RAPPORTER 84/14

REGRESJONSANALYSE MED ET STORT ANTALL VARIABLE

AV
ERLING SIRING OG EMIL SPJØTVOLL

STATISTISK SENTRALBYRÅ
CENTRAL BUREAU OF STATISTICS OF NORWAY

RAPPORTER FRA STATISTISK SENTRALBYRÅ 84/14

REGRESJONSANALYSE MED ET STORT ANTALL VARIABLE

AV
ERLING SIRING OG EMIL SPJØTVOLL

STATISTISK SENTRALBYRÅ
OSLO — KONGSVINGER 1984

ISBN 82-537-2122-6
ISSN 0332-8422

EMNEGRUPPE
Teori og metode

ANDRE EMNEORD
Teoretisk statistikk

SUMMARY

Regression analysis with a large number of explanatory variables is discussed. The main problem considered is the choice of a relevant subset of variables and a subset of cross products of these. It is proposed to proceed in three stages by first selecting the variables to be used in the regression function. This is made by a combination of standard stepwise methods and P-plots to determine the number of relevant variables. Next, the cross products of the variables carried on from the first stage is also treated by this technique. Finally, the selected linear terms and cross products are put together in one analysis to eliminate possible superfluous elements. The techniques are demonstrated on a real example with 36 initial explanatory variables.

FORORD

I regresjonsanalyse med et stort antall forklaringsvariable er det ofte et problem å avgjøre hvor mange og hvilke variable som bør være med i regresjonsfunksjonen. Størrelsesordenen av problemet øker hvis en vurderer også å ta med kryssproduktledd. Metoder for å behandle slike situasjoner er beskrevet i den foreliggende rapport. Det blir også presentert teknikker for å finne fram til enkeltobservasjoner som har stor innflytelse på resultatet av regresjonsanalysen.

Statistisk Sentralbyrå, Oslo, 29. oktober 1984

Arne Øien

INNHOLD

1. INNLEDNING

Regresjonsanalyse er en standard metode for å beskrive hvorledes én variabel -den avhengige variabel - varierer som funksjon av én eller flere andre variable -de såkalte uavhengige variable. Den brukes ofte i situasjoner hvor en har liten forhåndskunnskap om sammenhengene. En konsekvens av dette er at den statistiske metode som brukes blir svært avgjørende for resultatet av analysen.

Særlig er problemene store når det er svært mange uavhengige variable. Det gjelder da å avgjøre hvor mange og hvilke variable som skal brukes i regresjonsfunksjonen. Vanligvis gjøres dette ved trinnvise regresjonsmetoder. I dette arbeidet blir disse supplert med plottemetoder som gir et anslag for antall forklaringsvariable som bør være med.

Størrelsesordenen slike problemer kan ha, illustreres ved gjennomgangseksemplet som blir brukt i denne rapporten. Der er det opprinnelig 36 forklaringsvariable. Ved i tillegg å prøve alle kryssprodukt mellom disse kommer en opp i en regresjon med 666 mulige uavhengige variable.

Selv med en regresjonsfunksjon hvor mange variable er tatt med, kan en ofte være i tvil om en har fått så god tilpasning som det er mulig med de variable som er til rådighet. En slik vurdering kan gjøres ved å sammenligne restvariansen ved en såkalt intern estimering av feil.

Enkeltobservasjoner kan i visse tilfelle ha stor innflytelse på resultatet av regresjonsanalysen. Det gjelder å lokalisere slike observasjoner for å se nærmere på dem. Skyldes de feil eller er de ikke representative for resten av materialet eller er de bare en del av de tilfeldige variasjoner en må regne med? En kan finne fram til slike observasjoner ved hjelp av ulike mål for innflytelsen av enkeltobservasjoner.

2. PROBLEMSTILLING OG OPPSUMMERING

I denne artikkelen tas utgangspunkt i regresjonsmodellen

$$(1)\quad y_i = \beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip} + e_i \qquad i = 1,\ldots, n$$

hvor y er den avhengige variabel og $x_1, \ldots, x_p$ er p uavhengige variable eller forklaringsvariable. Tallet n er antall observasjoner. Leddet e representerer avviket mellom y og den beste beskrivelse av y med en lineær funksjon av x-ene. For at tester og konfidensintervall skal være tilnærmet gyldige bør e_i-ene være uavhengige, tilnærmet normalfordelte, ha forventning nær 0 og ha tilnærmet konstant varians. Variansen σ^2 antas ukjent. De ukjente koeffisientene $\beta_0, \beta_1, \ldots, \beta_p$ vil bli kalt førstegradskoeffisientene eller også av og til hovedeffektene.

Det antas at en starter med et stort antall p av x-er, men regner med at bare en del av disse bidrar vesentlig til å beskrive y. Det gjelder å finne fram til hvilke. Modellen kan også utvides ved å ta med kryssproduktledd. Videre må vurderes om en har fått så god modelltilpasning at den ikke kan forbedres ved hjelp av de foreliggende forklaringsvariable.

I kapittel 3 beskrives et datamateriale som blir brukt som gjennomgangseksempel for å illustrere de statistiske metodene. Ved hjelp av intern estimering av feil vises i kapittel 4 at det skulle være mulig å forbedre en tidligere brukt modell. For å få modellforutsetningene bedre oppfylt, foreslås i det påfølgende kapittel en logaritmetransformasjon av den avhengige variable. Det foretas også en gruppering innen de uavhengige variable.

Kapittel 6 behandler metoder for å finne fram til de mest betydningsfulle variable. Trinnvise metoder kombineres med et plott av P-verdiene (signifikannssannsynlighetene) knyttet til de enkelte regresjonskoeffisienter. Dette plottet gjør det mulig å anslå antall variable som bør være med.

I kapittel 7 beskrives diagnostiske mål for å finne frem til enkeltobservasjoner som kan ha stor innflytelse på resultatet av regresjonsanalysen. Mål for kolinearitet mellom forklaringsvariablene tas også med.

Modellen (1) utvides i kapittel 8 til å ta med kryssproduktledd mellom de variable som ble funnet betydningsfulle i kapittel 6. Modellen blir da

$$(2)\quad y = \beta_0 + \beta_1 Z_1 + \ldots + \beta_q Z_q + \beta_{12} Z_1 Z_2 + \ldots + \beta_{q-1,q} Z_{q-1} Z_q + e$$

hvor $Z_1, \ldots, Z_q$ er de variable som ble tatt med fra modellen (1) med bare førstegradsledd. Størrelsene $\beta_{12}, \ldots, \beta_{q-1,\,q}$ kaller vi kryssproduktkoeffisientene eller samspillene. Ved hjelp av P-plott estimeres igjen antall betydningsfulle samspill. Endelig i kapittel 9 foretas en totalvurdering av modellen med både førstegradsledd og kryssproduktledd.

I siste kapittel prøves metoder for å finne frem til innflytelsesrike observasjoner på dataene i gjennomgangseksemplet. Med så stort datamateriale som det her er tale om, ser det ut til at enkeltobservasjoner kan ha relativt liten innflytelse på det endelige resultat.

3. GJENNOMGANGSEKSEMPLET

For å illustrere de statistiske metoder er brukt et gjennomgangseksempel med data fra Ferieundersøkelsen 1974. Materialet har tidligere vært analysert av Mordal (1979) og Haldorsen (1981). Formålet med å bruke dette datamaterialet er ikke å komme lenger når det gjelder å analysere folks ferievaner, men å bruke det som et middel til å eksemplifisere de statistiske analysemetoder. Opprinnelig ble dette materialet også valgt fordi det inneholdt mange dikotome variable. Metodene som utvikles er imidlertid lite avhengig av dette.

