ARTIKLER

6

METODER I ANALYSEN AV FORBRUKSDATA

Av Arne Amundsen

METHODS IN FAMILY BUDGET ANALYSES

OSLO 1960

STATISTISK SENTRALBYRÅ

ARTIKLER FRA STATISTISK SENTRALBYRÅ NR. 6

SÆRTRYKK AV STATSØKONOMISK TIDSSKRIFT HEFTE 1, MARS 1960

METODER I ANALYSEN AV FORBRUKSDATA

Av Arne Amundsen

METHODS IN FAMILY BUDGET ANALYSES

INNHOLD

1. De vanlige prinsipper for bearbeiding av forbruksdata 4
2. Regresjonsmetoden — valget av funksjonsform 5
3. Gruppegjennomsnitt for matvareutgiften i et forbruksmateriale for jordbrukerfamilier 7
4. Matvareutgiften bestemt ved regresjonsberegning 9
5. Sammenlikning av regresjonsmetoden og gruppegjennomsnittsmetoden 10
6. Matvareutgiftens variasjon med totalutgiften når familiestørrelsen er gitt 13
7. Matvareutgiftens variasjon med familiestørrelsen når totalutgiften er gitt 14
8. Funksjonsformen, estimeringsprinsippet og estimatene 15
9. Noen beregningsresultater for andre utgiftsposter 20
Sammendrag på engelsk — English summary 21

Oslo 1960.

Forord

I denne artikkelen, som ble lagt fram på et nordisk møte av sosialøkonomer i Marstrand i mai 1959, er hovedvekten lagt på å drøfte metodiske spørsmål. En regresjonsmetode er brukt til å analysere matvareutgiftens variasjon med totalutgiften og familiestørrelsen i et materiale hentet fra Statistisk Sentralbyrås husholdningsundersøkelse for jordbrukerfamilier i 1954.

Statistisk Sentralbyrå, Oslo, 23. februar 1960.

Petter Jakob Bjerve

METODER I ANALYSEN AV FORBRUKSDATA[1])

Av *Arne Amundsen.*

Hovedformålet med denne artikkelen er å legge fram argumenter for en omlegging av bearbeidingsmåten for husholdningsundersøkelsenes forbruksdata. Argumentene vil imidlertid også ha gyldighet for andre statistikkområder.

Det er vanlig å presentere resultatene fra en forbruksundersøkelse i form av resultat-tabeller, som for grupper av husholdninger gjengir de aritmetiske gjennomsnitt av husholdningenes kjøp av forskjellige varer. Det synes å være en alminnelig oppfatning at denne måten å stille tabellene opp på representerer en forutsetningsløs presentasjon av grunnmaterialet. Men dette er ikke tilfelle. En resultat-tabell inneholder ikke de observerte data, den gir et konsentrat av dem — i dette tilfelle de aritmetiske gjennomsnitt for forbruksutgiften i ulike forbrukergrupper. Det er ikke forutsetningsløst å ta det som gitt at en «reduction of data» (for å bruke R. A. Fisher's terminologi) skal foretas på denne måten.

Når en bestemt bearbeidingsmåte velges, impliserer dette i virkeligheten et klart valg om hvilke av materialets opplysninger som er mest relevante for å belyse en bestemt problemstilling (i dette tilfelle forbruksstrukturen hos husholdninger), og om hvilken reduksjonsmåte — hvilke funksjoner av observasjonene — som er mest effektive til i konsentrert form å ta vare på materialets opplysninger. Med andre ord, bruk av en bestemt bearbeidingsmåte impliserer økonometriske forutsetninger om en økonomisk modell og om en statistisk estimeringsmetode.

[1] Under utformingen av denne artikkelen har jeg hatt god hjelp av kritikk fra Odd Aukrust.

Det følger av de synsmåter jeg her har hevdet at enhver reduksjon av økonomiske data for analyseformål — resultat-tabellene «i en vanlig statistisk publikasjon» ikke unntatt — har karakteren av økonometriske analyser. Konsekvensen av dette må være at vi ikke kan drøfte prinsipper for databearbeiding uten å gjøre det klart hvilke modellforutsetninger vi vil bygge på og uten å vurdere estimeringsmåten ut fra kriterier for statistisk effektivitet.

De argumenter jeg skal komme med for å endre bearbeidingsmåten for forbruksdata fra husholdningsundersøkelser, gjelder så å si utelukkende estimeringsmåten. Den estimeringsmetode jeg mener bør prøves, er en *regresjonsmetode* som bygger på at etterspørselsrelasjonene har en funksjonsform som impliserer at Engelelastisitetene er lineære funksjoner av logaritmene til etterspørselsrelasjonens forklaringsvariable.[1])

De modellforutsetninger jeg bygger på, er stort sett de samme som de som ligger til grunn for vanlig bearbeiding av forbruksdata, og jeg skal ganske kort referere disse forutsetningene i det følgende avsnitt. I avsnitt 2 skal jeg si litt om de overveielser som ligger til grunn for valget av funksjonsform for etterspørselsrelasjonen. Avsnittene 3–7 inneholder et eksempel på anvendelse av funksjonsformen på forbruksutgiften til matvarer i et norsk husholdningsregnskapsmateriale og en drøftelse av resultatene. Funksjonsformen og noen av dens egenskaper er det gjort rede for i avsnitt 8, og i avsnitt 9 gis resultatene av noen beregninger for andre forbruksutgiftsposter.

1. De vanlige prinsipper for bearbeiding av forbruksdata.

Den økonomiske modell som (implisitt eller eksplisitt) danner bakgrunnen for de fleste forbruksundersøkelser, har som forklaringsvariable først og fremst familieinntekten (eventuelt familiens totalutgift) og familiestørrelsen (idet prisene kan betraktes som

[1] Statistisk Sentralbyrå bearbeider for tiden materialet fra en forbruksundersøkelse foretatt i 1958. Bearbeidingen foretas på Byråets nye elektroniske databearbeidingsmaskin. Denne er meget effektiv og rask til regresjonsberegninger, og det er bearbeidingsteknisk ingen vanskeligheter ved å ta med et stort antall forklaringsvariable. Grove kostnadsberegninger viser at en regresjonsmetode for bearbeidingen ikke vil koste mer enn vanlig bearbeiding.

faste i et tverrsnittsmateriale). Ofte trekkes det inn også andre forklaringsvariable av mer sekundær karakter, f. eks. bosted, sosial gruppe osv. Funksjonsformen blir i alminnelighet ikke spesifisert eksplisitt. Det følger imidlertid av grupperingsmåten (klasseinndeling etter inntekt og familiestørrelse) at funksjonsformen er karakterisert over det aktuelle variasjonsområde ved et antall konstante parametre, like mange parametre som der er grupper hvor det forekommer observasjoner.

