Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 70/17 Oslo, 16. desember 1970

VARIGE KONSUMGODER I NORSKE HUSHOLDNINGER 1967

Med et vedlegg om bruk av dummyvariable

AV

OLAV VANNEBO

I N N H O L D

Dette arbeid er opprinnelig skrevet som spesialoppgave ved det sosialøkonomiske studium. Forfatteren har stått fritt i valg av opplegg og undersøkelsesmetoder. Arbeidet gjengis her en del forkortet og med en del endringer som forfatteren har ønsket å foreta. Synspunkter og konklusjoner står for forfatterens regning.

INNLEDNING

Jeg vil i dette notatet vise et litt større eksempel på bruk av dummy-variable i et variansanalyseproblem. Dette eksemplet bygger for en stor del på min spesialoppgave som omhandler husholdningenes beholdning av en del varige konsumgoder.

Deretter vil jeg gi en kort oversikt om bruk av dummyvariable (binære variable, null-en-variable) i forbindelse med regresjonsanalyse, variansanalyse- og kovariansanalyseproblem. De enkelte tema vil ikke bli utførlig behandlet (lesere som ønsker grundig innføring henvises til mer seriøse lærebøker), vi skal bare skissere noen hovedtrekk i metodene og forsøke å gripe noe av den underliggende tankegang. Det forutsettes dog at leseren er litt fortrolig med vanlig regresjonsanalyse. Enkle eksempel vil illustrere framstillingen.

OM HUSHOLDNINGENES BEHOLDNING AV EN DEL VARIGE KONSUMGODER

Vi vil foreta en økonometrisk undersøkelse av hvilke faktorer som er bestemmende for hvorvidt en husholdning har eller ikke har diverse varige forbruksgoder. I Forbruksundersøkelse 1967 {3} heter det i innledningen bl.a.:[1] "Et av hovedformålene med forbruksundersøkelser er å klarlegge hvorledes forbrukssammensetningen i husholdninger varierer med familiestørrelse, familieinntekt, familiens alderssammensetning, og med miljøfaktorer som er knyttet til kjennetegn som f.eks. sosialgruppe, geografisk område, befolkningstetthet osv." Dette arbeidet kan ses som et supplement til disse bestrebelser.

Å lage en teori for beholdningen av varige goder i husholdningene, som foruten å være skikkelig logisk oppbygd skal være autonom overfor endringer i utenforliggende forhold, er en svært vanskelig oppgave. Vi skal heller ikke ta mål av oss til å gjøre det, men foretrekke en enklere argumentasjon som mer direkte appelerer til troverdighet, som ramme for vår analyse. Inntekt og formue peker seg ut som opplagte forklaringsvariable, noe vi også ville kommet fram til ved å ta utgangspunkt i en modell hvor husholdningene driver en form for nyttemaksimering over tid. Priser tas ikke med som forklaringsvariable da vi har tverrsnittsdata og følgelig ingen variasjon i prisene. Vi er dessuten interessert i en del bakgrunnsvariable som klassifiserer husholdningene etter bosted, størrelse, hovedpersonens yrkesstatus og alder. Vår hypotese er at vi på denne måten får grupper som er nokså homogene med hensyn til preferanser, og at det er forskjellig preferansestruktur fra gruppe til gruppe.

1) Tallet i slyngeparantesen { } viser til litteraturlisten bak.

Vi antar nå at en husholdnings beholdning av et varig gode i en periode t er en funksjon av følgende variable i samme periode: Husholdningens inntekt, husholdningens formue, husholdningens størrelse (og alderssammensetning), bosted, husholdningens eierforhold til sin bolig og dens beholdning av andre varige goder, hovedpersonens alder og hovedpersonens yrkesstatus. Husholdningens beholdning av en del andre varige goder er tatt med som forklaringsvariable med følgende begrunnelse: De ulike goder prefereres ikke med samme styrke og da man vanligvis ikke kjøper alle varige goder samtidig, vil anskaffelsen av de ulike goder følge et visst mønster. Vi antar at dette mønstret har visse generelle trekk. Jamfør ideen om "priority patterns" hos Pyatt {9}.

Dette er et opplegg som kan behandles som et variansanalyse- eller kovariansanalyseproblem. Det er særlig to argumenter som taler for at vi skal velge et variansanalyseopplegg. En ting er at de uavhengige variable ikke burde referere seg bare til periode t, men også til en del foregående perioder, og forventninger om disse variables utvikling i noen perioder framover. (Siden vi tror at beslutningene om anskaffelse av varige goder egentlig bør ses i et dynamisk perspektiv). Det vi nå gjør er å bruke de angitte variable som "proxies" for tidsrekkene. Når vi videre deler inn forklaringsvariablene i grupper, gjøres dette bl.a. ut fra en antakelse om at bevegelse i de variable over tiden stort sett vil skje innen de grenser som er satt for gruppene. En annen grunn til at vi foretrekker et variansanalyseopplegg er at vi slipper å forutsette noe om formen på funksjonen.

Definisjon av de variable:

$z_{ji} = \begin{cases} 1 & \text{hvis husholdning nr. i har varig gode nr. j} \\ 0 & \text{ellers} \end{cases}$

j=17...21 for henholdsvis bil, kjøleskap, hjemmefryser, vaskemaskin og oppvaskmaskin. Nummereringen er gjort slik for å være i overensstemmelse med variabelnummereringen i regresjonsprogrammet.

i= 1,2,3,.....,4148 (observasjonsnummer)

$x_{1i} = \begin{cases} 1 & \text{hvis husholdning nr. i eier sin bolig} \\ 0 & \text{ellers} \end{cases}$

$x_{2i} = \begin{cases} 1 & \text{hvis hovedpersonen er selvstendig næringsdrivende} \\ 0 & \text{ellers} \end{cases}$

$x_{3i} = \begin{cases} 1 & \text{hvis hovedpersonen er lønnstaker} \\ 0 & \text{ellers} \end{cases}$

$x_{2i} = x_{3i} = 0$ dvs. hovedpersonen er ikke yrkesaktiv

$x_{4i} = \begin{cases} 1 & \text{hvis husholdning nr.i's inntekt} \leq \text{kr. 15 000,-} \\ 0 & \text{ellers} \end{cases}$

$x_{5i} = \begin{cases} 1 & \text{hvis kr. 15 000,- < inntekten} \leq \text{kr. 30 000,-} \\ 0 & \text{ellers} \end{cases}$

$x_{6i} = \begin{cases} 1 & \text{hvis kr. 30 000,- < inntekten} \leq \text{kr. 50 000,-} \\ 0 & \text{ellers} \end{cases}$

$x_{4i} = x_{5i} = x_{6i} = 0$ dvs. husholdningens inntekt er større enn kr. 50 000,-

$x_{7i} = \begin{cases} 1 & \text{hvis husholdning nr. i's formue} \leq \text{kr. 5 000,-} \\ 0 & \text{ellers} \end{cases}$

$x_{8i} = \begin{cases} 1 & \text{hvis kr. 5 000,- < formuen} \leq \text{kr. 50 000,-} \\ 0 & \text{ellers} \end{cases}$

$x_{7i} = x_{8i} = 0$ dvs. husholdning nr. i's formue er større enn kr. 50 000,-

$x_{9i} = \begin{cases} 1 & \text{hvis antall forbruksenheter i husholdning nr. i} \leq 2 \\ 0 & \text{ellers} \end{cases}$

$x_{10i} = \begin{cases} 1 & \text{hvis 2 < antall forbruksenheter} \leq 3 \\ 0 & \text{ellers} \end{cases}$

$x_{11i} = \begin{cases} 1 & \text{hvis 3 < antall forbruksenheter} \leq 4 \\ 0 & \text{ellers} \end{cases}$