Selv om vårt hovedformål ikke er å finne frem til nye konklusjoner når det gjelder Ferieundersøkelsen, må vi likevel for sammenhengens skyld gi en kort beskrivelse av datamaterialet. Det består av opplysninger fra yrkesaktive personer som var på minst én ferietur i 1974. Totalt inngår n = 980 personer. Som yrkesaktive er regnet personer med minst halvparten av vanlig arbeidstid i et yrke.

Ferietur er definert som opphold utenfor helårsboligen med helse- eller rekreasjonsformål, som inkluderer minst 4 overnattinger. Forretnings-/studiereiser og rekonvalesensopphold på sykehus, syke-/pleiehjem e.l. er ikke regnet som ferietur. "Helsereiser" (invalidereiser) o.l. som er lagt opp som ferietilbud, er derimot regnet som ferietur.

Den avhengige variabel er antall feriedager i løpet av et år. I alt 34 forklaringsvariable skal prøves. Disse er satt opp i vedlegg 1. Merk at den opprinnelige nummering av de variable er beholdt slik at den ikke går fra x_1 til x_{34}, men fra x_2 til x_{38} hvor x_8, x_9 og x_{19} ikke finnes.

I det følgende vil dette datamaterialet bli referert til som Eksemplet.

4. INTERN ESTIMERING AV FEIL

4.1. Den generelle tankegang

Når en regresjonsfunksjon er tilpasset et datamateriale, vil en være interessert i å vite om den gir så god tilpasning som mulig med de forklaringsvariable en har til rådighet. Kan en være sikker på at en annen regresjonsfunksjon hvor en kombinerer de forklaringsvariable på andre måter eller trekker inn andre transformasjoner av dem, ikke vil gi vesentlig bedre tilpasning? Hvis en har gjentatte observasjoner for visse variabelkombinasjoner, kan de brukes til å lage et variansestimat som er fri for eventuell modellfeil. Dette kan da sammenlignes med restvariansen fra den tilpassete modell. Hvis disse er noenlunde like, viser det at en ikke kan oppnå vesentlig bedre tilpasning. Er de svært forskjellige, tyder det på at modellen kan forbedres.

I ikke-eksperimentelle data har en vanligvis ikke gjentatte observasjoner med samme verdier på forklaringsvariablene. Daniel og Wood (1980) foreslår at en erstatter feilestimatet fra gjentatte observasjoner med et feilestimat fra "nære naboer". Med nære naboer menes observasjoner som ligger nær hverandre i rommet av forklaringsvarablene, dvs. at de har nesten samme verdier på disse. Mer presist, hvis den tilpassede regresjonsfunksjon er

$$\hat{r}(x) = b_o + \sum_{j=1}^{p} b_j x_j \quad ,$$

så måles avstanden mellom to punkter

$$x_i = (x_{i1} \ldots, x_{ip}) \qquad i = 1,2$$

ved

$$D_{12}^2 = \sum_{j=1}^{p} [b_j (x_{1j} - x_{2j})]^2$$

En vil regne med at to punkter som ligger nær hverandre, også vil ha y-verdier som ligger nær hverandre. Altså at de kan betraktes som nesten gjentatte observasjoner. Daniel og Wood (1980) bruker differenser av typen

$$y_1 - \hat{r}(x_1) - (y_2 - \hat{r}(x_2))$$

som utgangspunkt for estimering av feil fra nære naboer. De regner med at absoluttverdien av denne differensen har forventning tilnærmet forventningen til den absolutte differens mellom to uavhengige normale variable med samme forventning og varians lik den ideelle restvarians σ^2. Det antall nære naboer som skal inngå i feilestimeringen blir delvis en skjønnsssak, men når estimatet begynner å øke, tyder det på at en har tatt med for mange.

4.2. Modifisering for tilfellet med mange dikotome variable

Ved mange betydningsfulle dikotome forklaringsvariable kan en gå noe lengre enn ved fremgangsmåten til Daniel og Wood når det gjelder å gjøre seg uavhengig av modellforutsetninger. Anta at en ved en foreløpig analyse har funnet frem til en gruppe av viktige forklaringsvariable hvorav d er dikotome. Det er da 2^d mulige kombinasjoner av disse. Anta at det i en del av de tilsvarende 2^d grupper av observasjoner er nok observsjoner til å regne ut regresjon med hensyn til de resterende variable. I hver av disse delgruppene kan en så lage seg et internt estimat av feil basert på nære naboer. Disse feilestimatene vil da ikke være påvirket av modellspesifikasjonsfeil som angår de d dikotome variable. For eksempel vil alle samspill mellom disse variablene innbyrdes og med de resterende variable være eliminert fra feilestimatene.

Vi har gjennomført en slik fremgangsmåte for Eksemplet. Som foreløpig analyse er brukt de resultatene Haldorsen (1981) kom fram til. Ved trinnvis regresjon fant han frem til følgende variable etter 9 trinn (i rekkefølge i forlengs utvelgelse)

$$x_{35}, x_{33}, x_2, x_6, x_{23}, x_{17}, x_{12}, x_{36}, x_{10}.$$

Av disse er

$$x_{35}, x_{33}, x_2, x_{23}, x_{17}$$

dikotome. Vi delte da opp materialet i $2^5 = 32$ grupper etter verdiene på disse d = 5 dikotome variable. I stedet for å lage variansestimater fra næreste naboer i disse gruppene, nøydde vi oss med enklere prosedyrer. Den ene var å beregne variansestimatene etter full lineær regresjon på alle resterende 29 variable innenfor hver enkelt av de 32 gruppene. Den andre besto i å kjøre variansanalyse med hensyn på de viktige variablene x_6, x_{12}, x_{36} og x_{10} innenfor gruppene. Det siste for å prøve modeller som var uavhengig av spesielle funksjonsformer for disse 4 kvantitative variable. Bare hovedeffekter ble tatt med i variansanalysen. Resultatene er gitt i tabell 1. Den variable HJELP definert ved

$$HJELP = x_{23} + 2x_{17} + 4x_2 + 8x_{33} + 16x_{35}$$

kan brukes til å identifisere delgruppene. Gruppene med et lite antall observasjoner er ikke tatt med. Tomme plasser i tabellen svarer til situasjoner med for få observasjoner for å få utført variansanalysen eller den fulle regresjonen.

Det gjennomsnittlige variansestimat på grunnlag av kolonnen for full regresjon er 97,98. Den trinnvise regresjonen med 9 variable ga variansestimatet 139,92. Dette tyder på at det er muligheter for vesentlig forbedring av modellen.