Estimeringsmetoden er særdeles enkel. Parameteren for hver gruppe (tabellrute) estimeres som et aritmetisk gjennomsnitt av observasjonene i vedkommende gruppe (rute). Tallet på parametre som estimeres er som regel stort, og som en konsekvens av dette vil hvert enkelt estimat være en funksjon av et forholdsvis lite antall observasjoner. Det betyr at mange av gruppegjennomsnittene blir beregnet på grunnlag av så få observasjoner at utsagnskraften blir svært dårlig.

2. Regresjonsmetoden — valget av funksjonsform.

Ved bruk av regresjonsmetoden må ikke bare valget av forklaringsvariable, men også valget av funksjonsform skje eksplisitt. Dette valg må i noen grad avhenge av hva resultatene skal brukes til. Hvis f. eks. formålet er å belyse hvorledes etterspørselen etter forbruksgoder varierer med inntekt, priser osv. over et forholdsvis snevert område, kan det være fullt tilstrekkelig å få kjennskap til hvorledes etterspørselen varierer *i gjennomsnitt* over dette område. En enkel funksjonsform for etterspørselsrelasjonen, f. eks. en lineær funksjon eller en logaritmisk lineær funksjon, kan i slike tilfelle være en brukbar forutsetning som kan gi en tilstrekkelig god tilnærmelse til «den riktige» etterspørselsrelasjonen over det område vi er interessert i. Det vil da være liten grunn til å ofre spørsmålet om funksjonsform særlig oppmerksomhet.[1] [2])

[1] Det er grunn til å tro at i mange tilfelle spiller det en helt underordnet rolle om en velger en lineær funksjon eller en velger en logaritmisk lineær funksjon, når en skal anslå etterspørselens gjennomsnittsvariasjon på grunnlag av et observasjonsmateriale. I hvert fall peker en del forsøksberegninger, som er foretatt i Statistisk Sentralbyrå, i denne retningen. Med utgangspunkt i nasjonalregnskapsdata for perioden 1930–1939 og noen få etterkrigsår, ble etter-

Når formålet er å belyse etterspørselens variasjon over store områder for noen av de forklaringsvariable, f. eks. for inntekten, blir valget av funksjonsform langt viktigere. Imidlertid er det vanskelig å stille opp velbegrunnede økonomiske kriterier for valg av funksjonsform, og dette er vel hovedårsaken til at det er forholdsvis få arbeider som behandler dette spørsmålet. Bortsett fra de undersøkelser som bygger på Törnqvists system av etterspørselsrelasjoner, er det de helt enkle etterspørselsrelasjoner som behandles i de fleste etterspørselsanalyser som legger en eksplisitt spesifisert funksjonsform til grunn.[3])

Når jeg ikke uten videre har lagt Törnqvists formler til grunn, til tross for de oppmuntrende resultater de har gitt på forbruksdata — det gjelder f. eks. Wolds beregninger på svenske data[4]) —, har dette flere grunner. For det første har Törnqvists formler den statistisk sett ubekvemme egenskap ikke å være lineære i de ukjente parametrene. Det vanskeliggjør beregningene regneteknisk sett, og — noe som kanskje er vel så vesentlig — det avskjærer oss fra å gjøre bruk av det statistiske apparat (spredningsmål osv.) som er utviklet for det lineære tilfelle, men ikke for det ikke-lineære. For det annet er det ikke uten videre opplagt hvorledes en skal generalisere Törnqvists formler til å gjelde problemer hvor det forekommer *flere* forklaringsvariable. Og endelig kan det vel reises tvil om det er ønskelig at funksjonens asymptotiske egenskaper skal komme så sterkt i forgrunnen som tilfellet er med Törnqvists formler. Selv

spørselselastisiteter med hensyn på relativ pris og med hensyn på total konsumutgift beregnet for vel 100 konsumutgiftsgrupper både på grunnlag av en lineær funksjon og på grunnlag av en logaritmisk lineær funksjon. Gjennomgående var det så små forskjeller i resultatene fra de to beregningsmetodene at det var uten praktisk betydning. (En del av resultatene finnes i «A Preliminary Report on Regression Studies of Consumers' Expenditures in Norway», stensilert memorandum, Oslo, mai 1958.)

[2] Det begrensede gyldi hetsområde for resultatene av slike undersøkelser er sterkt presisert bl. a. av Wold, jfr. Herman Wold, Demand Analysis, Stockholm 1953, pp. 258–259, p. 275.

[3] Det bør nevnes at en del engelske forfattere har eksperimentert med forskjellige funksjonsformer for etterspørselsrelasjoner i de senere årene, se f. eks. S. J. Prais, Nonlinear Estimates of the Engel Curves, Review of Economic Studies (1952), 20, p. 87; J. A. C. Brown, The consumption of food in relation to household composition and income, Econometrica, 1954, 4, p. 444; Prais and Houthakker, The Analysis of Family Budgets, Cambridge, 1955; J. Aitchison and J. A. C. Brown, A Synthesis of Engel Curve Theory, Review of Economic Studies (1954/55), 22, p. 35.

[4] Jfr. Herman Wold, op. cit. Ch. 16. Törnqvists formler er gitt i avsnitt 9

om vi er interessert i et stort variasjonsområde, kan funksjonens egenskaper for helt ekstreme verdier av de variable være av liten interesse. Jfr. avsnitt 8.

Den funksjonsform vi har basert forsøksberegninger på, er et annengradspolynom, hvor de variable i regelen er *logaritmene* til de observerte variable.[1]) Denne funksjonsformen gir elastisiteter som er lineære funksjoner av logaritmene til de forklaringsvariable, og en kan si at det er denne egenskapen ved funksjonen som mer enn noen annen har vært bestemmende for valget av den.

3. Gruppegjennomsnitt for matvareutgiften i et forbruksmateriale for jordbrukerfamilier.

For å vise hva en regresjonsmetode for bearbeiding av et husholdningsregnskapsmateriale i praksis vil lede til, skal jeg i det følgende gjengi noen resultater fra en prøveberegning utført på grunnlag av en norsk forbruksundersøkelse for jordbrukere.[2]) Den opprinnelige undersøkelse omfattet 585 regnskaper. For å redusere regnearbeidet, ble det for vårt formål trukket et utvalg av disse (alle regnskaper med null eller fem som siste siffer), slik at vårt materiale omfatter 96 regnskaper.

