregnet slik at gjennomsnittet for husholdningene blir lik total observert forbruksutgift).

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

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

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

$$\hat{\sigma}^2_n = (21079)^2$$

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

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

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 husholdningene og binære variable 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.

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

For å finne anslag på variansene behøves det som sagt også anslag på b. Vi antar en to-deling av forbruket, slik at $b_1 + b_2 = 1$. For varegruppe 1 tas en lineær regresjon med total forbruksutgift som forklaringsvariabel:

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

hvor α_1 og β_1 betegner konstanter, X er total forbruksutgift ($X_1 + X_2 = X$) og V_1 et restledd.[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 fotskrift i betegner husholdning nr. i og n er antall observasjoner. Vi setter inn

$$X_{1i} = a_1 + b_1 R'_i + U_{1i},$$

$$X_{2i} = a_2 + b_2 R'_i + U_{2i} \text{ og}$$

$$X_{1i} + X_{2i} = X_i \quad (i = 1, 2, \ldots, n),$$

og får

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

Det er visse problemer med å finne eksakte uttrykk for forventning og varians til $\hat{\beta}_1$ 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 - $U^2_{R'}$ - konvergerer mot en konstant (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)]

$$\rho \lim \hat{\beta}_1 = \frac{b_1 (b_1 + b_2) m^2_{R'} + \sigma^2_{u_1}}{(b_1 + b_2)^2 m^2_{R'} + \sigma^2_u}$$

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

$$\rho \lim \hat{\beta}_1 = \frac{b_1 m^2_{R'} + \sigma^2_{u_1}}{m^2_{R'} + \sigma^2_u}$$

som gir

$$b_1 = \frac{(m^2_{R'} + \sigma^2_u)(\rho \lim \hat{\beta}) - \sigma^2_{u_1}}{m^2_{R'}}$$

---

1) $V_1 = (a_1 - \alpha_1) + (b_1 - \beta_1) R + (U_1 - \beta_1 (X''_1 + X''_2))$.

For å kunne anslå b_1 behøves også anslag på $m^2_{R'}$. Vi har tatt sentralmomentet for R^e som anslag på $m^2_{R'}$, altså

$$\hat{m}^2_{R'} = \frac{1}{n-1} \sum_{i=1}^{n} (R^e_i - \bar{R}^e)^2$$

Som anslag på b_1 er nyttet

$$\hat{b}_1 = \frac{(\hat{m}^2_{R'} + \hat{\sigma}^2_u)\,\hat{\beta} - \hat{\sigma}^2_{u_1}}{\hat{m}^2_{R'}}$$

Metoder ved målefeil

Det er utviklet forskjellige metoder for estimering av parametre i relasjoner når det er målefeil i data. Metodene vil selvsagt ofte variere med hvilke typer målefeil man tror gjelder. En gruppe av metodene tar særlig utgangspunkt i fordelingsegenskapene til feilleddene[1]. En viss ulempe ved slike metoder kan være at det er vanskelig å vurdere egenskapene på estimatorene dersom forutsetningene om fordelingsegenskapene er mindre bra oppfylt.

Andre metoder som har vært nyttet ved feil i variablene er datagrupperings- og instrumentvariabelmetoden. Disse vil her bli omtalt meget kort.

a. Datagrupperingsmetoden

Modellen er den samme som i kapittel II, avsnitt 3.

$$X_j = a_j + b_j R' + E_j$$

$$X_j = X'_j + X''_j$$

$$R = R' + R''$$

Ved målefeil er det blitt foreslått[2] å ordne enhetene etter stigende verdier på den uavhengige (observerbare) variabel R_1 og danne gjennomsnitt av R for enheter med verdi under en viss størrelse k_1 og over en viss størrelse k_2 - henholdsvis $\bar{R}_1$ og $\bar{R}_2$. (Det antas $k_1 \leq k_2$). En finner tilsvarende verdier for X_j, h.h.v. $\bar{X}_{j1}$ og $\bar{X}_{j2}$ og estimerer b ved

$$b^*_j = \frac{\bar{X}_{j2} - \bar{X}_{j1}}{\bar{R}_2 - \bar{R}_1} \quad (j = 1, 2, \ldots, m)$$

Vi ser at rankeringen og grupperingen av enhetene foretas med hensyn på en variabel med målefeil. Målefeilene vil kunne påvirke hvor de enkelte enhetene "havner" ved grupperingen, og estimatorene vil normalt ikke være konsistente[2]. Det spiller forøvrig en rolle hvordan k_1 og k_2 settes[3].

Det har vært foreslått å nytte mulige andre variable enn R som grupperingsvariabel[2]. Variabelen må være slik at feilleddene X'' og R'' er uavhengig av variabelens størrelse, samtidig som det må være samvariasjon mellom denne variablen og (den sanne) inntekten R'. Tankegangen bak denne framgangsmåten er analog med den som ligger bak instrumentvariabelmetoden (jfr. neste avsnitt).

b. Bruk av instrumentvariabel

Vi betrakter modellen fra avsnittet foran, og vil først illustrere metoden ved å forutsette at det ikke er målefeil i variablene.

Vi betrakter avvik fra gjennomsnittet:

$$X'_{ji} - \bar{X}'_j = b_j (R'_i - \bar{R}') + (E_{ji} - \bar{E}_j)$$

Vi tenker oss videre en annen ikke-stokastisk variabel Z hvor vi tar dens avvik fra gjennomsnittet $(Z_i - \bar{Z})$, multipliserer med denne på begge sider og summerer over alle enhetene:

$$\sum_{i=1}^{n} (X'_{ji} - \bar{X}'_j)(Z_i - \bar{Z}) = b_j \sum_{i=1}^{n} (R'_i - \bar{R}')(Z_i - \bar{Z}) + \sum_{i=1}^{n} (E_{ji} - \bar{E}_j)(Z_i - \bar{Z})$$

1) Se f.eks. J. Johnston: Econometric Methods, ch. 6-4 (1960) om "The Classical Approach".
2) Se E. Malinvand: Statistical Methods of Econometrics, ch. 10,9 (1968). 3) Se f.eks. J. Johnston: Econometric Methods, ch. 6-4 (1960).