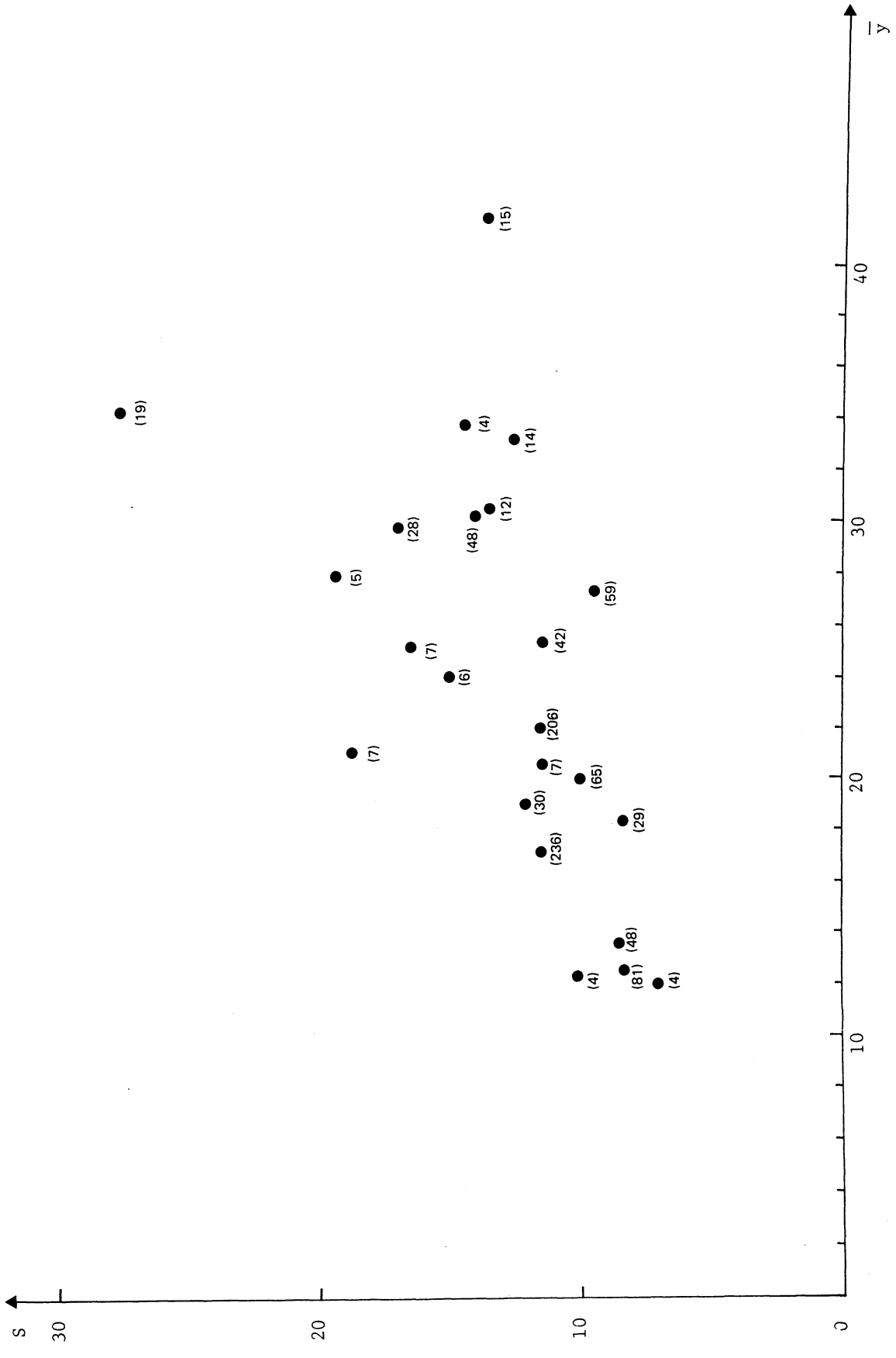
Tabell 1. Variansen for antall ferieturdager innen delgrupper

HJELP	Varians innen grupper	Variansanalyse m.h.p. x6,x12, x36 og x10	Full regresjon på 29 variable	Antall obs.
0	68,6	54,4	65,8	81
4	74,5	81,0	92,1	48
8	129,0	113,1	93,0	236
9	72,6			29
10	100,0	86,8	83,8	65
12	132,7	116,1	112,2	206
13	775,0			19
14	84,9	92,3	81,9	59
24	146,3	83,6		30
25	289,3	282,6		28
27	179,5			12
28	130,8	80,2		42
29	196,2	211,8	157,9	48
30	156,1			14
31	189,5			15

4.3. Varians i delgrupper

Oppdelingen i delgrupper har også den fordel at en kan få et inntrykk av om variansen er konstant i ulike områder av x-rommet. Tallene i tabell 1 kan tyde på at det er en viss variasjon mellom gruppene utover det som kan skyldes tilfeldigheter.

For å undersøke om variansene har en tendens til å øke med forventet verdi av den avhengige variable er estimert standardavvik i gruppene (kolonne 1 i tabell 1) plottet mot gruppegjennomsnittene i figur 1. Det ser ut til at standardavviket øker tilnærmet lineært. Dette tyder på at en burde transformere dataene for å få en mer stabil varians. Dette er emne for neste kapittel.



Figur 1. Standardavvik, S, til antall ferieturdager etter størrelsen av gjennomsnittet, $\bar{y}$, i forskjellige grupper. Tallene i parentes angir antall observasjoner