$x_{9i} = x_{10i} = x_{11i} = 0$ dvs. antall forbruksenheter er større enn fire

$x_{12i} = \begin{cases} 1 & \text{hvis hovedpersonens alder} \leq 25 \text{ år} \\ 0 & \text{ellers} \end{cases}$

$x_{13i} = \begin{cases} 1 & \text{hvis 25 år < hovedpersonens alder} \leq 35 \text{ år} \\ 0 & \text{ellers} \end{cases}$

$x_{14i} = \begin{cases} 1 & \text{hvis 35 år < hovedpersonens alder} \leq 50 \text{ år} \\ 0 & \text{ellers} \end{cases}$

$x_{12i} = x_{13i} = x_{14i} = 0$ dvs. hovedpersonen er eldre enn 50 år

$x_{15i} = \begin{cases} 1 & \text{hvis husholdningen er bosatt i Oslo, Bergen eller Trondheim} \\ 0 & \text{ellers} \end{cases}$

$x_{16i} = \begin{cases} 1 & \text{hvis husholdningen bor i by eller tettsted utenom Oslo, Bergen eller Trondheim} \\ 0 & \text{ellers} \end{cases}$

$x_{15i} = x_{16i} = 0$ dvs. husholdningen bor i et spredtbygd strøk

Det er selvsagt noe tilfeldig hvordan klassifiseringen er foretatt og hvilke grupper vi har valgt som basisgrupper. Det er dessuten mange muligheter for valg av basisgruppe og dummyvariable selv innen samme klassifisering.

Modellen vil nå se slik ut:

$$(10) \qquad z_{ji} = \beta_{jo} + \sum_{k=1}^{16} \beta_{jk} x_{ki} + \sum_{n=17}^{21} \beta_{jn} z_{ni} + u_{ji} \qquad (n \neq j)$$

$j = 17, 18, .. 21$

$i = 1, 2, ... 4148$

u_{ji} er et stokastisk restledd som svarer for den variasjon i z_{ji} som ikke kan forklares ved hjelp av de øvrige variable. Vi vil ikke spesifisere annet om fordelingen til u_{ji} enn at den har forventning null og ikke-konstant varians.

Vi har i denne modellen forutsatt at vi ikke har samspill. Det vil si at bidraget (til forklaring av z) av å tilhøre en gruppe langs en grupperingsvei (f.eks. inntekt) er uavhengig av hvilken gruppe telleenheten tilhører langs en annen grupperingsvei (f.eks. bosted). Dette innebærer videre at den totale virkning på z kan skrives som summen av gruppevirkningene (addivitetsforutsetningen). Modellen (10) er å betrakte som en hybrid form (dvs. en blanding av strukturform og redusert form), hvor vi tror å ha funnet fram til relevante forklaringsvariable[1].

Som vi ser er også venstresidevariablene dummies, og da støter vi på slike problem som er omtalt på sidene 22-25. Vi skulle gjerne transformert de variable for å få et restledd med konstant varians. Men vi skal estimere og teste modellen vår med et standard regresjonsprogram og dette standardprogrammet egner seg dårlig for en slik prosedyre. Vi skal derfor ignorere de vanskeligheter som følger av at vi har heteroscedastisitet. At estimatet $\hat{z}$ kan falle utenfor intervallet $[0,1]$ skal vi heller ikke bry oss så meget om. For ett av de varige godene skal vi se litt nærmere på metode C) på side 25.

Datamaterialet

Beholdningsdataene for de varige godene er fra innledningsintervjuene for Forbruksundersøkelsen, ellers er datamaterialet som brukes i denne oppgaven stort sett årsintervjuer fra 1967 innsamlet av Statistisk Sentralbyrå i forbindelse med Forbruksundersøkelsen 1967. 4 148 av landets husholdninger er blitt intervjuet. De venstresidevariable er forsåvidt greie, en svakhet er nok at man ikke får med opplysninger om mengde eller kvalitet, bare om husholdningen har eller ikke har diverse varige goder. Inntekts- og formuesdata er ligningstall fra skattestatistikken, og man kan nok gå ut fra at inntektstallene her er for lave i forhold til virkelig inntekt. Når det gjelder familiestørrelse, har vi på den tapen vi skal bruke, antall personer i hver av aldersklassene: 0-6 år, 7-13 år, 14-16 år, 17-19 år, 20-29 år, 30-39 år, 40-49 år, 50-59 år, 60-69 år, 70 år og over. Vi har (noe vilkårlig) benyttet følgende skala for forbruksenheter: En person i alderen 0-6 år regnes som 0,25 forbruksenheter, personer i alderen 7-13 år regnes som 0,5 forbruksenheter, personer i alderen 14-16 år regnes som 0,75 forbruksenheter, det gjør også personer over 60 år.

---

1) Estimering på en slik hybrid form reiser imidlertid problemer som neglisjeres her.

En person i alderen 17-59 år regnes som 1 forbruksenhet. Deretter er husholdningene gruppert etter størrelse som angitt på side 4. Vi har her forutsatt at virkningen av å ha personer i forskjellige aldersgrupper er additiv. Det kan godt tenkes (og kan påvises) at i visse tilfelle går virkningen på forbruksutgiften til et gode, når husholdningen øker med ett medlem, i motsatt retning om familiemedlemmet er et barn enn om det er voksent. (Se f.eks. {3} tabell 1 om drikkevarer og tobakk). Hva vi formodentlig burde gjort var å gruppere husholdningsstørrelsen langs flere veier. F.eks. som i Forbruksundersøkelsen hvor man har gruppert etter størrelse i hver av aldersgruppene 0-15 år, 16-69 år, 70-105 år.

Resultater og tolkning av første kjøring

Regresjonene er kjørt med Statistisk Sentralbyrås standardprogram for regresjonsanalyse på Byråets maskin IBM 360/40. Estimeringsmetode er minste kvadraters metode.

Hvis restleddsvariansen hadde vært konstant lik σ^2, hadde vi kunnet regne med at hver av observatorene $\hat{\beta}_{ji}$ var tilnærmet normalfordelt (se Johnston {6} side 116). Da ville det videre kunne vises at $\frac{\hat{\beta}_{ji} - \beta_{ji}}{STD\hat{\beta}_{ji}}$ er tilnærmet t-fordelt med n-k frihetsgrader (n - antall observasjoner, k - antall regresjonskoeffisienter). $STD\hat{\beta}_{ji}$ er det empiriske standardavviket for $\hat{\beta}_{ji}$ (se Johnston {6} side 118). Likeledes ville $\frac{\hat{\beta}_{ji} - \hat{\beta}_{jk} - (\beta_{ji} - \beta_{jk})}{STD(\hat{\beta}_{ji} - \hat{\beta}_{jk})}$ være t-fordelt med n-k frihetsgrader.

I vårt tilfelle har vi ikke konstant varians, og når vi likevel vil bruke estimatorene $\frac{\hat{\beta}_{ji} - \beta_{ji}}{STD\hat{\beta}_{ji}}$ til å teste hypoteser om β_{ji}, vil det si at vi opererer med en slags gjennomsnittsvarians for restleddene.

Vi vil nå teste hver av hypotesene: $\beta_{ji} = 0$ mot $\beta_{ji} \neq 0$ for alle i og j (ikke i=0 og i=j). Nå har vi ikke metoder for å gjøre dette simultant. Vi må derfor gjennomføre en rekke enkelttester på samme materialet, og det kan bli vanskelig å vite hvilket endelig nivå man får for hele testen. La $t_{ji} = \frac{\hat{\beta}_{ji}}{STD\hat{\beta}_{ji}}$. Vi vil nå forkaste hypotesen om at $\beta_{ji} = 0$ dersom $|t_{ji}| > t_{1-\xi,\, n-k}$ hvor $t_{1-\xi,n-k}$ er $(1-\xi)$-fraktilen i t-fordelingen med n-k frihetsgrader.