De kjennetegn som ligger til grunn for bearbeidingen er:

x = matvareutgift, når egenproduserte varer verdsettes til hva familien måtte ha betalt hvis den skulle ha kjøpt tilsvarende varer. Måleenhet 1000 kroner før transformasjon til logaritmer.

y = familiestørrelsen, målt ved tallet på forbruksenheter.

z = familiens totale forbruksutgift, når egenproduserte varer verdsettes til hva familien måtte ha betalt hvis den skulle ha kjøpt tilsvarende varer. Måleenhet som for x.

$a = \frac{x}{z} \cdot 100$ = budsjettprosenten for matvarer.

[1] Som vi skal se, gjør vi unntak fra regelen om å transformere til logaritmer for en variabel som familiestørrelsen (målt ved tallet på forbruksenheter), hvor måleenheten til en viss grad har en konvensjonell karakter. I slike tilfelle har det ikke særlig god mening å drøfte *relative* endringer i den variable, og følgelig har det heller ikke god mening å transformere til logaritmer.

[2] Husholdningsregnskaper for jordbrukerfamilier 1954, Norges offisielle statistikk, XI, 274.

Tabell 1. Budsjettprosenter for matvarer etter familiestørrelse og familiens totale forbruksutgift bestemt som gruppegjennomsnitt. (Tallet på regnskaper e oppført i parrentes.)

Grupper etter familiestørr. målt ved tallet på forbruksenheter	Totalutgiftsgrupper, kr./år							
	3 160–4 460	4 460–6 310	6 310–8 910	8 910–12 590	12 590–17 780	17 780–25 120	25 120–35 480	I alt
	1	2	3	4	5	6	7	8
1,5–2,0	47,5 (1)	45,8 (2)	36,6 (3)	..	24,2 (1)	..	..	38,2 (7)
2,0–2,5	..	50,8 (1)	37,8 (5)	33,9 (3)	..	30,8 (1)	..	36,9 (10)
2,5–3,0	67,0 (1)	55,3 (3)	46,5 (4)	35,7 (10)	27,0 (1)	25,5 (3)	..	38,6 (22)
3,0–3,5	..	..	44,4 (5)	44,3 (7)	31,0 (6)	27,6 (2)	..	37,9 (20)
3,5–4,0	..	..	42,8 (2)	46,0 (4)	42,7 (2)	31,0 (3)	..	40,3 (11)
4,0–4,5	..	..	67,9 (2)	47,1 (6)	39,7 (6)	..	..	43,9 (14)
4,5–5,0	..	..	..	57,4 (1)	47,1 (1)	43,1 (1)	24,7 (2)	37,2 (5)
5,0–	..	..	..	54,3 (3)	48,4 (2)	33,9 (2)	..	45,9 (7)
I alt	56,4 (2)	51,2 (6)	42,1 (21)	42,3 (34)	36,4 (19)	30,1 (12)	24,7 (2)	39,5 (96)

Verdsettingen av egenproduserte varer til innkjøpspris er valgt av hensyn til sammenlikninger med liknende beregninger for andre forbruksundersøkelser (arbeidere, funksjonærer, fiskere, alderstrygdede).

I tabell 1 er materialet gruppert på vanlig måte etter familiestørrelse og forbruksutgift. Klasseintervallet for familiestørrelse er $^1/_2$ forbruksenhet. For totalutgiften ligger øvre klassegrense ca. 41 % over nedre klassegrense.[1]) For å dekke materialets variasjonsområder, trenges det da 8 størrelsesklasser og 7 utgiftsklasser. Det gir en tabell med 56 grupper (ruter), og dessuten grupper for kolonnegjennomsnittene og linjegjennomsnittene.

Til tross for det ganske store tall på grupper, må gruppeinndelingen karakteriseres som grov, iallfall når det gjelder utgiften. F. eks. kommer familier med totalutgift i intervallet 3160–4460 kr./år i samme gruppe, og familier med totalutgift i intervallet 25 120–35 480 kr./år kommer i samme gruppe.

[1] Denne metoden for fastsettelse av klassegrenser for inntekten ble brukt første gang på norske data i «Husholdningsregnskaper mai 1947–april 1948», Norges offisielle statistikk, XI, 23. Metoden kan sies å være motstykket til å bruke *logaritmen* til forbruksutgiften som forklaringsvariabel i en etterspørselsrelasjon.

Tabellen omfatter som nevnt bare 96 blant undersøkelsens 585 regnskaper. Med et ca. seks ganger så stort materiale ville representasjonen i mange av gruppene bli atskillig forbedret. Men de samme svakhetene ville fortsatt være til stede, bare i noe svakere grad (mange tabellruter ville være representert ved et lite antall regnskaper).

Istedenfor å gi tallene for matvareutgifter i kronebeløp, er *budsjettprosenten* for matvarer brukt som kjennetegn, idet det vel letter lesingen av tabellene å få presentert relative tall.

4. Matvareutgiften bestemt ved regresjonsberegning.

Tabell 2 gir budsjettprosenter for matvarer bestemt ved regresjon for de samme grupper som i tabell 1. Istedenfor «I alt»-kolonnen og «I alt»-linjen i tabell 1 er det i tabell 2 tatt med en bunnlinje som viser matvareprosentens variasjon for en familie hvis størrelse svarer til familienes gjennomsnittsstørrelse i hele materialet, og en kolonne til høyre som viser matvareprosentens variasjon for en totalutgift svarende til den gjennomsnittlige totalutgift i hele materialet. Bunnlinjen og høyreside-kolonnen i tabell 2 har altså

Tabell 2. Budsjettprosenter for matvarer etter familiestørrelse og familiens totale forbruksutgift bestemt ved regresjon. (I ruter hvor materialet ikke har observasjoner er tallene satt i parentes.)

Grupper etter familiestørr. målt ved tallet på forbruksenheter	Totalutgiftsgrupper kr./år							
	3 160–4 460	4 460–6 310	6 310–8 910	8 910–12 590	12 590–17 780	17 780–25 120	25 120–35 480	Gjennomsnitt
	1	2	3	4	5	6	7	8
1,5–2,0	48,4	42,1	35,8	(29,9)	24,4	(19,6)	(15,5)	29,3
2,0–2,5	(54,3)	47,1	40,1	33,4	(27,4)	22,0	(17,3)	32,9
2,5–3,0	60,4	52,4	44,6	37,2	30,4	24,4	(19,2)	36,5
3,0–3,5	(66,5)	(57,7)	49,0	40,9	33,4	26,9	(21,1)	40,2
3,5–4,0	(72,8)	(62,9)	53,6	44,6	36,5	29,2	(23,0)	43,8
4,0–4,5	(78,7)	(68,2)	57,9	48,2	39,5	(31,6)	(24,8)	47,4
4,5–5,0	(84,5)	(73,3)	(62,1)	51,8	42,3	33,9	26,6	50,8
5,0–	(92,5)	(80,0)	(67,9)	56,5	46,3	36,9	(29,0)	55,5
Gjennomsnitt ..	67,8	58,7	49,9	41,6	34,0	27,3	21,5	40,8

en tolking som prinsipielt sett er helt analog med tolkingen av de øvrige linjer og kolonner i tabellen. (Dette er ikke tilfelle i tabell 1, hvor bunnlinjen viser variasjonen med totalutgiften for en ujamt varierende familiestørrelse, og kolonnen til høyre viser variasjonen med familiestørrelsen for en ujamt varierende totalutgift. Det er klart at tolkingen av slike tall kan være vanskelig, for ikke å si vill-ledende.)