5. TRANSFORMASJONER

5.1. Transformasjoner av den avhengige variable

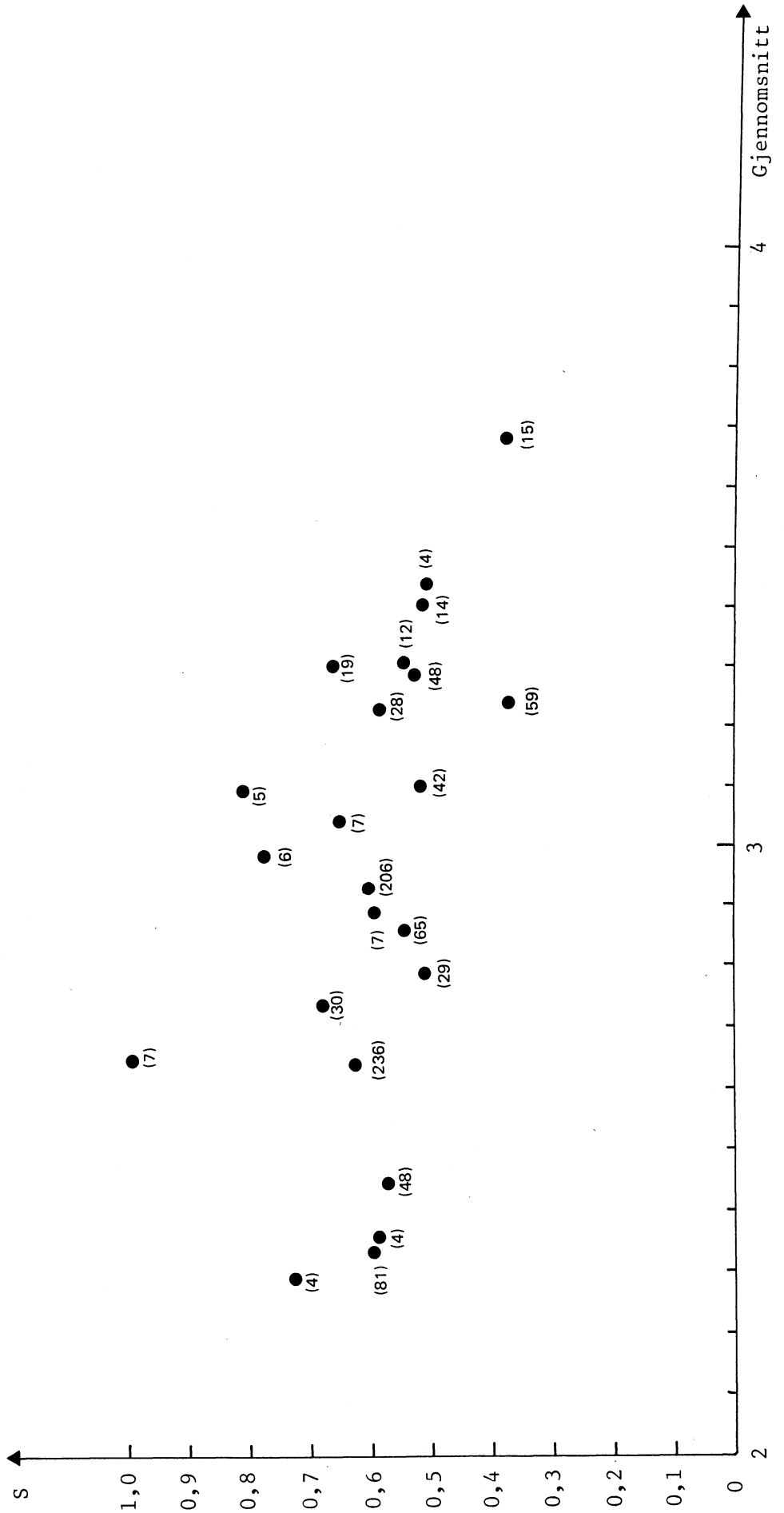
5.1.1. Logaritmetransformasjon

For å estimere regresjonskoeffisientene vil minste kvadraters metode bli brukt. Da er det en fordel at variansene er noenlunde like i hele variasjonsområdet for observasjonene. Figur 1 i forrige kapittel viser at det er en tendens til at standardavviket øker lineært med forventningen. I et slikt tilfelle kan en få mer stabile varianser ved å ta logaritmen til observasjonene. Tabell 2 viser resultatet av tilsvarende beregninger som i kapittel 1, men med utgangspunkt i logaritmen til antall feriedager. Variansen ser nå ut til å være mer stabile, iallfall for de største gruppene. Plott av standardavvik mot gruppegjennomsnittet er gitt i figur 2. Nå kan det synes som det er en tendens til at standardavviket avtar med gruppegjennomsnittet. Dette inntrykket blir imidlertid skapt av noen få punkter med lite antall frihetsgrader, punkter som ikke bør tillegges så stor vekt.

Tabell 2. Variansen for logaritmen til antall feriedager innen delgrupper

HJELP	Totalt	Variansanalyse m.h.p. x6,x12, x36 og x10	Full regresjon på 29 variable	Antall obs.
0	0,353	0,301	0,336	81
4	0,330	0,322	0,417	48
8	0,395	0,374	0,295	236
9	0,257			29
10	0,298	0,260	0,280	65
12	0,362	0,324	0,306	206
13	0,439			19
14	0,140	0,143	0,133	59
24	0,463	0,334		30
25	0,334	0,279		28
27	0,296			12
28	0,263	0,179		42
29	0,278	0,259	0,192	48
30	0,265			14
31	0,145			15

Figur 2. Standardavvik (S) til logaritmen til antall ferieturdager etter størrelsen av gjennomsnittet i forskjellige grupper. Tallene i parentes angir antall observasjoner



I tillegg til like varianser er det også en fordel om observasjonene er tilnærmet normalfordelte rundt sine forventninger. To mål for avvik fra normalitet er gitt gjennom skjevhet og kurtosis. For en serie observasjoner $Y_1, Y_2, \ldots, Y_n$ med

$$\bar{Y} = \frac{1}{n} \Sigma Y_i \ , \ S^2 = \frac{1}{n} \Sigma (Y_i - \bar{Y})^2$$

er målet for skjevhet

$$\frac{1}{n} \Sigma (Y_i - \bar{Y})^3 / S^2$$

og for kurtosis

$$\frac{1}{n} \Sigma (Y_i - \bar{Y})^4 / S^4 - 3$$

For en normalfordeling er de teoretiske verdier for begge størrelser lik 0.

I tabell 3 er gitt de beregnete verdier av skjevhet og kurtosis i de enkelte gruppene. Dette er gjort både for de opprinnelige data og for logaritmen til dem. Stort sett ligger verdiene nærmere 0 for de transformerte tallene. Spesielt gjelder dette de store gruppene hvor estimatene er mest pålitelige. I tillegg til plottene over standardavvikene styrker dette den oppfatning at en ved regresjonsberegningene bør basere seg på logaritmen til antall ferieturdager.

Tabell 3. Skjevhet og kurtosis i forskjellige grupper

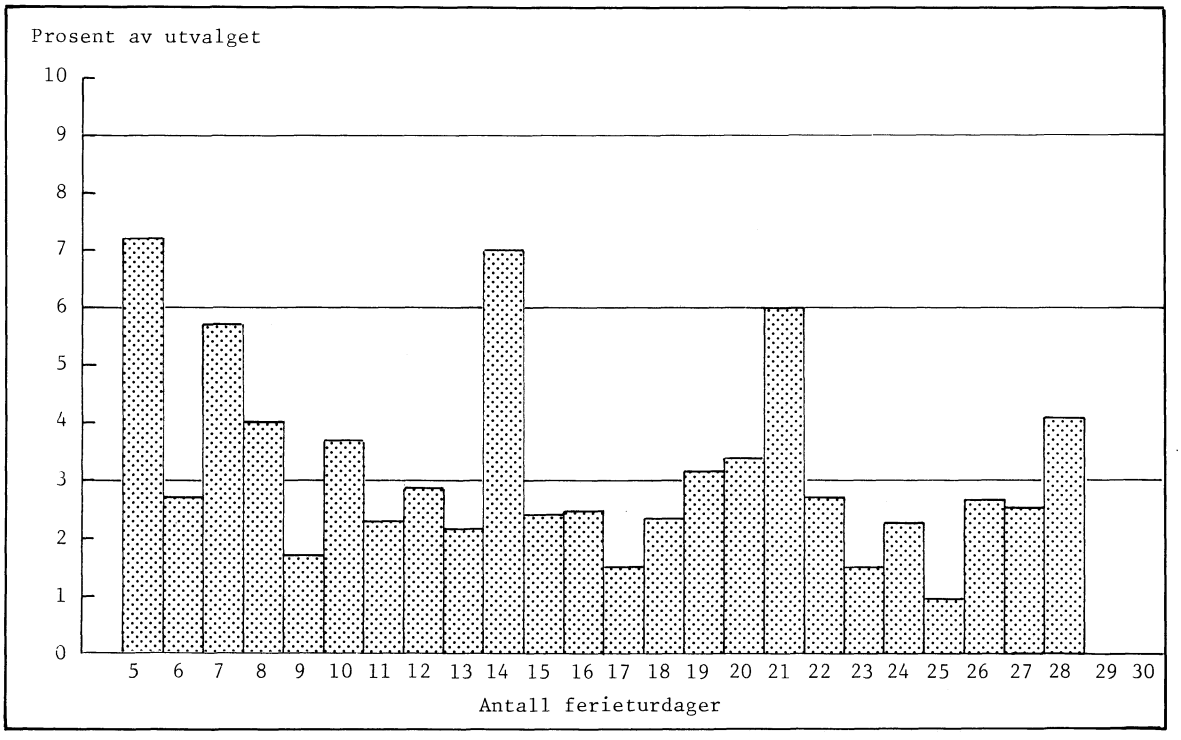
	Antall ferieturdager		Logaritmen til antall ferieturdager		
HJELP	Skjevhet	Kurtosis	Skjevhet	Kurtosis	Ant. obs.
0	1,7	3,5	0,4	-0,8	81
4	1,6	3,5	0,4	-0,8	48
8	2,0	7,2	0,0	-0,6	236
9	0,4	-0,6	-0,3	-0,8	29
10	0,6	-0,11	-0,4	-0,5	65
12	0,7	0,67	-0,6	-0,2	206
13	2,4	6,1	0,4	1,2	19
14	0,2	-0,17	-0,7	0,1	59
24	1,3	2,4	-0,2	-0,8	30
25	1,1	0,5	-0,1	-0,2	28
27	0,3	0,3	-1,3	2,8	12
28	0,8	1,6	0,8	-0,8	42
29	0,5	-0,2	-0,7	0,3	48

5.1.2. Gruppering av den avhengige variable

For å få en oversikt over et datamateriale kan det i første omgang være nyttig å lage histogrammer over de enkelte variable. I figur 3 er tegnet et histogram over fordeling på antall feriedager for personene i utvalget. Av plasshensyn er tatt med de med 28 eller færre ferieturdager. Totalt er det 235 personer av utvalgets 980 med flere enn 28 ferieturdager.