Tabell nr. 1 på side 8 angir hvilke regresjonskoeffisienter som er signifikant forskjellig fra null når vi velger $\xi = 5\%$. Minustegn angir negative og plusstegn angir positive estimater. At vi velger $\xi = 5\%$ vil si at

vi skal forkaste hypotesen dersom $|t_{ji}| > 1{,}96$. Det at vi tester flere hypoteser på samme materialet vil bidra til at "over-all"-nivået for hele testen vil bli større enn 5%.

Koeffisientene i tabell 2 angir hvordan beholdningen av et varig gode for en gruppe, er i forhold til basisgruppen for denne grupperingsvei. F.eks. når det gjelder inntekt, er basis den høyeste inntektsgruppen (over kr. 50 000,-). Da skulle vi vente negative tall for koeffisientene nr. 4,5 og 6, og slik at $|\beta_{j4}|>|\beta_{j5}|>|\beta_{j6}|$, noe som stort sett også er tilfelle. Tolkningen av koeffisientene $\beta_{j17}.....\beta_{j21}$ er noe problematisk. Man kan neppe trekke noen slutninger om prioriteringsmønsteret på grunnlag av disse.

For variabel nr. 19, hjemmefryser, får vi overhodet ikke signifikante koeffisienter for inntektsvariablene. Likeledes får vi for variabel nr. 21, vaskemaskin, ikke signifikante koeffisienter for inntektsvariablene nr. 4 og nr. 6, men en signifikant positiv koeffisient for variabel nr. 5. Siden inntekten er den viktigste av våre forklaringsvariable vil vi slutte at vi har en dårlig teori for de varige godene hjemmefryser og vaskemaskin. Når det gjelder forklaringsvariablene eierforhold til bolig og yrkesstatus, er resultatene litt motstridende og til dels ikke signifikante. De bidrar neppe noe særlig til å forklare hvorvidt en husholdning har eller ikke har forskjellige varige goder. For bil later det til at følgende er viktige forklaringsvariable: inntekt, formue, husholdningens størrelse og hovedpersonens alder. Koeffisienten for beholdning av kjøleskap er også signifikant positiv. For kjøleskap ser variablene inntekt, formue, husholdningens størrelse, boligstrøk og dessuten beholdning av bil, fryseskap og vaskemaskin ut til å være de viktigste forklaringsvariablene. For oppvaskmaskin ser det ut til at inntekt, formue og boligstrøk er de viktigste forklaringsvariablene. Variablen beholdning av fryseskap er også signifikant. Noen helt klar tale synes likevel ikke å ligge i disse tallene.

Vi er dessuten interessert i å teste om det er signifikant forskjell mellom koeffisientene innen inntekts-, formues-, størrelses- og aldersgruppene. Altså ikke bare i forhold til basis. Til dette trenger vi kovarianser mellom estimatene. Alternativt kunne vi kjørt flere ganger og stadig valgt nye basisgrupper, og testet de nye koeffisientene i forhold til den nye basis. Dette vil kreve mye maskintid i vårt tilfelle. Siden kovariansene mellom estimatene ikke er standardoutput i programmet vårt, har vi laget et lite tilleggsprogram for å regne ut disse. Denne prosedyren er såpass arbeidskrevende at vi bare skal gjennomføre den for ett av de varige godene.

TABELL NR. 1

	Eier-forhold til bolig	Hoved-personens yrkes-status		Inntekt			For-mue		Hushold-ningens størrelse			Hoved-personens alder			Bolig-strøk		Bil	Kjøle-skap	Frys-skap	Opp-vask-maskin	Vaske-maskin	Multi-ppel korrela-sjons-koeff.
Variabel nr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Bil 17	-	+	+	-	-		-	-		+	+	+	+	+				+				0,50
Kjøle-skap 18		-		-	-		-	-	+	+	+	-			+	+	+		+		+	0,42
Frys-skap 19	+						-	-						+	-	-	+	+		+	+	0,38
Oppvask-maskin 20		+		-	-	-		-			+			+	+	+			+			0,25
Vaske-maskin 21					+		-	-	-	-		-			-			+	+			0,49

TABELL NR. 2

OVERSIKT OVER RESULTATENE AV FØRSTE KJØRING

Høyreside-variable / Venstreside-variable		Eier-forhold til bolig	Hovedpersonens yrkesstatus		Inntekt			Formue	
		1	2	3	4	5	6	7	8
Bil	17	-0,01108	0,04856	0,09407	-0,37631	-0,26719	-0,04767	-0,25291	-0,12084
		(0,01709)	(0,02431)	(0,02232)	(0,03216)	(0,02721)	(0,02698)	(0,02561)	(0,02204)
Kjøleskap	18	0,03061	-0,08886	-0,00187	-0,21921	-0,12568	-0,04444	-0,04963	-0,04330
		(0,01596)	(0,02266)	(0,02089)	(0,03033)	(0,02562)	(0,02518)	(0,02418)	(0,02064)
Hjemmefryser	19	0,06415	0,03143	-0,02621	-0,05861	-0,03763	-0,01635	-0,22053	-0,14827
		(0,01736)	(0,02472)	(0,02274)	(0,03322)	(0,09727)	(0,02743)	(0,02612)	(0,02237)
Oppvaskmaskin	20	0,00472	0,01332	-0,00599	-0,06644	-0,07340	-0,06967	-0,00846	-0,02029
		(0,00457)	(0,00649)	(0,00597)	(0,00867)	(0,00726)	(0,00712)	(0,00692)	(0,00590)
Vaskemaskin	21	0,02058	0,02522	0,00127	-0,01994	0,05773	0,01347	-0,11973	-0,06167
		(0,01599)	(0,02273)	(0,02091)	(0,03056)	(0,02571)	(0,02522)	(0,02415)	(0,02066)

		Husholdningens størrelse			Hovedpersonens alder			Boligstrøk	
		9	10	11	12	13	14	15	16
Bil	17	-0,01657	0,06196	0,05942	0,17112	0,18314	0,15175	-0,03034	-0,01131
		(0,02831)	(0,02645)	(0,02722)	(0,03862)	(0,02268)	(0,01767)	(0,02035)	(0,01625)
Kjøleskap	18	0,17698	0,15145	0,11449	-0,23708	0,00225	0,02003	0,19952	0,16213
		(0,02628)	(0,02459)	(0,02536)	(0,03595)	(0,02134)	(0,01664)	(0,01875)	(0,01496)
Hjemmefryser	19	-0,03816	-0,00479	-0,04391	-0,02271	-0,04275	0,06318	-0,09375	-0,13887
		(0,02877)	(0,02691)	(0,02768)	(0,03936)	(0,02323)	(0,01810)	(0,02064)	(0,01638)
Oppvaskmaskin	20	0,00122	0,00949	0,01806	0,00095	-0,00510	0,01046	0,02048	0,01741
		(0,00756)	(0,00706)	(0,00727)	(0,01034)	(0,00610)	(0,00476)	(0,00543)	(0,00433)
Vaskemaskin	21	-0,27970	-0,08655	-0,03410	-0,13949	-0,03044	0,02460	-0,19076	0,00791
		(0,02610)	(0,02470)	(0,02545)	(0,03612)	(0,02137)	(0,01666)	(0,01879)	(0,01519)

		Bil	Kjøleskap	Hjemme-fryser	Oppvask-maskin	Vaske-maskin	Multippel korr-koeff.	Konstant-ledd	Estimert varians
		17	18	19	20	21			
Bil	17		0,08518	0,03006	-0,04737	0,02521	0,50014	0,55204	0,42653
			(0,01662)	(0,01530)	(0,05830)	(0,01668)			
Kjøleskap	18	0,07423		0,03041	0,06053	0,17804	0,41757	0,49519	0,39819
		(0,01449)		(0,01429)	(0,05442)	(0,01530)			
Hjemmefryser	19	0,03107	0,3606		0,24987	0,13995	0,38280	0,41407	0,43362
		(0,01582)	(0,01694)		(0,05915)	(0,01679)			
Oppvaskmaskin	20	-0,00338	0,00495	0,01723		0,00665	0,25336	0,05716	0,11388
		(0,00416)	(0,00445)	(0,00408)		(0,00444)			
Vaskemaskin	21	0,02203	0,17849	0,11831	0,08152		0,48996	0,73754	0,39869
		0,01455	(0,01534)	(0,01419)	(0,05448)				

Tallene uten parantes angir estimat for regresjonskoeffisientene $(\hat{\beta}_{ji}\ \begin{matrix} j = 17,...21 \\ i = 1,2,..)j(.21 \end{matrix})$

Tallene i parantes angir standardavviket for estimatene.