De regresjonsbestemte matvareprosentene er regnet ut for klasseintervallenes midtpunkt (som geometriske gjennomsnitt).

5. Sammenlikning av regresjonsmetoden og gruppegjennomsnittsmetoden.

Tabell 3 viser i hvilken grad det er overensstemmelse mellom gruppetallene bestemt ved de to metoder. Den gir for hver tabellrute det gruppegjennomsnittsbestemte utgiftsbeløp i prosent av det regresjonsbestemte utgiftsbeløp.[1])

Tabell 3. Gruppegjennomsnittsbestemte utgiftsbeløp for matvarer i prosent av regresjonsbestemte utgiftsbeløp. (Tallet på regnskaper er oppført i parentes.)

Grupper etter familiestørr. målt ved tallet på forbruksenheter	Totalutgiftsgrupper kr./år							
	3 160–4 460 1	4 460–6 310 2	6 310–8 910 3	8 910–12 590 4	12 590–17 780 5	17 780–25 120 6	25 120–35 480 7	I alt/gj.snitt 8
1,5–2,0	92,7 (1)	93,8 (2)	97,9 (3)	..	84,3 (1)	..	..	73,8 (7)
2,0–2,5	..	112,5 (1)	101,9 (5)	104,9 (3)	..	126,8 (1)	..	95,5 (10)
2,5–3,0	95,9 (1)	114,0 (3)	99,6 (4)	100,5 (10)	76,4 (1)	92,3 (3)	..	92,5 (22)
3,0–3,5	..	..	97,5 (5)	105,4 (7)	85,5 (6)	97,3 (2)	..	97,7 (20)
3,5–4,0	..	..	88,1 (2)	101,1 (4)	101,9 (2)	99,3 (3)	..	107,2 (11)
4,0–4,5	..	..	93,5 (2)	101,2 (6)	100,6 (6)	..	..	102,1 (14)
4,5–5,0	..	..	..	113,5 (1)	119,4 (1)	116,2 (1)	91,8 (2)	131,5 (5)
5,0–	..	..	..	103,0 (3)	102,8 (2)	82,2 (2)	..	106,9 (7)
I alt/gj.snitt .	75,1 (2)	86,7 (6)	88,1 (21)	104,9 (34)	101,2 (19)	100,9 (12)	113,8 (2)	96,8 (96)

[1] Dette svarer ikke helt til det vi får om vi regner ut rutetallene i tabell 1 i prosent av tallene i korresponderende ruter i tabell 2. Uoverensstemmelsene skyldes at budsjettprosentene i tabell 1 er regnet i forhold til *observert* totalutgift i hver rute, mens budsjettprosentene i tabell 2 er regnet i forhold til *midtpunktet* for rutens totalutgiftsintervall.

Vi ser at det stort sett er god overensstemmelse mellom materialets gruppegjennomsnitt og de regresjonsbestemte tall i ruter hvor tallet på regnskaper er forholdsvis stort, mens overensstemmelsen er sterkt varierende der hvor det er ganske få regnskaper. Dette er nærmere belyst ved følgende oppstilling:

Når tallet på regnskaper i gruppen er:	har materialets gruppegjennomsnitt følgende verdier (regnet i prosent av de regresjonsbestemte utgiftsbeløp)
10	100,5
7	105,4
6	85,5, 100,6, 101,2
5	97,5, 101,9
4	99,6, 101,1
3	92,3, 97,9, 99,3, 103,0, 104,9, 114,0
2 og 1	fra 76,4 til 126,8

Ser vi nærmere på tallene i tabell 1 og tabell 2 i de rutene hvor uoverensstemmelsen er av noen betydning, og prøver ut fra en generell vurdering å bedømme hvilke resultater som er plausible, synes det å være grunn til å la avgjørelsen gå i favør av de regresjonsbestemte tallene i tabell 2. I og for seg er det selvsagt ikke noe sterkt argument at resultatene «er plausible». Men andre argumenter har større tyngde.

Svakhetene ved den metoden som ligger til grunn for oppstillingen av tabell 1, sett ut fra et statistisk estimeringssynspunkt, kan illustreres ved et resonnement som kanskje ikke er helt stringent i alle detaljer, men som til gjengjeld kan ha en intuitiv appell. Metoden innebærer, kan en si, at materialet splittes opp i et større antall del-materialer, som så behandles som atskilte undersøkelser. Tallet på del-undersøkelser er lik tallet på ruter i tabellen hvor det forekommer tall, dvs. i det foreliggende tilfelle lik 32, jfr. tabell 1. For hver del-undersøkelse regnes ut et resultat (som i det foreliggende tilfelle er den gjennomsnittlige matvareprosent for regnskapene i vedkommende del-materiale). *Dette resultat regnes ut uten hensyn til hvilke resultater som kommer fram i tabellens øvrige ruter.* Ved å gå fram på denne måten unnlater vi å ta hensyn til en særdeles viktig del av de opplysninger materialet inneholder, nemlig at *variasjonen* fra rute til rute innenfor en linje og innenfor en

kolonne i tabellen har en bestemt tendens. Tar vi for oss f. eks. første linje i tabell 1, finner vi at matvareprosenten synker fra rute til rute, fra 47,5 i den laveste inntektsgruppen til 24,2 i den høyeste inntektsgruppen som er representert. Etterat vi har konstatert det, kan vi stille spørsmålet: Hvor meget synker matvareprosenten *gjennomsnittlig* over et bestemt antall ruter på linjen, f. eks. over de tre første rutene? Dette kan vi finne ut ved hjelp av våre data. Og når vi har regnet ut resultatet, har vi fått en opplysning som øker vår viten om rutegjennomsnittene i de tre rutene, nemlig om hvor stor *avstand* det gjennomsnittlig er mellom dem. Denne opplysningen kan vi bruke til å *korrigere* de rutegjennomsnittene vi opprinnelig hadde regnet ut. Slik kunne vi — i prinsippet — tenke oss å fortsette, feltvis linje for linje eller kolonne for kolonne, tvers igjennom hele tabellen.