Histogrammet viser en flertoppet fordeling med topper på 5, 7, 14, 21 og 28 feriedager. Grunnen til toppen på 5 dager er antagelig at det er det minste antall som blir regnet med i denne undersøkelsen. Ellers er det altså en tendens til at når folk reiser på ferietur, blir de borte et helt antall uker.

Figur 3. Histogram over utvalgets fordeling på antall ferieturdager

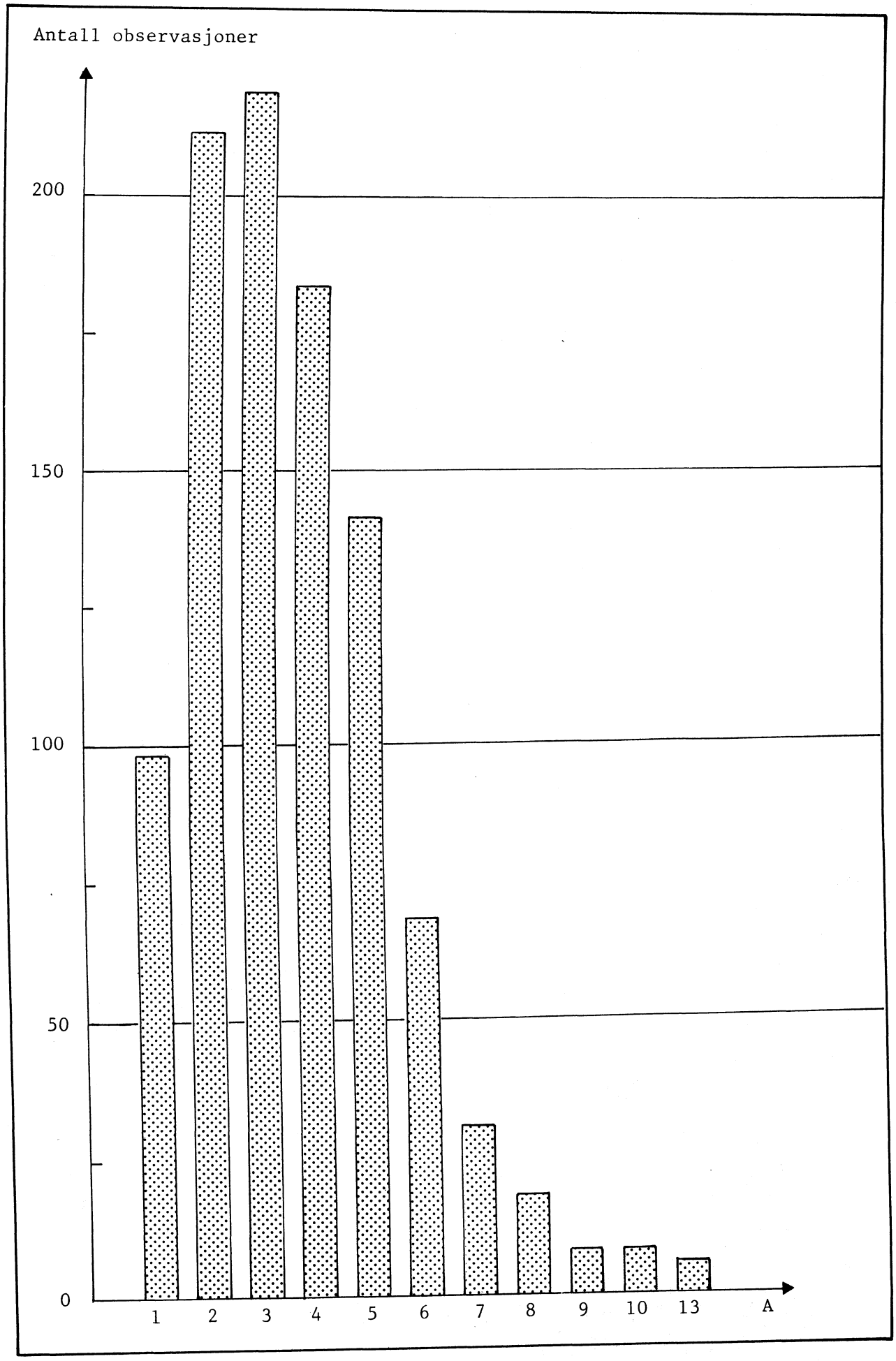


Det kunne derfor være rimelig å ta utgangspunkt i antall ferieuker i stedet for antall feriedager. Vi har derfor sett på en variabel A definert på følgende måte.

A = 1	dersom	5	≤	ant.	ferieturdager	≤	6
A = 2	"	7	≤	"	"	"	≤	13
A = 3	"	14	≤	"	"	"	≤	20
A = 4	"	21	≤	"	"	"	≤	27
A = 5	"	28	≤	"	"	"	≤	34
A = 6	"	35	≤	"	"	"	≤	41
A = 7	"	42	≤	"	"	"	≤	48
A = 8	"	49	≤	"	"	"	≤	55
A = 9	"	56	≤	"	"	"	≤	62
A = 10	"	63	≤	"	"	"	≤	69
A = 13	"	70	≤					

Histogrammet av A i figur 4 viser en entoppet fordeling. Det kunne hende at en ville få klarere sammenhenger og konklusjoner ved å basere analysen på antall uker. Vi har likevel valgt - delvis for lettere å kunne sammenligne med tidligere resultater - å ta utgangspunkt i antall ferieturdager.

Figur 4. Histogram over antall ferieturdager regnet i uker



5.2. Transformasjoner av de uavhengige variable

I sin alminnelighet kan det være aktuelt å transformere også de uavhengige variable. Ved hensiktsmessige transformasjoner kan en få frem variable som mest mulig direkte beskriver forventningen av den avhengige variable. Valg av transformasjoner vil som regel være basert på den apriori kunnskap en har om sammenhengene.

De fleste av variablene i Eksemplet er dikotome. Disse er det ikke noe behov for å transformere. Når det gjelder de resterende variable, har vi valgt å gruppere og dikotomisere også disse. Det er gjort på følgende måte:

$$x_{601} = \begin{cases} 0 \text{ hvis } x_6 = 1 \\ 1 \text{ hvis } 2 \leq x_6 \leq 7 \end{cases}$$

$$x_{121} = \begin{cases} 1 \text{ hvis } x_{12} = 1 \\ 0 \text{ ellers} \end{cases}$$

$$x_{126} = \begin{cases} 1 \text{ hvis } x_{12} = 6 \\ 0 \text{ ellers} \end{cases}$$

$$x_{361} = \begin{cases} 0 \text{ hvis } 4 \leq x_{36} \leq 6 \\ 1 \text{ hvis } 7 \leq x_{36} \leq 12 \end{cases}$$

$$x_{371} = \begin{cases} 0 \text{ hvis } 2 \leq x_{37} \leq 3 \\ 1 \text{ hvis } 4 \leq x_{37} \leq 6 \end{cases}$$

$$x_{501} = \begin{cases} 1 \text{ hvis } 1 \leq x_5 \leq 2 \\ 0 \text{ ellers} \end{cases}$$

$$x_{506} = \begin{cases} 1 \text{ hvis } 6 \leq x_5 \leq 7 \\ 0 \text{ ellers} \end{cases}$$

Opprinnelig var det to motiver for å innføre disse nye variablene. Det ene var at vi hadde tenkt å utvikle et metodeverktøy for regresjon med bare dikotome variable. Etter hvert fant vi ut at vi ikke ville begrense oss bare til denne situasjonen. Det viktigste motivet var likevel at selv om mange av disse variable kan anta flere verdier, har disse karakter av å være indekser. Det er ikke noen grunn til å vente lineær økning (eller minskning) for antall ferieturdager over hele variasjonsområdet til variablene.