Av tabell nr. 3, ses at alle inntektskoeffisientene (nr. 4,5 og 6) er signifikant forskjellig fra hverandre. Likeledes er formueskoeffisientene (nr. 7 og 8) signifikant forskjellig fra hverandre. Når det gjelder størrelsesvariablene (9,10 og 11) ser vi at regresjonskoeffisienten knyttet til variabel nr. 9 er signifikant mindre enn koeffisientene foran variablene nr. 10 og nr. 11. Men koeffisient nr. 10 er ikke signifikant forskjellig fra koeffisient nr. 11. Videre husker vi fra tabell nr. 1 at koeffisientene nr. 10 og 11 var signifikant positive i forhold til basis. Ut fra dette vil vi stille oss litt skeptisk til familiestørrelsen som forklaringsfaktor for hvorvidt en husholdning har eller ikke har bil. Det kunne forøvrig tenkes at nettopp mellomstore familier har en relativt høy tendens til å ha bil i forhold til små eller særlig store familier. Og det kan føres rasjonelle argumenter for at det er slik. Når størrelsen øker, får dette en inntektseffekt som skulle tilsi nedsatt forbruk. En annen effekt er substitusjon fra individuelle goder til fellesgoder. Effektene går altså i motsatt retning, og de kan godt ha forskjellig styrke i forskjellige områder. Videre ser vi at når det gjelder hovedpersonens alder, er koeffisient nr. 12 forskjellig fra nr. 13 og nr. 14. Derimot er koeffisient nr. 13 ikke signifikant forskjellig fra nr. 14. Når det gjelder boligstrøk ser en at koeffissient nr. 15 ikke er utsagnskraftig forskjellig fra nr. 16. (Vi husker også at ingen av disse var signifikant forskjellig fra null).

TABELL NR. 3

TABELL OVER $\frac{\hat{\beta}_{ji} - \hat{\beta}_{jk}}{STD(\hat{\beta}_{ji} - \hat{\beta}_{jk})} = t^{1}_{ik}$

i \ k	5	6	8	10	11	13	14	16
4	-5,45	-13,98						
5		-12,68						
7			-7,90					
9				-4,26	-3,15			
10					0,32			
12						-0,30	0,49	
13							1,40	
15								-0,98

Forkastingsregel:

Forkast hypotesen om at $\beta_{ji} = \beta_{jk}$ dersom $|t^{1}_{ik}| > 1,96$

Annen kjøring - Estimering av sannsynligheter

Vil nå se litt videre på det varige godet bil, (som nå blir variabel nr. 9) med formål å estimere den betingede sannsynligheten for at en husholdning skal være i besittelse av dette godet. Vi vil benytte den metoden som er skissert på side 25. Som høyresidevariabel vil vi nå bruke inntekt, (tidligere variabelnr. 4,5 og 6, nå nr. 1,2 og 3) formue (før nr. 7 og 8, nå nr. 4 og 5) og alder (før variabelnr. 12,13 og 14, nå nr. 6,7 og 8). Vi har delt materialet inn i grupper som er homogene med hensyn på de høyresidevariable. Gruppene er angitt på side 12.

Deretter har vi estimert koeffisientene i den nye regresjonsligningen på vanlig måte. Vi har nå lagt inn en rutine som angir hvilken gruppe en observasjon tilhører. Foruten resultatene av estimeringen (se tabell nr. 4 side 14) vil vi få utskrevet fortløpende gruppenummer og estimerte restledd på hver sin tape.

Når det gjelder tallene i tabell nr. 4, ser en at forholdet mellom de estimerte koeffisienter og de respektive standardavvik ser ut til å være langt gunstigere fra et signifikanssynspunkt enn ved første kjøring. Nå er det klart at vi ikke kan teste koeffisientene ved denne kjøringen og samtidig uttale oss om nivået på testen. Man kan likevel bruke disse tallene til forsvar for ikke å ta med de variablene som nå er utelatt.

Gruppering etter verdi på de høyresidevariable:

Var.nr. / Gruppe nr.	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0
3	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	1
5	0	0	0	1	0	0	0	0
6	0	0	0	1	0	1	0	0
7	0	0	0	1	0	0	1	0
8	0	0	0	1	0	0	0	1
9	0	0	0	0	1	0	0	0
10	0	0	0	0	1	1	0	0
11	0	0	0	0	1	0	1	0
12	0	0	0	0	1	0	0	1
13	1	0	0	0	0	0	0	0
14	1	0	0	0	0	1	0	0
15	1	0	0	0	0	0	1	0
16	1	0	0	0	0	0	0	1
. . . .								
47	0	0	1	0	1	0	1	0
48	0	0	1	0	1	0	0	1

Ved hjelp av restleddene og gruppenumrene har vi kjørt ut en tabell over:

Z ESTIMAT $\hat{Z}_k = \hat{\beta}_o + \sum_{j=1}^{8} \hat{\beta}_i X_{ik}$

for hver gruppe, dvs. k=1,2.....48,

GJENNOMSNITT AVVIK $F(\hat{Z}_k) = \frac{1}{n_k} \sum^{n_k} (\hat{Z}_k - Z_i)$

for hver gruppe. n_k er antall observasjoner i gruppe nr. k. Z er renummerert slik at i starter på 1 i hver gruppe.

KORRIGERT ESTIMAT $\tilde{Z}_k = \hat{Z}_k + F(\hat{Z}_k)$

for hver gruppe. Gruppene er beskrevet på side 12. Hvis n_k er null er $F(\hat{Z}_k)$ satt lik null.

Tabellen har vi på side 15. En ser av tabellen at en av de estimerte z-verdier er mindre enn null og at to er lik null. Når det gjelder de to nullene, skyldes dette at ingen observasjoner har falt i gruppene 10 eller 38. For disse gruppene finner vi:

$$\hat{Z}_{10} = \tilde{Z}_{10} = \hat{\beta}_0 + \hat{\beta}_5 + \hat{\beta}_6 = 0{,}79302$$

$$\hat{Z}_{38} = \tilde{Z}_{38} = \hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_6 = 0{,}81825$$

La oss se litt nærmere på de KORRIGERTE ESTIMATER som er blitt lik 1,ooooo. Det gjelder gruppene 2,3,14 og 46. De to første har høyeste inntekts- og formuesgruppe og henholdsvis laveste og nestlaveste aldersgruppe. Nummer 14 har laveste inntektsgruppe, høyeste formuesgruppe og laveste aldersgruppe. Den siste har nesthøyeste inntekts- og formuesgruppe og nestlaveste aldersgruppe. Tallene virker ikke helt urimelige.

Videre ses at korreksjonene er til dels ganske kraftige (opp til 3/2 av opprinnelig verdi). (Tallene i tabell 5 kan godt brukes til å illustrere at forventningsretthet og konsistens er forholdsvis svake krav til estimatorer). Når en ser på de korrigerte estimatverdier, ser en at alle ligger i intervallet $[0,1]$, men at hele 5 (7-2) verdier ligger på intervallgrensene. Vi vil på grunnlag av disse tallene stille oss litt skeptiske til metoden.