Av flere grunner er den framgangsmåten som her er skissert, ikke særlig anvendbar for praktiske formål, bl. a. fordi den ikke gir noen klar regel for hvorledes vi skal regne ut den gjennomsnittlige avstand mellom rutegjennomsnittene eller for hvorledes vi skal foreta korreksjonene. Men *framgangsmåten illustrerer et prinsipielt synspunkt, hvis konsekvens er at en regresjonsmetode skulle gi muligheter for bedre estimater.* De rutetallene vi får når vi bruker en regresjonsmetode er også gjennomsnittstall. De er ikke vanlige aritmetiske gjennomsnitt, men det er ikke noe ved problemstillingens karakter som tilsier at de aritmetiske gjennomsnitt er «de riktigste». Det vi ønsker å få fram i tabellene er «den typiske variasjon i materialet», og for dette formål vil regresjonsbestemte tall kunne være «bedre» gjennomsnitt enn de vanlige aritmetiske gjennomsnitt.

Argumenteringen ovenfor innebærer klart nok at regresjonsberegninger kan ha en viss plass på databearbeidingens område. Når vi oppfatter regresjonen som en annen måte å regne ut gjennomsnittstall på — og det kan vi altså gjøre —, så kan metoden ikke oppfattes som noe markert utslag av radikalisme. Det er heller ikke holdbart å si at vanlige aritmetiske gruppegjennomsnitt presentert i en tabell som tar sikte på å belyse samvariasjon, er forutsetningsløse, mens regresjonsestimering bygger på at visse forutsetninger er oppfylt. I begge tilfelle bygger vi på a-prioriske forutsetninger. Det finnes selvsagt ingen måte å gi et konsentrat av et

materiale på, som er forutsetningsløs. De vanlige aritmetiske gjennomsnitt er intet unntak i så måte. Det er nok å vise til at for utpreget *skjeve* fordelinger kan kjennskapet til det aritmetiske gjennomsnitt *alene* gi oss en villedende opplysning om fordelingens karakter, fordi vi — når intet annet er opplyst — stilltiende forutsetter at vi har med en symmetrisk fordeling å gjøre.

6. *Matvareutgiftens variasjon med totalutgiften når familiestørrelsen er gitt.*

Matvareprosentens variasjon med totalutgiften, når familiestørrelsen er konstant, kommer til uttrykk i tallene på hver enkelt linje i tabell 2. Regelmessigheten i tallenes forløp kommer klarere til uttrykk om vi regner ut Engelelastisiteten for hver rute i tabellen, altså for den totalutgift og for den familiestørrelse som hver av rutene representerer. Det viser seg at i det foreliggende tilfelle har Engelelastisiteten praktisk talt den samme variasjon med totalutgiften på hver eneste linje, eller med andre ord at den er på det nærmeste *uavhengig av familiestørrelsen*. Dette er et resultat som intuitivt synes å være plausibelt, men ikke på noen måte selvsagt. Det vil være av betydelig interesse å få undersøkt om tilsvarende beregninger på andre forbruksdata og for andre utgiftsgrupper leder til den samme konklusjon.

Engelelastisitetens variasjonsområde går fra 0,61 i den laveste totalutgiftsgruppen til 0,28 i den høyeste. Variasjonen for øvrig går fram av følgende oppstilling:[1])

Gruppens kol.nr. i tabellen	1	2	3	4	5	6	7	8
Gruppens gj.sn. totalutg. (klassemidtpunkt), kroner, avrundet	3 800	5 300	7 500	10 600	15 000	21 100	29 900	10 900
Engelelastisitet	0,61	0,56	0,50	0,45	0,39	0,33	0,28	0,44

Det er av interesse å sammenlikne Engelelastisitetene i oppstillingen ovenfor med resultater fra andre undersøkelser. For

[1]) Forskjellen mellom Engelelastisitetene innenfor en kolonne i tabellen (dvs. for varierende familiestørrelse) er ikke i noe tilfelle større enn en enhet i den siste oppgitte desimal.

arbeiderfamilier i 1951/1952, med gjennomsnittsinntekt på ca. kr. 12 000, som i 1954-priser svarer til noe over kr. 14 000, er Engelelastisiteten for matvarer beregnet til 0,35. For høyere funksjonærfamilier, 1952/1953, med gjennomsnittsinntekt på ca. kr. 21 000 som i 1954-priser svarer til kr. 24–25 000, er den samme elastisistet beregnet til 0,27. Disse resultatene må sies å føye seg pent inn i oppstillingen ovenfor. Det samme gjør resultatene av Törnqvists og Wolds analyser av finske og svenske data.[1])

7. *Matvareutgiftens variasjon med familiestørrelsen når totalutgiften er gitt.*

Denne variasjonen kommer til uttrykk i tallene i en kolonne i tabell 2. Som mål for denne variasjonen kan vi bruke den prosentvise endring i matvareutgiften ved en øking i familiestørrelsen med en forbruksenhet. Analogt med tilfellet med Engelelastisiteten finner vi at denne prosentendringen har praktisk talt samme variasjon i hver eneste kolonne, eller med andre ord at den nesten er *uavhengig av totalutgiftsnivået.* Dette er for så vidt en refleks av uavhengigheten for Engelelastisitetens vedkommende, idet den brukte formel ikke åpner mulighet for samtidig å ha uavhengighet i det ene tilfelle og avhengighet i det andre tilfelle.

Variasjonsområdet for den prosentvise matvareutgiftsøking pr. enhet øking i familiestørrelsen, når totalutgiften er konstant, går fram av følgende oppstilling:

Grupper etter familiestørrelse målt med tallet på forbruksenheter:	Prosentvis matvareutgiftsøking pr. enhet øking i familiestørr. når totalutgiften er konstant:
1,5–2,0	26,5
2,0–2,5	24,4
2,5–3,0	22,2
3,0–3,5	20,1
3,5–4,0	18,0

[1] Törnqvists beregninger gir en Engelelastisitet for matvarer (arbeidere) på 0,63 for en lav inntekt, og på 0,36 for en tre ganger så høy inntekt. (Ekonomisk Tidskrift, Nr. 2, 1941, p. 223). En finner stort sett det samme variasjonsområde for Engelelastisiteten for matvarer i Wolds beregninger, jfr. Herman Wold, op. cit., p. 272.