Vi har derfor valgt å dele opp hver enkelt av disse variable i 2 eller 3 grupper. Som eksempel se på variabel x_6 som er knyttet til tallet på ganger personen utøvde idretts- og mosjonsaktiviteter i løpet av året. Der har vi delt opp verdiene i to grupper: De som ikke utøvde slike aktiviteter ($x_6 = 1$) og de som gjorde det. Det er lite trolig at en lineær sammenheng her ville gjelde over hele variasjonsområdet for x_6 (fra 1 til 7). Andre variable som x_5 og x_{12} har vi delt i 3 grupper. Denne grupperingen av variable er delvis gjort ut fra skjønn, delvis ved å studere plott av deler av materialet (ikke gjengitt her). En kunne selvfølgelig ha gått enda lengre ved å gjøre som i variasjonsanalysen, hvor en lar hvert mulig nivå av variablene være representert ved en egen effekt. Men dette ville ha ført til mange parametere med små effekter knyttet til hver enkelt.

6. ANTALL BETYDNINGSFULLE FØRSTEGRADSLEDD

6.1. Innledning

For å komme fram til en regresjonsmodell vil bli brukt en strategi som består av flere trinn. Med uavhengige variable $x_1, \ldots, x_p$ tas utgangspunkt i en modell av formen

$$(3) \quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \text{feil}$$

Altså en modell som inneholder bare førstegradsledd (eller hovedeffekter) i forklaringsvariablene. Da en regner med at bare en del av disse bidrar til å forklare den avhengige variable y, gjelder det å bestemme hvor mange og hvilke det er. Tradisjonelt blir dette gjort ved hjelp av ulike trinnvise metoder. I neste avsnitt vil bli gitt en beskrivelse av slike teknikker. Deretter skal vi introdusere to plottemetoder for å behandle dette problemet. Vi vil argumentere for at plottemetoder kan være statistisk mer holdbare enn trinnvise metoder. De kan nemlig gi et estimat av antall signifikante ledd i regresjonen.

I et senere kapittel blir det vist hvorledes en ut fra de effekter som finnes i modellen (3), kan utvide modellen for å ta med eventuelle kryssproduktledd (eller samspill) i de variable.

6.2. Trinnvise metoder

I det følgende gis en kortfattet beskrivelse av 5 forskjellige prosedyrer for trinnvis regresjon. De finnes alle i programpakken SAS.

6.2.1. FORWARD

Denne teknikken starter med bare konstantleddet og tar inn variable én etter én. Prosedyren beregner F-observatoren som reflekterer tilleggsbidraget som hver variabel gir til modellen dersom den tas med. F-ene sammenlignes med et på forhånd spesifisert signifikansnivå. Dersom ingen er større enn den kritiske verdi, stopper prosedyren. Ellers tas variablen med den største F-en inn i modellen. Prosedyren tar så inn variable inntil ingen har signifikant F. En variabel som er tatt inn i modellen, blir værende.

6.2.2. BACKWARD

Denne teknikken begynner med alle de uavhengige variablene i modellen. Deretter blir variablene tatt ut av modellen én etter én inntil alle variablene som er igjen i modellen, er signifikante (m.h.p. F) på et på forhånd spesifisert signifikansnivå. På hvert trinn blir variablen med den minste F-en utelatt.

6.2.3. STEPWISE

Denne metoden er en modifisert utgave av forlengsmetoden (FORWARD). Etter at en ny variabel er tatt inn i modellen, blir alle variablene som er i modellen fra før vurdert. Variable som ikke er signifikante (m.h.p. F) blir utelatt. Deretter vurderer prosedyren nye variable. Når alle variable i modellen er signifikante og ingen utenfor er signifikante, stopper prosedyren.

6.2.4. MAXR

Denne metoden prøver å finne den beste én- variabel modellen, den beste to- variable modellen osv. Best er definert som modellen med størst multippel korrelasjonskoeffisient R. Den er imidlertid ikke garantert å finne modellen med den største R på hvert trinn. Teknikken er en forlengsprosedyre. Den starter med den variabel som gir størst R. Så tar den inn variabelen som gir størst økning i R. Deretter bytter den ut én og én variabel i modellen med alle variablene som ikke er i modellen, og vurderer om det kan oppnås høyere R. Slik fortsetter den på alle trinn.

6.2.5. MINR

Denne teknikken er en baklengsvariant av MAXR. Den starter med modellen hvor alle variable er med, og fjerner variable én etter én, analogt til at MAXR tar inn variable etter tur. I prinsippet skulle MINR og MAXR gi samme resultat, nemlig modellene med størst R for gitte antall variable. Men i og med at ikke alle delmodeller er undersøkt, er en ikke garantert dette. En ulempe med MINR er at den produserer mye utskrifter, noe som vanskeliggjør oversikten.

6.2.6. Sammenligning

Ved kjøringer på Eksemplet har vi stort sett brukt BACKWARDS og MAXR. For samme antall variable i regresjonsfunksjonen gav disse to teknikkene samme modell i mange tilfelle. På trinn der det var uoverenstemmelser mellom de to prosedyrene, dreidde dette seg oftest om én variabel. Med mange variable i modellen produserte BACKWARDS-prosedyren i enkelte tilfelle en større R enn det MAXR gjorde. Med få variable i modellen var det MAXR som hadde en tendens til å produsere den største R. I alle tilfellene var det svært liten forskjell på R-verdiene ved de to prosedyrene.