TABELL NR. 4

VARIABLE NO.	MEAN	STANDARD DEVIATION	CORRELATION X VS Z	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEF.	COMPUTED T VALUE
1	0.26736	0.44263	-0.35451	-0.49970	0.02837	-17.61372
2	0.39682	0.48930	-0.03152	-0.30908	0.02675	-11.55382
3	0.25892	0.43809	0.27845	-0.07254	0.02756	-2.63201
4	0.35366	0.47816	-0.16222	-0.22667	0.02275	-9.96204
5	0.51688	0.49978	0.06182	-0.09777	0.02106	-4.64175
6	0.03592	0.18612	-0.01435	0.18309	0.03752	4.88036
7	0.13862	0.34559	0.08946	0.21868	0.02132	10.26533
8	0.31148	0.46315	0.20638	0.17059	0.01592	10.71254
DEPENDENT						
9	0.39200	0.48825				

INTERCEPT	0.70770
MULTIPLE CORRELATION	0.47781
STD. ERROR OF ESTIMATE	0.42933

ANALYSIS OF VARIANCE FOR THE REGRESSION

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F VALUE
ATTRIBUTABLE TO REGRESSION	8	225.70242	28.21280	153.06197
DEVIATION FROM REGRESSION	4139	762.91185	0.18432	
TOTAL	4147	988.61427		

TABELL NR. 5

GRUPPE	Z ESTIMAT	GJENNOMSNITT AVVIK	KORRIGERT ESTIMAT
1	0.70770	0.00063	0.70833
2	0.89079	0.10921	1.00000
3	0.92638	0.07362	1.00000
4	0.87829	-0.19647	0.68182
5	0.48103	0.08419	0.56522
6	0.66412	-0.16412	0.50000
7	0.69971	-0.07471	0.62500
8	0.65162	-0.08495	0.56667
9	0.60992	0.11981	0.72973
10	0.0	0.0	0.0
11	0.82860	-0.32860	0.50000
12	0.78052	0.05510	0.83562
13	0.20800	0.02729	0.23529
14	0.39109	0.60891	1.00000
15	0.42667	-0.42667	0.0
16	0.37859	-0.07859	0.30000
17	-0.01867	0.03938	0.02071
18	0.16442	0.05780	0.22222
19	0.20001	0.06086	0.26087
20	0.15192	0.11731	0.26923
21	0.11022	-0.04087	0.06935
22	0.29331	0.04002	0.33333
23	0.32890	-0.00890	0.32000
24	0.28082	-0.08329	0.19753
25	0.39862	-0.01337	0.38525
26	0.58171	-0.08171	0.50000
27	0.61730	0.04937	0.66667
28	0.56922	0.05300	0.62222
29	0.17195	-0.05290	0.11905
30	0.35505	-0.08231	0.27274
31	0.39063	-0.00601	0.38462
32	0.34255	0.00656	0.34911
33	0.30085	0.01293	0.31378
34	0.48394	0.20837	0.69231
35	0.51953	0.10311	0.62264
36	0.47144	-0.01361	0.45783
37	0.63516	0.08447	0.71963
38	0.0	0.0	0.0
39	0.85384	-0.10384	0.75000
40	0.80576	-0.03527	0.77049
41	0.40849	-0.03640	0.37209
42	0.59159	0.00841	0.60000
43	0.62717	-0.10604	0.52113
44	0.57909	0.02231	0.60140
45	0.53739	-0.03160	0.50579
46	0.72048	0.27952	1.00000
47	0.75607	-0.00265	0.75342
48	0.70798	0.03408	0.74206

Tredje kjøring - Testing av samspillseffekter

Vi er interessert i å teste hvorvidt det er samspillseffekter tilstede mellom gruppene. Det vil si om virkningen av å tilhøre en gruppe langs en grupperingsvei er avhengig av hvilken gruppe man tilhører langs en annen grupperingsvei. Vi velger bil som venstresidevariabel og forklaringsvariablene inntekt, formue og hovedpersonens alder. For å redusere antall variable brukes bare tre inntektsgrupper og to formuesgrupper. Vi har fire aldersgrupper som tidligere.

Definisjon av de variable:

$x_{1i} = \begin{cases} 1 \text{ hvis husholdningens inntekt} \leq 25\,000 \text{ kroner} \\ 0 \text{ ellers} \end{cases}$

$x_{2i} = \begin{cases} 1 \text{ hvis kr. } 25\,000 < \text{inntekten} \leq \text{kr. } 50\,000 \\ 0 \text{ ellers} \end{cases}$

$x_{1i} = x_{2i} = 0$ dvs. inntekten er større enn 50 000 kroner

$x_{3i} = \begin{cases} 1 \text{ hvis husholdningens formue} \leq \text{kr. } 5\,000 \\ o \text{ ellers} \end{cases}$

x_{4i}, x_{5i}, x_{6i} er definert på samme måte som henholdsvis x_{12i}, x_{13i} og x_{14i} på side 4.

$x_{7i} = x_{1i} * x_{3i}$

$x_{8i} = x_{2i} * x_{3i}$

$x_{9i} = x_{1i} * x_{4i}$

$x_{10i} = x_{1i} * x_{5i}$

$x_{11i} = x_{1i} * x_{6i}$

$x_{12i} = x_{2i} * x_{4i}$

$x_{13i} = x_{2i} * x_{5i}$

$x_{14i} = x_{2i} * x_{6i}$

$x_{15i} = x_{3i} * x_{4i}$

$x_{16i} = x_{3i} * x_{5i}$

$x_{17i} = x_{3i} * x_{6i}$

$x_{18i} = x_{7i} * x_{4i}$

$x_{19i} = x_{7i} * x_{5i}$

$x_{20i} = x_{7i} * x_{6i}$

$x_{21i} = x_{8i} * x_{4i}$

$x_{22i} = x_{8i} * x_{5i}$

$x_{23i} = x_{8i} * x_{6i}$

* angir multiplikasjon

z_{24i} er definert på samme måte som z_{17i} på side 3.

Modellen vår ser nå slik ut:

$$(11) \qquad z_{24i} = b_0 + \sum_{j=1}^{23} b_{24j} x_{ji} + v_i$$

Hvor variablene $x_{7i}, x_{8i}, \ldots, x_{23i}$ er samspillsvariable og koeffisientene $b_{24,7}, b_{24,8}, \ldots\ldots, b_{24,23}$ gir uttrykk for samspillseffekter. Vår null-hypotese er at hver av koeffisientene $b_{24,7}, \ldots\ldots, b_{24,23}$ er lik null.

Resultatene av kjøringen er angitt i tabell nr. 6. Vi ser at koeffisientene stort sett ikke er signifikant forskjellige fra null (en av sytten koeffisienter er større enn 1,96 i tallverdi). Vi kan selvsagt ikke derav slutte at det ikke er samspillseffekter tilstede, men er ikke bekymret over å ha forutsatt at slike effekter ikke forekommer.