4,0–4,5	16,0
4,5–5,0	13,9
5,0–	11,0
3,3 = gjennomsnitt	19,7

For diskusjonene om «hva det koster» å ha stor familie, er det kanskje lettere å studere en tabell som har utgifts*endringer* i absolutte tall. La oss til illustrasjon grovt regne et barn ekvivalent med $^1/_2$ forbruksenhet. Da gir en beregning på grunnlag av regresjonsformelen følgende tall for økingen i matvareutgiftene i de forskjellige totalutgiftsgruppene, når familien øker med ett barn, uten at det skjer noen øking i totalutgiften (inntekten):

Gruppens kol.nr. i tabellen	1	2	3	4	5	6	7	8
Gruppens gj.sn. totalutgift (klassemidtpunkt), kroner	3 800	5 300	7 500	10 600	15 000	21 100	29 900	10 900
Øking i matvareutgift ved øking i fam.størr. på $^1/_2$ forbr.enhet ved konstant totalutgift, kroner	225	280	335	395	455	510	565	390

Tallene varierer litt med familiestørrelsen, f. eks. fra 220 til 235 kroner i kolonne 1 og fra 530 til 580 kroner i kolonne 7, men disse variasjonene har liten utsagnskraft. Variasjonen med totalutgiftsnivået er langt sikrere bestemt.

Det vil være av interesse å kjenne til hvilke poster som *reduseres* som følge av en øking i matvareutgiften under konstant totalutgift, og hvorledes reduksjonene varierer med totalutgiftsnivået. Dette vil kunne gi verdifulle opplysninger til å vurdere forskjellene mellom små og store familiers økonomi. Beregninger med sikte på dette er planlagt.

8. *Funksjonsformen, estimeringsprinsippet og estimatene.*

De variable som inngår i modellen er matvareutgiften x, familiestørrelsen y og totalutgiften z, idet vi bruker de betegnelsene som er innført i avsnitt 3. Den forbruksrelasjonen som er lagt til grunn for beregningene, har følgende form:

$$(1) \quad \log x = a_0 + a_1 y + a_2 \log z + a_3 y^2 + a_4 (\log z)^2 + a_5 y \cdot \log z$$

hvor a-ene er ukjente konstanter, i alt seks, som skal estimeres på grunnlag av observasjonsmaterialet, og hvor «log» betegner «logaritmen til». Av grunner som allerede er nevnt, transformeres ikke den variable y til logaritmer, jfr. fotnote 1), side 7.

Ved partiell derivasjon av (1) med hensyn på logaritmen til z får vi Engelelastisiteten. Den er

$$E = a_2 + 2a_4 \log z + a_5 y \tag{2}$$

Ved partiell derivasjon av (1) med hensyn på y får vi den relative tilvekst i matvareutgiften (når totalutgiften er konstant) ved en absolutt tilvekst i familiestørrelsen. Den er

$$F = a_1 + 2a_3 y + a_5 \log z \tag{3}$$

(Derivasjonene forutsetter at vi bruker de naturlige logaritmer. Hvis Brigg-ske logaritmer brukes i beregningene, må det etterpå foretas en transformasjon til naturlige logaritmer for å få en enkel overgang til prosentvise tilvekster.)

Formel (1) gir i det spesialtilfelle da y er konstant og a_4 og a_5 er lik null, en logaritmisk lineær funksjon i to variable. Ut fra det synspunkt at denne funksjon er en tilnærmelse til «den riktige» funksjonsformen, hvor bare første ledd i en rekkeutvikling etter Taylor's formel er tatt med, kan formel (1) tolkes som en tilnærmelse som tar med de to første ledd i en slik utvikling. Av formel (2) ser vi at en gitt prosentvis endring i totalutgiften, under konstant familiestørrelse, gir samme absolutte tilvekst i Engelelastisiteten uansett hvilket totalutgiftsnivå vi velger som utgangspunkt. Dette er en egenskap som intuitivt er lett å tolke, og som derfor skulle være egnet som grunnlag for en vurdering av plausibiliteten av funksjonsformen. Umiddelbart kan denne egenskapen virke som en stram forutsetning, men klart nok mindre stram enn en forutsetning om konstant Engelelastisitet over hele variasjonsområdet for totalutgiften. Formel (3) har stort sett en analog tolking.

Det kan være av interesse å foreta en sammenlikning med Törnqvists funksjonsform for et nødvendighetsgode, når inntekten er den eneste forklaringsvariable. Den har følgende form:

$$x = a \frac{z}{z+b} \tag{4}$$

hvor a og b er ukjente konstanter og x og z har samme tolking som tidligere. Engelelastisiteten er gitt ved

$$E = \frac{b}{z+b} \tag{5}$$

Ved derivasjon av dette uttrykket med hensyn på (den naturlige) logaritmen til z får vi

$$\frac{\partial E}{\partial \log z} = -\frac{b\,z}{(z+b)^2} \qquad \text{(i formel 5)} \tag{6}$$

mens vi ved samme derivasjon i formel (2) får en *konstant*, nemlig

$$\frac{\partial E}{\partial \log z} = 2\,a_4 \qquad \text{(i formel 2)} \tag{7}$$

Til tross for den høyst forskjellige karakter av disse to uttrykkene, viser det seg at (6) varierer så lite over det aktuelle variasjonsområde for totalutgiften at uttrykket med god tilnærmelse kan settes lik en konstant. Settes b i (5) lik 8,6, som svarer til at Engelelastisiteten er lik 0,44, når de forklaringsvariable antar de respektive observerte gjennomsnittsverdier, vil høyresiden i (6) stige litt til å begynne med, fra ca. 0,5 for det laveste totalutgiftsnivå til vel 0,6 for et noe høyere nivå, for så å synke til ca. 0,4 for det høyeste totalutgiftsnivå. Avvikene fra den konstante verdi av høyresiden i (7), 0,44, er ikke så store at uoverensstemmelsen kan ha særlig praktisk betydning. Jfr. også figur A, hvor Engelelastisitetens variasjon med logaritmen til z er tegnet inn. I dette tilfelle, når inntekten er den eneste forklaringsvariable som viser variasjon, ser det altså ut til at det spiller liten rolle hvilken av de to funksjonsformene vi velger.

Estimeringsmåten som er brukt for parametrene i (1) er minste kvadraters metode. Det skulle ikke være særlig betenkelig i dette tilfelle å forutsette at logaritmen til x har stokastiske fordelings-

egenskaper som tilsier at minste kvadraters metode gir gode estimater for parametrene.