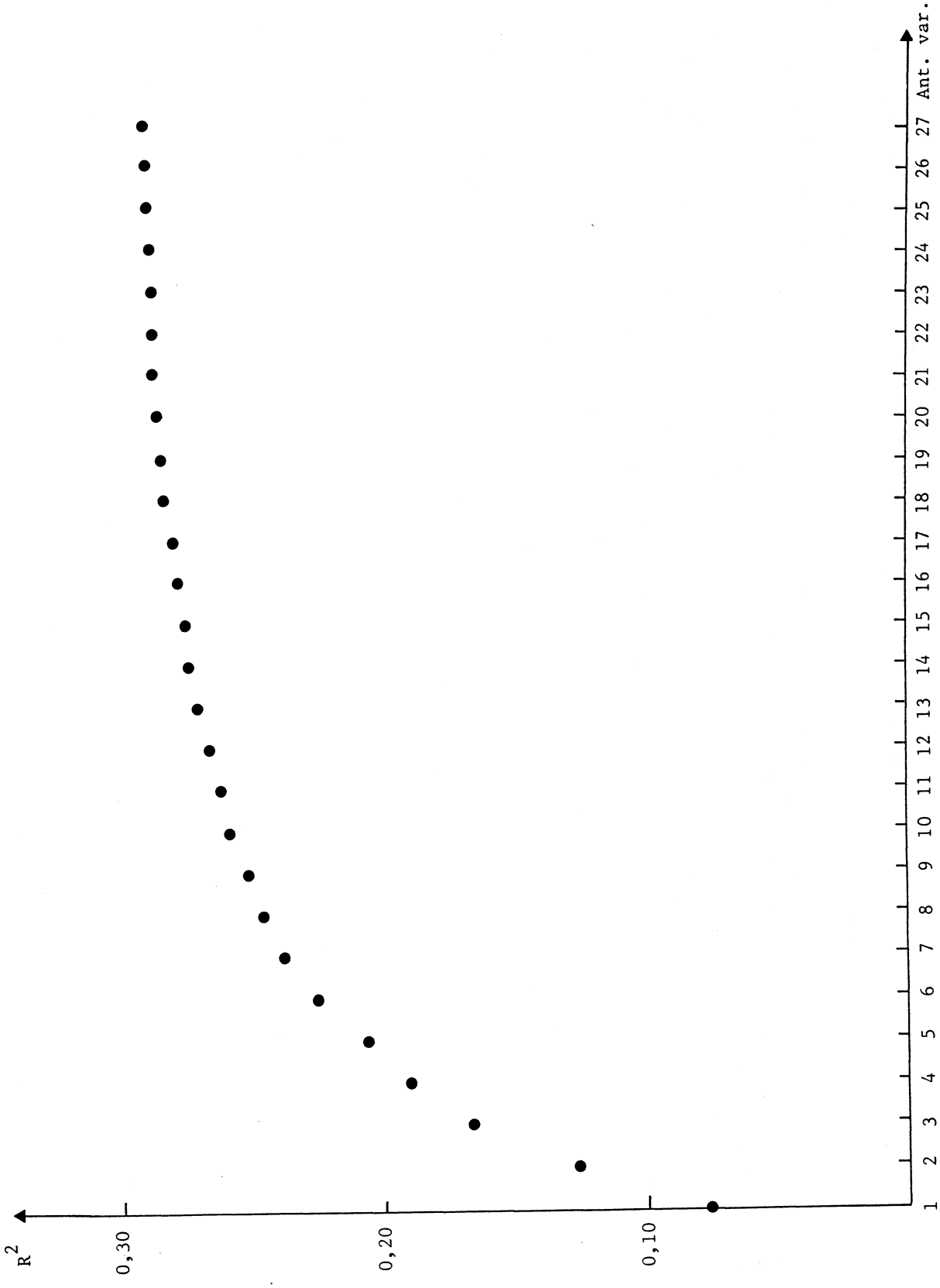
6.3. R^2 -plott

Den kvadrerte multiple korrelasjonskoeffisient R^2 forteller hvor stor del av variasjonen i den avhengige variable som kan beskrives av de uavhengige variable i modellen. Vanligvis vil R^2 som funksjon av antall variable i modellen øke til å begynne med for deretter å flate ut. Punktet hvor den begynner å flate ut eller et punkt hvor det er et knekk i forløpet av funksjonen kan indikere at da har en fått med de viktigste variable. De resterende variable er lite betydningsfulle eller bare har tilfeldig innflytelse.

I figur 5 er R^2 plottet mot antall variable for Eksemplet. Plottet har et relativt jevnt forløp. Det er et lite brudd i forløpet etter 17 variable. Fra 18 variable og ut er det nesten lineært med små endringer i R^2. Også ved 7 variable ser det ut til å være et knekkpunkt. Det er nesten lineære deler før og etter 7 variable.

Vi skal ikke utnytte dette plottet på noe mer formell måte, men bare peke på at det kanskje ligger informasjon i det som burde utnyttes bedre. Som hjelp til dette kan et arbeide av Zirphile (1975) være nyttig. Han studerer de asymptotiske egenskapene til R^2 i tilfellet med ortogonale uavhengige variable.

Figur 5. Trinnvis regresjon (MAXR) på de 36 hovedeffektene. R^2 som funksjon av antall variable i modellen



6.4. Normalplott

I det tilfellet hvor en har mange estimerte regresjonskoeffisienter kan en prøve å plotte disse på normalfordelingspapir for å avsløre hvilke som har forventning forskjellig fra 0. De med forventning 0 bør ligge rundt en rett linje, mens de knyttet til reelle effekter avviker fra denne linjen. Daniel (1959) brukte halvnormalt plott (dvs. plott av absuluttverdiene) for å finne signifikante hovedeffekter og samspill i 2^p forsøk. I det tilfelle er de estimerte effekter uavhengige. De estimerte regresjonskoeffisienter er ikke uavhengige, men i forventning skulle likevel plottet være rettlinjet for de koeffisienter som har forventning 0. Det viktigste for at plottet skal være effektivt, er at en god del av koeffisientene har forventning 0. Dermed får en anslått en rett linje svarende til en normalfordeling rundt 0. Koeffisientene som virkelig bidrar til regresjonen vil en vente avviker fra denne linjen.

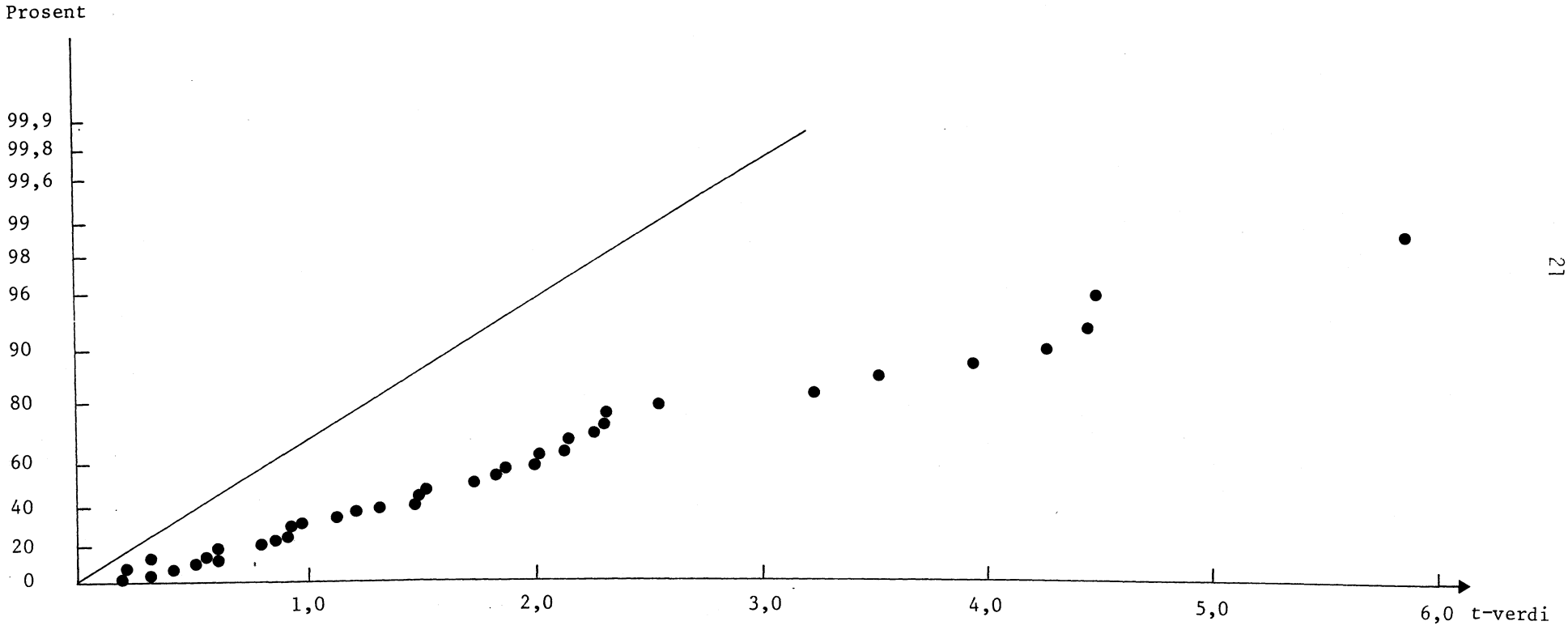
Ut fra den fulle regresjonen på alle 36 variable i vedlegg 2 er det halvnormale plottet for de 36 regresjonskoeffisientene i Eksemplet tegnet opp i figur 6. Regresjonskoeffisientene er standardisert slik at de har like varianser. I dette konkrete tilfellet er også variansen omtrentlig lik 1, idet vi har dividert med restvariansen fra regresjonen. Etter som det her er et stort antall frihetsgrader, vil derfor de plottete koeffisientene være omtrentlig normalfordelte med varians 1.