TABELL NR. 6

VARIABLE NO.	MEAN	STANDARD DEVIATION	CORRELATION X VS Z	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEF.	COMPUTED T VALUE
1	0.53881	0.49855	-0.38150	-0.50177	0.05255	-9.54870
2	0.38428	0.48648	0.28746	-0.14526	0.05428	-2.67609
3	0.69720	0.45952	-0.15762	-0.10898	0.07282	-1.49660
4	0.03592	0.18612	-0.01435	0.25610	0.43729	0.58564
5	0.13862	0.34559	0.08946	0.13110	0.16099	0.81430
6	0.31148	0.46315	0.20638	-0.00861	0.07129	-0.12075
7	0.40164	0.49029	-0.33953	-0.05017	0.07723	-0.64972
8	0.25699	0.43703	0.16401	-0.11750	0.08051	-1.45941
9	0.02724	0.16281	-0.03123	0.10177	0.47903	0.21245
10	0.07015	0.25544	0.00180	0.24216	0.18342	1.32023
11	0.11813	0.32280	-0.04143	0.21448	0.08397	2.55421
12	0.00796	0.08885	0.02259	0.14526	0.61708	0.23540
13	0.06268	0.24242	0.11222	-0.07974	0.17688	-0.45081
14	0.15791	0.36470	0.23054	0.14572	0.08188	1.77977
15	0.03423	0.18185	-0.02081	-0.39102	0.53729	-0.72775
16	0.12078	0.32591	0.06457	-0.20352	0.20181	-1.00849
17	0.21842	0.41322	0.11576	0.10786	0.10233	1.05404
18	0.02604	0.15926	-0.03515	0.23721	0.57348	0.41364
19	0.06389	0.24458	-0.00985	0.12088	0.22218	0.54403
20	0.08799	0.28332	-0.05940	-0.09808	0.11473	-0.85488
21	0.00771	0.08750	0.01951	0.11750	0.69620	0.16877
22	0.05304	0.22414	0.09863	0.37545	0.21791	1.72294
23	0.11138	0.31464	0.16308	-0.00025	0.11419	-0.00221
DEPENDENT						
24	0.39200	0.48825				

INTERCEPT	0.74390
MULTIPLE CORRELATION	0.46034
STD. ERROR OF ESTIMATE	0.43465

ANALYSIS OF VARIANCE FOR THE REGRESSION

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARES	F VALUE
ATTRIBUTABLE TO REGRESSION	23	209.49991	9.10869	48.21403
DEVIATION FROM REGRESSION	4124	779.11436	0.18892	
TOTAL	4147	988.61427		

LITT OM BRUK AV DUMMYVARIABLE

Regresjonsanalyse, variansanalyse, kovariansanalyse

Ordinære regresjonsligninger er oftest av typen: $y = x + \sum_{i=1}^{k} \beta_i x_i + u$, hvor alle de variable y og x-ene er observerbare og kvantifiserbare. U - det stokastiske restledd - er ikke-observerbart (latent). Formålet med regresjonsanalyse er svært ofte å estimere og teste størrelsen på koeffisientene x og β_i (i=1,...k). Å teste $\beta_j = 0$ er det samme som å teste hvorvidt endringer i forklaringsvariabel nr. j virkelig bidrar noe til å forklare variasjoner i y.

Variansanalyse er en prosedyre for å teste relevansen av en klassifisering. Tabelloppstillinger har også til formål å vise samvariasjon mellom variable som er klassifisert etter visse kriterier. Disse metoder atskiller seg bl.a. ved at variansanalysen krever mer av forutsetninger, til gjengjeld fås skarpere konklusjoner. La oss gruppere noen observasjoner i klasser etter kjennetegn (kvantitative eller kvalitative). Vi kan gruppere en eller flere veier. En grupperingsvei vil si en fullstendig klassifisering etter ett av kjennetegnene. Hvis vi f.eks. grupperer personer etter alder og yrkesstatus har vi en toveisgruppering. Tror vi videre at f.eks. gjennomsnittsinntekten for de forskjellige grupper varierer, har vi en situasjon som kan behandles med et variansanalyseopplegg. Hvis det nå er slik at virkningen (på inntekten) av å tilhøre en gruppe langs én grupperingsvei, er den samme uansett hvilken gruppe man tilhører langs en annen vei, sier vi at vi ikke har samspill. I motsatt fall, når vi foruten direkte-virkningene får tilleggsvirkninger ved at grupper langs forskjellige veier opptrer samtidig, sier vi at vi har samspill mellom gruppene.

For å forsøke å få fram noe av hovedideen ved variansanalysen, skal vi se på en enveisgruppering. Anta at vi har et sampel av inntektstakere som er gruppert etter alder. Vi vil teste om det er forskjell i gjennomsnittsnivået på inntekten i gruppene når variansen antas å være konstant fra gruppe til gruppe[1)]. Det vil intuitivt være rimelig å forkaste en hypotese om at alle gruppegjennomsnittene er like dersom den empiriske variansen mellom gruppegjennomsnittene er stor i forhold til variansen i hele samplet, eller om man vil, er stor i forhold til gjennomsnittet av gruppevariansene. For utledning av testobservatorer se H.T. Amundsen {2}. Kap. 9.6

Kovariansanalyse er en kombinasjon av variansanalyse og regresjonsanalyse. En situasjon som egner seg for et slikt opplegg kan f.eks. være når konstantleddet i en regresjonsligning er avhengig av hvilken klasse en

---

1) Eksemplet er litt uheldig valgt, siden forutsetningen om konstante varianser neppe holder

observasjon tilhører, mens de øvrige regresjonskoeffisienter er like for alle grupper.

Dette er i overensstemmelse med den vanlige bruk av betegnelsene variansanalyse og kovariansanalyse. (Jamfør forfattere som H.T. Amundsen, Malinvaud, Goldberger).

Både variansanalyse og kovariansanalyse kan bringes over til et regresjonsproblem ved hjelp av dummyvariable.

Et eksempel på kovariansanalyse med dummyvariable

Anta at et individs forbruk er en lineær funksjon av individets inntekt. Dvs. (1) $Y = \alpha + \beta x + U$, Y er forbruksutgift, x er disponibel inntekt, U er et stokastisk restledd[1], α og β er konstanter.

En hypotese kan være at nivået på konsumet er høyere i bystrøk enn på landet. Dvs. at α er større i tettsted enn i grisgrendte strøk.

(2) $Y_i = \alpha_1 + \beta x_i + U_i$ hvis individ nr. i bor i by

(3) $Y_i = \alpha_2 + \beta x_i + U_i$ hvis individ nr. i bor på landet.

Anta at vi er interesserte i å estimere koeffisientene α_1, α_2 og β og at vi vil teste $\alpha_1 = \alpha_2$ mot $\alpha_1 \neq \alpha_2$. En mulighet er selvsagt å kjøre to regresjoner, en for bybefolkningen og en for landbefolkningen. Vi ville da neppe fått samme estimat for β, men kunne godt testet om α_1 og α_2 var like. Ved å innføre dummyvariable kan vi klare oss med å kjøre en regresjon og kan estimere β ved å bruke alle observasjonene samtidig. Dette kan gjøres slik:

Vi innfører variablene $z_{1i} = \begin{cases} 1 & \text{hvis individ nr. i bor i by} \\ 0 & \text{ellers} \end{cases}$

og $z_{2i} = \begin{cases} 1 & \text{hvis individ nr. i bor på landet} \\ 0 & \text{ellers} \end{cases}$

(Det er her bare to muligheter, enten bor man i en by eller så bor man på landet, altså $z_1 + z_2 \equiv 1$).

Vi ser at vi kan skrive ligningene (2) og (3) enten som:

(4) $Y_i = \alpha_1 z_{1i} + \alpha_2 z_{2i} + \beta x_i + U_i$ eller som:

(5) $Y_i = \alpha_2 + (\alpha_1 - \alpha_2) z_{1i} + \beta x_i + U_i$

Ved å sette inn verdier for z_1 og z_2 ser en at det stemmer.

Vi kan velge en av disse to ligningene ((4) eller (5)), og estimere koeffisientene α_1, α_2 og β på vanlig måte med minstekvadraters metode. Fordelen er at vi får alt på en ligning, og at vi bruker alle observasjonene når vi estimerer β.

---

1) Også her er forutsetningen om konstant varians noe tvilsom.

Betrakt muligheten for å skrive ligningene (2) og (3) slik:

(6) $Y_i = \gamma + \eta_1 z_{1i} + \eta_2 z_{2i} + \beta x_i + U_i$

hvor $\eta_j = (\alpha_j - \gamma)$, $j = 1,2$

En rimelig tolkning av koeffisientene her vil være: γ som et gjennomsnittsnivå og η_j som avvik fra dette gjennomsnittsnivået. Vi ser at vi også her får tilbake (2) og (3) når vi setter inn henholdsvis z_1 og z_2. Men vi får her inn en ubestemthet i γ, for vi ser at $Y_i = \delta_0 + \delta_1 z_{1i} + \delta_2 z_{2i} + \beta x_i + U_i$ hvor $\delta_i = (\alpha_i - \delta_0)$, $(i = 1,2)$, passer like godt.