Når den estimerte verdi av parametrene innsettes i (1), får vi følgende estimeringsfunksjon for logaritmen til matvareutgiften:

$$\begin{aligned}(8)\qquad \log x = {} & \underset{\pm\,0{,}007}{0{,}645} + \underset{\pm\,0{,}060}{0{,}131}\,(y - 3{,}337) + \underset{\pm\,0{,}258}{0{,}833}\,(\log z - 1{,}039) \\ & - \underset{\pm\,0{,}006}{0{,}008}\,(y^2 - 11{,}136) - \underset{\pm\,0{,}159}{0{,}185}\,(\log^2 z - 1{,}080) \\ & - \underset{\pm\,0{,}053}{0{,}002}\,(y \cdot \log z - 3{,}467)\end{aligned}$$

Parentesuttrykkene i formelens høyreside uttrykker de forklaringsvariable som avvik fra den observerte gjennomsnittsverdi, og 0,645 er den observerte gjennomsnittsverdi av logaritmen til x (Brigg-ske). [1]) Standardavvik for regresjonskoeffisientene er oppført, med fortegn $\pm$, under formelen. Den residuale spredning (for avvikene fra regresjonslinjen) er 0,062, som har antilogaritmen 1,15. Det betyr grovt regnet at observasjonenes gjennomsnittlige avvik fra regresjonsflaten er ca. 15 %. Da den marginale spredning for $\log x$ er 0,140, som har antilogaritmen 1,38, kan vi i en viss forstand si at regresjonen «forklarer» differensen mellom 38%s spredning og 15%s spredning for matvareutgiften. Det ligger imidlertid ikke mer i dette utsagn enn den opplysning at den multiple korrelasjonskoeffisient er 0,91.

Estimeringsfunksjonen for Engelelastisiteten er

$$(9)\qquad E = 0{,}440 - 0{,}002\,(y - 3{,}337) - 0{,}371\,(\log z - 1{,}039)$$

og for familiestørrelsekoeffisienten er den

$$(10)\qquad F = 0{,}079 - 0{,}015\,(y - 3{,}337) - 0{,}002\,(\log z - 1{,}039)$$

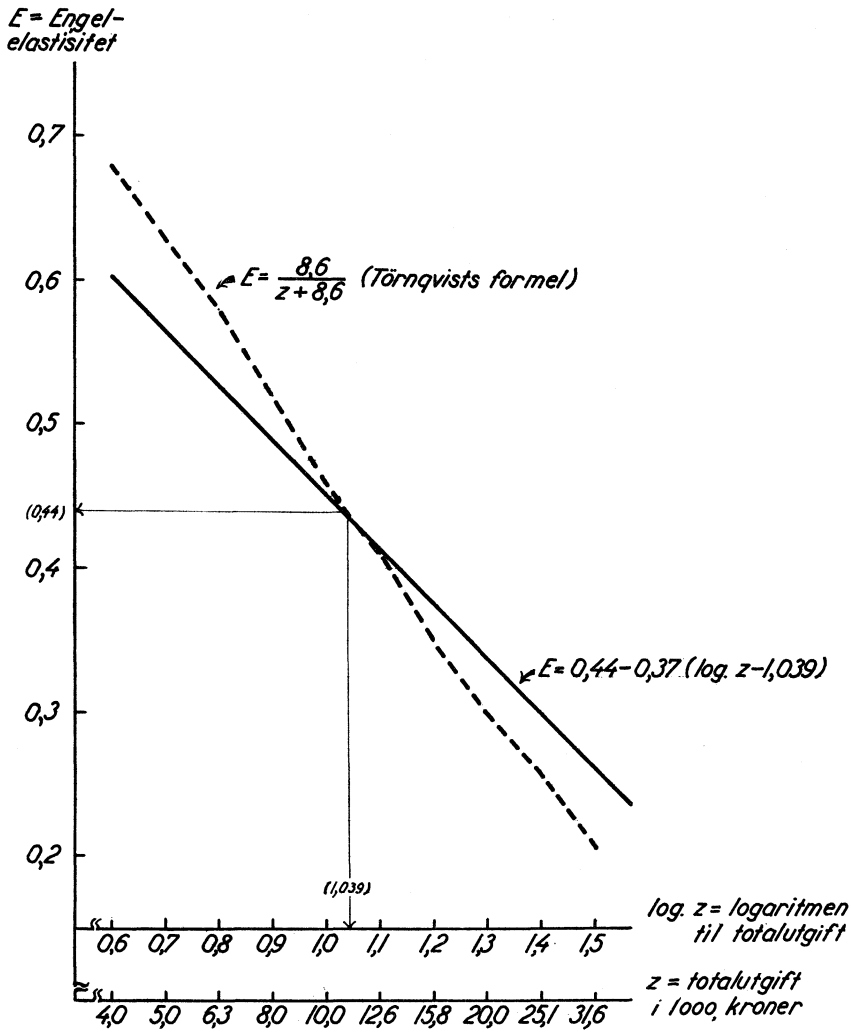
når de forklaringsvariable måles som avvik fra gjennomsnittsverdien, og når vi bruker Brigg-ske logaritmer. (Ved transformasjon til naturlige logaritmer kommer det inn en proporsjonalitetsfaktor, som er 2,3026.)

[1]) Merk at $\log^2 z$ er en forenklet skrivemåte for $(\log z)^2$. Gjennomsnittet av kvadratene av en observert variabel er i alminnelighet ikke lik kvadratet av gjennomsnittet av den variable. I (8) er de to annengradsvariable målt fra kvadratet av gjennomsnittet av den variable.

For *logaritmen* til budsjettprosenten for matvarer, a, utledes lett en estimeringsfunksjon fra (8), idet vi har

$$\log a = \log 100 + \log x - \log z \tag{11}$$

jfr. definisjonen av a i avsnitt 3.



Figur A. Engelelastisitetens variasjon med totalutgiften. Matvareutgift.

9. Noen beregningsresultater for andre utgiftsposter.

Beregninger er foretatt for fire utgiftsposter til, på grunnlag av det samme forbruksmaterialet. Det gjelder postene margarin, egg, poteter og posten kjøtt og flesk. A priori var det grunn til å tro at disse poster hadde svært stor spredning i et materiale av jordbrukerhusholdninger, og de ble valgt nettopp av denne grunn.

Tabell 4 gir estimeringsfunksjonene for de fire varepostene, og tabell 5 gir Engelelastisitetene. Størrelsen av koeffisientene, og spredningsmålene for dem, viser at forbruket av disse varer ikke er godt forklart ved den modell som er brukt.

Tabell 4. Estimeringsfunksjoner for fire matvareposter, når funksjonen skrives på formen (8).