Divisjon med $\sum_{i=1}^{n} (R'_i - \bar{R}) (Z_i - \bar{Z})$ (som antas $\neq 0$) gir

$$b_j = \frac{\sum_{i=1}^{n} (X'_{ji} - \bar{X}_j) (Z_i - \bar{Z})}{\sum_{i=1}^{n} (R'_i - \bar{R}') (Z_i - \bar{Z})} - \frac{\sum_{i=1}^{n} (E_{ji} - \bar{E}_j) (Z_i - \bar{Z})}{\sum_{i=1}^{n} (R'_i - \bar{R}') (Z_i - \bar{Z})}$$

Det forutsettes at E_j er ukorrelert med Z. Da vil estimatoren

$$\tilde{b}_j = \frac{\sum_{i=1}^{n} (X'_{ji} - \bar{X}'_j) (Z_i - \bar{Z})}{\sum_{i=1}^{n} (R'_i - \bar{R}') (Z_i - \bar{Z})}$$

være forventningsrett med varians

$$\text{var } \tilde{b}_j = \frac{\tau^2_{E_j} \sum_{i=1}^{n} (Z_i - \bar{Z})^2}{\left[\sum_{i=1}^{n} (R'_i - \bar{R}') (Z_i - \bar{Z})\right]}$$

hvor $\sigma^2 E_j$ betegner variansen til E_j.
Vi innfører symboler for sentralmomentene:

$$M^2_Z = \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z})^2$$

$$M_{R'Z} = \frac{1}{n} \sum_{i=1}^{n} (R'_i - \bar{R}') (Z_i - \bar{Z})$$

$$M^2_{R'} = \frac{1}{n} \sum_{i=1}^{n} (R'_i - \bar{R}')^2$$

og kan skrive

$$\text{var } \tilde{b}_j = \frac{M^2_Z}{n\, M^2_{R'Z}}$$

Korrelasjonen mellom R' og Z ($r_{R'Z}$) er pr. definisjon lik

$$r_{R'Z} = \frac{M_{R'Z}}{\sqrt{M^2_Z \cdot M^2_{R'}}}$$

Variansen til $\tilde{b}_j$ kan dermed skrives

$$\text{var } \tilde{b}_j = \frac{\sigma^2_{u_j}}{n \cdot r^2_{R'Z} \cdot \sqrt{M^2_{R'}}}$$

Variansen blir mindre jo mer korrelert R' og Z er med hverandre (jo større $|r_{R'Z}|$ er). Siden $r_{R'R'} = 1$, finnes det selvsagt ingen bedre instrumentvariabel enn R' selv. [Med R' som instrumentvariabel blir estimatoren $\tilde{b}$ lik minstekvadratersestimatoren,

$$\tilde{b}_j = \frac{\sum_{i=1}^{n} (X'_{ji} - \bar{X}') (R'_i - \bar{R}')}{\sum_{i=1}^{n} (R'_i - \bar{R}')^2}\Bigg]$$

Ved tilfeldige målefeil i forbruk og inntekt blir instrumentvariablen

$$\tilde{b}_j = \frac{\sum_{i=1}^{n} (X_{ji} - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^{n} (R_i - \bar{R})(Z_i - \bar{Z})}$$

som kan skrives[1]:

$$\tilde{b}_j = \frac{b_j \sum_{i=1}^{n} (R'_i - \bar{R}')(Z_i - \bar{Z}) + \sum_{i=1}^{n} (U_{ji} - \bar{U}_j)(Z_i - \bar{Z})}{\sum_{i=1}^{n} (R'_i - \bar{R}')(Z_i - \bar{Z}) + \sum_{i=1}^{n} (R''_i - \bar{R}'')(Z_i - \bar{Z})}$$

Ved å la M betegne sentralmoment kan uttrykket skrives

$$\tilde{b}_j = \frac{b_j M_{R'Z} + M_{u_j Z}}{M_{R'Z} + M_{R''Z}}$$

$\tilde{b}_j$ vil være asymtotisk konsistent dersom $M_{u_j Z}$ og $M_{R''Z}$ konvergerer mot null mens $R_{R'Z}$ ikke konvergerer mot null ved økende antall observasjoner.

En noe spesiell form for bruk av instrumentvariabelmetoden er foreslått av Liviathan (1961)[2] for forbruksdata. Han foreslo å nytte inntekten som instrumentvariabel når han skulle finne inntekts-deriverte (egentlig utgiftsderiverte). Nærmere bestemt foreslo han følgende estimator:

$$\tilde{\tilde{b}}_j = \frac{\sum_{j=1}^{n} (X_{ji} - \bar{X}_j)(R_i - \bar{R})}{\sum_{j=1}^{n} (X_i - \bar{X})(R_i - \bar{R})}$$

hvor altså X betegner total forbruksutgift. Som det går fram av kap. II, avsnitt 4 er dette nøyaktig den samme estimatoren som er symbolisert ved $\hat{\hat{b}}_j$ i kap. II, altså

$$\tilde{\tilde{b}}_j = \hat{\hat{b}}_j$$

---

1) $U_j = E_j + X''_j$. 2) N. Liviathan: Errors in Variables and Engel Curve Analysis. Econometrica, juli 1961.