Det vil si at koeffisientene γ og $\eta_i (i = 1,2)$ ikke er identifiserbare, og grunnen til dette er at vi har en lineær sammenheng mellom z_1 og z_2. Hvis vi forsøker å sette regresjonsligningen på formen (6), må vi innføre en tilleggsbeskranking for å få determinerthet. Det kan vi si å ha gjort i (4) hvor vi har valgt å sette $\gamma = 0$, og i (5) hvor vi har satt inn $z_2 = (1 - z_1)$ i (6). Dette er de vanlige måter å innføre restriksjoner på, men det kan tenkes andre. Se Suits {10} side 549. I (5) sier vi at vi har valgt gruppe nr. 2 (landsbefolkningen) som basisgruppe eller referansegruppe.

Hvis vi har en toveisgruppering uten samspill, f.eks. med bosted og yrkesstatus som grupperingsveier, kan vi gjøre det slik når vi innfører dummies:

a) dropper konstantleddet og en dummyvariabel langs en av grupperingsveiene
b) tar med konstantleddet og dropper en dummy langs hver av grupperingsveiene.

Det er ingen realitetsforskjell mellom disse metoder.

Anta at vi har en toveisgruppering med samspill og at grupperingen ser slik ut:

Bosted / Yrkesstatus	BY	LAND
SELVSTENDIG		
ANSATT		
IKKE YRKESAKTIV		

Som man ser av tabellen er det i alt 6 (2x3) mulige grupper. Man kan da velge en av disse som basis, og innføre fem dummyvariable for de andre gruppene. (Eventuelt sløyfe konstantleddet og innføre seks dummies). Alternativt kunne vi innføre en dummy mindre enn antall grupper langs hver grupperingsvei (som når vi ikke regnet med samspill) og i tillegg til dette innføre samspillsvariable som er hver dummy langs en vei multiplisert med hver dummy langs den andre grupperingsveien.

Definer $z_1 = \begin{cases} 1 \text{ hvis telleenheten er en byborger} \\ 0 \text{ ellers} \end{cases}$

$z_2 = \begin{cases} 1 \text{ hvis telleenheten er selvstendig næringsdrivende} \\ 0 \text{ ellers} \end{cases}$

$z_3 = \begin{cases} 1 \text{ hvis telleenheten er ansatt} \\ 0 \text{ ellers} \end{cases}$

$z_2 = z_3 = 0$ dvs. telleenheten er ikke yrkesaktiv. Disse er da variablene for egenvirkningene (z_1 for bosted, z_2 og z_3 for yrkesstatus). Vi innfører i tillegg samspillsvariablene $z_4 = z_1 * z_2$ og $z_5 = z_1 * z_3$ ($*$ angir multiplikasjon). Dette kan lett generaliseres til grupperinger langs flere veier enn to. Når det gjelder samspillsvariablene vil vi ved fullstendig utfylling foruten annenordens kryssprodukt, få kryssprodukt av alle ordner opp til det antall grupperingsveier vi har. Regner en med ufullstendig samspill, tas med kryssprodukt opp til en lavere orden enn antall grupperingsveier.

La oss vende tilbake til eksemplet med konsumfunksjonen. Vi kan også godt teste en hypotese om at β (den marginale forbrukstilbøyelighet) er større i byer enn på landet.

(7) $Y_i = \alpha + \beta_1 x_i + U_i$ for bybefolkningen

(8) $Y_i = \alpha + \beta_2 x_i + U_i$ for landbefolkningen[1)]

Vi ser at vi kan skrive (7) og (8) som en ligning slik:

(9) $Y_i = \alpha + \beta_2 x_i + (\beta_1 - \beta_2) Z_{1i} x_i + U_i$

Vi kan nå estimere α, β_1 og β_2 ved hjelp av minstekvadraters metode og deretter teste om $\beta_1 = \beta_2$. Vi bruker her hele materialet når vi estimerer α. Ifølge Clopper Almons terminologi ({1} side 134) kalles testing av likhet i konstantleddet for variansanalyse og testing av likhet i vinkelkoeffisienter for kovariansanalyse. Hvis vi imidlertid tror at både nivå og vinkelkoeffisient er forskjellige, bruker vi data like godt om vi kjører to separate regresjoner som når vi bruker dummies.

---

1) NB! Samme fordeling på restleddet i de to tilfellene.

Variansanalyse med dummyvariable

Anta at vi har en variabel som vi vil forsøke å forklare variasjoner i. Hvis vi har arrangert alle relevante forklaringsfaktorer- eller variable i grupper i et en- eller flerveissystem, har vi et opplegg for variansanalyse. Vi definerer dummyvariable for gruppene som angitt i forrige avsnitt, alt etter om vi tror å ha samspill eller ikke. Hvis vi ikke har samspill, vil virkningen på den avhengige variable bli lik summen av egenvirkningene. Dette omtales ofte som additivitetsforutsetningen. Den avhengige variable kan være en dummy eller en vanlig kvantifiserbar variabel. Vi kan nå regneteknisk behandle dette som et vanlig regresjonsproblem.

Spesielle problem i forbindelse med bruk av dummies som venstresidevariable

Det er spesielt to problem som reiser seg når man har en binær variabel på venstre side i en regresjonsligning.

1) Man kan ikke forutsette homoscedastisitet (dvs. konstant varians på restleddet). Goldberger ({5} side 249 eller {6} side 227) har vist at variansen på restleddet er avhengig av verdien på de høyresidevariable.

Anta at $y = \alpha + \beta x + u$, og at y er en null-en-variabel. For en gitt x-verdi er $u = y - (\alpha + \beta x)$ dvs. enten lik $1 - (\alpha + \beta x)$ eller $- (\alpha + \beta x)$. Sannsynlighetene for at disse begivenheter skal inntreffe er henholdsvis $(\alpha + \beta x)$ og $1 - (\alpha + \beta x)$. Hvis $E(u) = 0$ blir variansen til u:

$$E(u^2) = \left[1 - (\alpha + \beta x)\right]^2 (\alpha + \beta x) + (\alpha + \beta x)^2 \left[1 - (\alpha + \beta x)\right] =$$

$$(\alpha + \beta x)\left[1 - (\alpha + \beta x)\right] = E\left[y(1 - y)\right]$$

Følgelig har vi heteroscedastisitet (dvs. ikkekonstant varians).

Når man vil benytte minste kvadraters metode ved estimeringen av koeffisientene i en regresjonsligning, er det vanlig å forutsette homoscedastisitet. Om man bruker denne metoden når man positivt vet at man har heteroscedastisitet, vil estimatorene ha følgende egenskaper:

a) forventningsretthet

b) konsistens

c) generelt ikke BLUE (best linear unbiased estimators)

Mer alvorlig er det formodentlig når vi skal teste hypoteser om koeffisientene. Malinvaud sier på side 256 i {7}: "Heteroscedasticity also affects tests of hypotheses. Various authors have studied its influence on the analysis of variance and have shown that it seriously affects the significance level and the power of the tests especially when there is great variation in size from one class to another."