Varepost	Konstantledd	Førstegradsledd		Annengradsledd		
	a_0	a_1	a_2	a_3	a_4	a_5
Margarin	(0,234−1)	+0,205	−0,102	+0,013	+0,262	−0,146
	±0,015	±0,145	±0,630	±0,015	±0,387	±0,129
Egg	(0,267−1)	−0,337	−0,288	−0,002	+0,101	+0,328
	±0,030	±0,287	±1,244	±0,031	±0,764	±0,255
Poteter	(0,451−1)	+0,085	+0,379	+0,003	—0,058	+0,009
	±0,017	±0,159	±0,691	±0,017	±0,424	±0,142
Kjøtt, flesk	0,073	+0,136	+1,316	−0,019	−0,488	+0,076
	±0,014	±0,132	±0,575	±0,014	±0,354	±0,118

Tabell 5. Engelelastisiteter for fire matvareposter (skrevet på formen 9).

Vareposter	Konstantledd lokalisert til gj.sn.punktet	Koeff. for (y − 3,337) (fam.størr.)	Koeff. for ($\log z$ − 1,039) (totalutgift)
Margarin	−0,045	−0,146	+0,524
Egg	+1,016	+0,328	+0,202
Poteter	+0,287	+0,009	−0,116
Kjøtt, flesk	+0,556	+0,076	−0,976

English summary

The paper discusses possibilities offered by regression methods in the processing of mass data from family budget surveys. A study of the expenditure on food in Norwegian farmer households (1954) exemplifies the discussion.

The sources of data for most regression studies of family expenditures are the summary tabulations prepared and published by the agencies that conduct the surveys. Analysts outside such agencies have no choice but to use the data in the published form. For a data-collecting agency, however, there is the possibility to employ regression methods directly on data for individual families.

High processing costs and the view that a data-collecting agency should keep clear of presumptuous *a priori* restrictions on the form of the underlying expenditure structure are probably the main arguments held against the use of regressions in large-scale data-processing. The cost argument has lost most of its weight in recent years as a consequence of the availability of modern computers that can handle the computations involved quite efficiently, at high speed. Neither is the disinclination to impose *a priori* restrictions too well founded. It follows from general principles of estimation that «it pays» to impose (valid) *a priori* restrictions. In the author's opinion, the traditional tabulation methods do not take into account well known features of expenditure functions. Regression methods are more flexible in this respect; it is possible to specify forms of regression functions that do take into account *a priori* restrictions on functional forms without being unduly restrictive. There is a wide range of possibilities. Intuitively this is made plausible by the fact that ordinary tabulation methods can be considered as special forms of regression functions.

The conclusion to be drawn is that «ordinary processing of

family budget data» — or, indeed, of any type of economic data — requires model assumptions to be specified, and principles of estimation to be considered, in order to justify the choice of a particular processing method.

The model assumptions underlying tabulations of family budget data are implicit in the choice of classification variables and of range of class intervals. A typical case, based on a sample of 96 Norwegian farmer household budgets, is shown in table 1. The classification is by family size (vertically, 8 size intervals) and by total family expenditure (horizontally, 7 expenditure intervals). It is seen that as many as 56 parameters, one for each cell, are needed to characterise the food expenditure function within the range of variation covered by the sample. The method of estimation implies that each cell parameter is a function (arithmetic average food expenditure) only of observations in that particular cell; no estimates are obtained in empty cells. (In table 1 — and table 2 — food expenditures are expressed as percentages of the average total budgets in the respective cells, to get rid of cumbersome value dimensions.)

The weaknesses of tabulation methods are well brought out by table 1. Only seven estimates are based on more than four observations and as many as seventeen on only one or two observations (the number of observations in each cell is given in parentheses). A sample many times as large would be needed in order to obtain moderately reliable estimates. Even with a sample ten times as large one would probably have a large number of estimates based on very few observations. To reduce the number of parameters by making class intervals wider is not recommendable, since the intervals are already rather wide.

Table 2 illustrates the use of a regression method for the purpose of estimating the 56 cell parameters defined by the classification in table 1. The table has been prepared on the basis of the regression function given in formula (1), with numerical coefficients as in formula (8). The notation is as follows:

x: family expenditure on food, Kroner per annum,
y: family size, measured by the number of equivalent adults,
z: total family expenditure, Kroner per annum.

The function log x, quadratic in y and log z, has 6 parameters.

The sample is used to estimate the parameters. Insertion of appropriate values (interval midvalues) for the right-hand-side variables in formula (8) gives estimates of the 56 cell parameters.

A comparison of tables 1 and 2 reveals that estimates obtained by the two methods differ on the whole only slightly in cells containing more than three households. Differences are mainly found where one must expect that the estimates in table 1 are highly influenced by sampling errors.

It his hard to find evidence that the model restrictions underlying the estimates in table 2 have caused any harm. The regularity of variation of food expenditure from cell to cell in table 2 is, of course, to some extent a consequence of the properties of the regression function employed. However, the introduction of this regularity amounts to little more than taking explicitly into account what everybody takes for granted. Six parameters seem sufficient in the present case to characterise the structure. To neglect this, and base estimation on the model characterised by 56 independent parameters, is a vaste of information.

The form chosen for the regression function is not considered as having particular advantages in the present case; the important criterion seems to be the number of independent parameters required to specify the form. In cases like the present, characterised by a fairly large sample covering the whole range of variation that is of interest to the analyst (no need for «extrapolation»), it is likely that many other functional forms characterised by equally many independent parameters would lead to almost identical results. Estimation criteria, however, favour forms linear in the unknown parameters — as is formula (1). The Engel elasticity — formula (2) — is linear in y and log z, and so is the «size coefficient» — formula (3). A comparison with the Törnqvist function (4), which is non-linear in unknown parameters, and has the Engel function (5), reveals that — for a given value of the size variable and with appropriate coefficients inserted — the Engel functions (2) and (5) are rather similar in shape (Figure A).

The estimate of the Engel function is given in formula (9). The elasticity of the expenditure on food (with respect to total

family expenditure) varies from 0.61 in the lowest expenditure group to 0.28 in the highest (the size variable held fixed, its influence is, however, negligible). These results compare well with those obtained by Törnqvist and by Herman Wold. Formula (10) gives an analogous estimate of the «size coefficient».

Formula (1) has also been employed on data for margarine, eggs, potatoes and meat, pork. These items have in common that they are likely to be subject to large errors of measurement in farmer family household budgets, and they were chosen for that reason. Results are given in tables 4 and 5.

Til salgs hos alle bokhandlere.
Pris kr. 2,00.

Grøndahl & Søns boktrykkeri, Oslo.