Det finnes prosedyrer som reduserer de ubehagelige effekter av heteroscedastisiteten. Her skal nevnes to: Den ene er foreskrevet av Glejser på sidene 316-318 i {4}, den andre av Goldberger på side 250 i {5}. Den første er beregnet for testing og estimering av heteroscedastisitet, den andre er spesielt beregnet for situasjoner med dummies som venstresidevariable. Det som skiller disse metoder er bl.a. at man ved Goldbergers metode vet noe om formen på heteroscedastisiteten (jamfør utledningen på side 22). Glejsers metode er noe mer generell, den starter med å forutsette noe om formen på heteroscedastisiteten (dvs. postulerer en funksjon $\sigma_u^2 = f(x)\sigma^2$). Neste step i begge metoder er estimering av varians-kovarians-matrisen til restleddet. Goldberger gjør det ved å estimere Ey_i på vanlig måte (dvs. est $\{Ey_i\} = \hat{\alpha} + \hat{\beta}x_i$). Vi betegner estimatet $\hat{y}_i$. Varianskovariansmatrisen Ω^* estimeres slik:

$$\Omega^* = \begin{vmatrix} \hat{y}_1(1-\hat{y}_1) & 0 & \cdots & 0 \\ 0 & \hat{y}_2(1-\hat{y}_2) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \hat{y}_n(1-\hat{y}_n) \end{vmatrix}$$

Hos Glejser ville den sett slik ut:

$$\Omega^{**} = \sigma^2 \begin{bmatrix} \widehat{f(x_1)} & 0 & \cdots & 0 \\ 0 & \widehat{f(x_2)} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \widehat{f(x_n)} \end{bmatrix}$$

En mulig test her er $\widehat{f(x_i)} = 1$ for alle i

Deretter transformeres de variable med den estimerte varianskovariansmatrisen. De variable ved observasjon nr. i divideres med kvadratroten av i'te diagonaleldd i Ω^{est}. Restleddet i den nye regresjonsligningen er konstant. Endelig estimeres koeffisientene ved hjelp av minstekvadraters metode på de transformerte variable.

2) Når den avhengige variable er dikotomisk (y er lik én hvis kjennetegnet forekommer og lik null hvis ikke) kan $y_i = \alpha + \beta x_i + u_i$ sies å være av typen "linear probability function". Dette fordi $\hat{y}_i$ kan tolkes som estimat for en betinget sannsynlighet. Denne utformingen av modellen tillater $\hat{y}_i$ å falle utenfor intervallet $[0,1]$, noe som ikke er i overensstemmelse med definisjonen av en sannsynlighet. Jeg skal kort nevne noen metoder for å unngå slike vanskeligheter:

A) "The Probit Analysis Model"

Denne modellen har sin opprinnelse i biologiske undersøkelser av levende organismers svar på stimuli av varierende styrke. Modellen er slik at dersom forklaringsvariablene overstiger visse kritiske verdier, så vil visse kjennetegn framtre ved det fenomen man vil undersøke.

La f.eks. $X = G(Y)$, og la videre $G(Y) = \begin{cases} 0 \text{ hvis } Y < Y_0 \\ 1 \text{ hvis } Y \geq Y_0 \end{cases}$

Y kan igjen være en indeks som er en lineær funksjon (regresjon) av forklaringsvariablene. Ytterligere stokastiske element kan trekkes inn ved å innføre en sannsynlighetsfunksjon for Y_0. Se Goldberger {5} side 250.

B) "The Linear Logit Model"

Anta at vi er interessert i å estimere sannsynligheten for at et kjennetegn skal forekomme. En alternativ spesifikasjon til den lineære sannsynlighetsfunksjonen kan være[1]:

$$\log\left(\frac{p}{1-p}\right) = a + b\log x + w$$

$\log\left(\frac{p}{1-p}\right)$ som kalles "the logit" korresponderende til p, kan variere fritt, og det vil alltid være slik at $0 \leq p \leq 1$. Nå er p ikke observerbar, her settes derfor inn diverse gjennomsnitt av observerte y-verdier. Hvis man har dummygrupperinger blir framgangsmåten ved estimeringen slik: Man deler materialet inn i grupper som er homogene med hensyn på de høyresidevariable (f.eks. ved hjelp av en tabell) og regner ut $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i$ for hver gruppe. n_j er antall observasjoner i gruppe nr. j, og nummereringen i starter på 1 i hver gruppe.

Modellen blir da: $\log\left(\frac{\bar{y}_j}{1-\bar{y}_j}\right) = a + b\overline{\log x_j} + \bar{w}_j$ hvor $\overline{\log x_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \log x_i$ og $\bar{w}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} w_i$. a og b estimeres ved hjelp av minstekvadraters metode og vi

---

1) NB! Denne modellen kan ikke avledes av den lineære sannsynlighetsfunksjonen.

får da $\text{est}\{\log(\frac{\bar{y}_j}{1-\bar{y}_j})\} = \hat{a} + \hat{b}\ \overline{\log x}_j$.

$\log(\frac{\hat{p}_j}{1-\hat{p}_j})$ settes deretter lik $\text{est}\{\log(\frac{\bar{y}_j}{1-\bar{y}_j})\}$. Se forøvrig Theil {11} kap. 3.4 og 3.5.

C) Korrigering etter kjøring

Denne metoden er foreslått av Orcutt, Greenberger, Korbel og Rivlin på side 250 i {8}. Prosedyren er enkel og går ut på følgende: Man kjører først gjennom med minstekvadraters metode på den lineære modellen. Deretter klassifiseres enhetene i grupper som er homogene med hensyn på de høyresidevariable (og m.h.p. estimatet $\hat{y}_j$) på samme måte som under avsnitt B). Deretter beregnes gjennomsnittet av $(y_i - \hat{y}_j)$ for hver slik gruppe. Dette gjennomsnittet betegner vi $F(\hat{y}_j)$ (i er nummerert fra 1 til n_j i hver gruppe). Deretter beregnes det endelige estimatet slik: $\tilde{y}_j = \hat{y}_j + F(\hat{y}_j)$. Anta f.eks. at $\hat{y}_j > 1$, da må vi få $y_i - \hat{y}_j < 0$ og $F(\hat{y}_j) < 0$. Likeledes hvis $\hat{y}_j < 0$ må vi få $F(\hat{y}_j) > 1$. Man innser lett at $\tilde{y}_j$ er en forventningsrett og konsistent estimator.

Siden har man gjort meg oppmerksom på at $\tilde{y}_j = \hat{y}_j + \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i - \hat{y}_j) = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i = \bar{y}_j$.

Altså et resultat man ikke trenger regresjon for å komme fram til.

D) Dummygruppering med samspill

Anta at vi har klassifisert våre høyresidevariable langs flere grupperingsveier. Desto mer fullstendig vi fyller ut for mulige samspillseffekter, jo mindre trolig er det at vi vil få problemer med at estimatet faller utenfor intervallet $[0,1]$. Det bør være nok å ta med 2. og 3. ordens kryssprodukt. En vanskelighet som melder seg om man vil kjøre med samspill er at antall variable blir snart svært stort.

E) Beskrankninger på koeffisientene

Nok en mulighet er å legge beskrankninger på koeffisientene under estimeringen. Dette medfører nokså vanskelige problemer for tolkingen av estimatorene.

REFERANSER

{1} Clopper Almon jr.: "Matrix Methods in Economics" 1967

{2} H.T. Amundsen: "Innføring i teoretisk statistikk" Hefte III Oslo 1963

{3} Forbruksundersøkelse 1967 hefte II SSB Oslo 1969

{4} H. Glejser: "A New Test for Heteroskedasticity" Journal of the American Statistical Association. March 1969

{5} Goldberger: "Econometric Theory" N.Y. 1964

{6} Johnston: "Econometric Methods" London 1963

{7} Malinvaud: "Statistical Methods of Econometrics" Amst. 1968

{8} Orcutt, Greenberger, Korbel & Rivlin: "Microanalysis of Socioeconomic Systems. A Simulation Study". N.Y. 1961

{9} Pyatt: "Priority Patterns and the Demand for Household Durable Goods," Cambridge 1964

{10} Suits: "Use of Dummy Variables in Regression Equations" Journal of the American Statistical Association. Dec. 1957

{11} Theil: "Economics and Information Theory" Amst. 1967