Figur 6 gir imidlertid ikke det forventede resultat. Alle punktene ligger stort sett på én rett linje. Det skulle tyde på at alle kom fra samme normalfordeling. Men en kan likevel ikke tolke dette som at ingen i virkeligheten er signifikant forskjellig fra 0. Da burde punktene ha ligget rundt den rette linjen tegnet inn på figuren. Den svarer til normalfordelingen N (0,1).

Vi tolker det observerte plottet slik at det egentlig er få av koeffisientene som har forventning 0, men at de har forventningsverdier av forskjellig størrelsesorden med en jevn variasjon fra 0 og oppover i absoluttverdi. På den måten kan fordelingen av regresjonskoeffisientene oppfattes som en blanding av to fordelinger hvor den ene er knyttet til feilleddenes fordeling og den andre til de sanne koeffisientenes fordeling.

I dette tilfellet er konklusjonen at det halvnormale plottet ikke gir særlig stor informasjon om hvor mange variable regresjonen bør inneholde.

Figur 6. Regresjon med 36 variable. Halvnormalplott av t-verdiene til regresjonskoeffisientene



6.5. P-plott

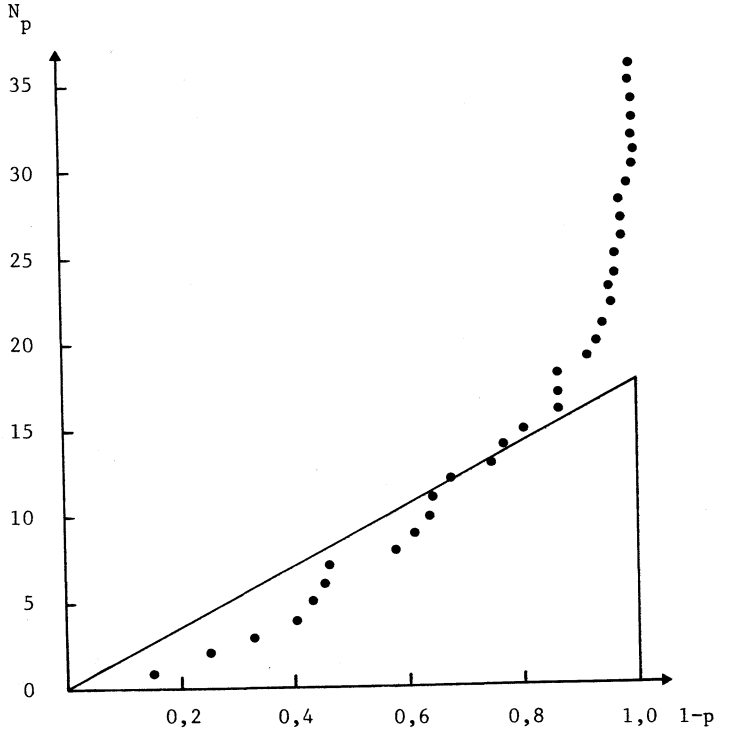
En annen teknikk for å behandle et stort antall tester simultant er foreslått av Schweder og Spjøtvoll (1982). Den består i å lage et plott av de beregnete signifikanssannsynlighetene (P-verdiene) for testene. Det som utnyttes er at for de sanne nullhypoteser er P-verdiene rektangulært fordelt i intervallet [0,1], mens for de andre vil P-verdiene stort sett være små.

Anta at vi har T nullhypoteser H_t (t=1, --, T) og at H_t blir forkastet når observatoren Z_t (i vårt tilfelle t-verdien) er stor. La F_t være den kumulative fordelingsfunksjonen under H_t. P-verdien, dvs. signifikanssannsynligheten for H_t er $P_t = 1 - F_t(Z_t)$. La T_0 være det ukjente antall sanne nullhypoteser, og la N_p være antall P-verdier større enn p. Siden P-verdien skulle være liten for en feil nullhypotese vil, når p er stor,

$$E(N_p) \approx T_0 (1-p)$$

Et plott av N_p mot 1-p skulle derfor for store p indikere en rett linje med stigningskoeffisient T_0. For små verdier av p, har vi at $E(N_p) > T_0 (1-p)$ siden falske nullhypoteser blir medregnet i N_p.

Figur 7. P-plott av de 36 regresjonskoeffisientene



Figur 7 viser p-plottet av signifikannssannsynlighetene for de 36 testene for at regresjonskoeffisientene er 0 i Eksemplet. Hvis ingen regresjonskoeffisienter var forskjellig fra null, skulle punktene gruppere seg rundt en rett linje. Det gjør de tydeligvis ikke. Det gjelder da å identifisere en rett linje ut fra den første del av plottet. De 4 første punktene ligger nesten på linje. Men 4 punkter er litt lite til at en kan stole på et estimat bare på dem. På skjønn har vi lagt inn en linje på figuren. Den er ment å representere hovedtendensen i første del av plottet. Den skjærer N_p - aksen et sted mellom $N_p = 17$ og $N_p = 18$.
Dette blir da estimatet av antall sanne hypoteser. Estimatet av antall regresjonskoeffisienter forskjellig fra 0 blir dermed 18 eller 19.

Et slikt estimat er selvfølgelig beheftet med usikkerheter. I Schweder og Spjøtvoll (1982) er angitt en mulig framgangsmåte for å gi et omtrentlig anslag på disse. Ved å gå inn i plottet for $p = 0.3$ får en estimert antall sanne nullhypoteser til å være $\hat{T}_0 = 17.1$. Ved å regne som om tekstene var uavhengige får dette estimatet variansen $\hat{T}_0 \cdot 0{,}3 \,/\, (1\text{-}0{,}3) = 7.5$. Altså et standardavvik lik 2.7 for det estimerte antall sanne nullhypoteser, og dermed også for det estimerte antall ikke sanne hypoteser. Dette skulle da gi et omtrentlig 95 prosent konfidensintervall for antall betydningsfulle effekter til å gå fra 13 til 24.

En annen bruk av det estimerte antall sanne hypoteser er følgende. Etter som antall sanne hypoteser er estimert til å være 17 bør en bruke et signifikansnivå $\alpha/17$ når en tester de enkelte hypoteser. Da vil det være maksimal sannsynlighet på α for minst en feilaktig forkastning.

Hvis en i Eksemplet bruker $\alpha = 0.05$ blir $\alpha/17 = 0.0029$. Fra vedlegg 2 finnes da at variablene

$$x_2,\ x_{17},\ x_{33},\ x_{34},\ x_{35},\ x_{121} \text{ og } x_{501}$$

er signifikant på dette nivået.

6.6. Resultatet av den foreløpige analyse for Eksemplet

Formålet med det nåværende trinn i analysen er å finne frem til de uavhengige variable som kan være betydningsfulle, samtidig som de med svært liten eller ingen innflytelse tas ut av modellen. På neste trinn skal vi se på kryssproduktledd mellom de variable som er beholdt. Til slutt lages en endelig modell hvor flere variable elimineres, men de med de viktigste førstegradseffekter og kryssproduktledd beholdes. For å unngå at en mister eventuelle viktige kryssproduktledd er det en fordel å beholde en god del variable fra første trinn, selv om en ikke er overbevist om deres statistiske signifikans.

Vi valgte å ta med 18 av de uavhengige variable til neste trinn i analysen. Dette var basert på

(i) P-plottet ga et estimat rundt 18 variable.

(ii) R^2-plottet gjorde et lite hopp ved antall variable lik 18, og det er liten endring i R^2 ved å ta inn flere variable.

(iii) De trinnvise prosedyrene BACWARDS og MAXR ga samme modell ved 18 variable. Selv om dette ikke er noe hovedpoeng, gjorde det lett å velge de 18 variable som skulle brukes.

Den beregnete regresjonsfunksjonen med disse 18 variable er gitt i tabell 4.

Tabell 4. Regresjon på de 18 variable valgt i første trinn

Variabel	Regresjons-koeffisient	Standardavvik	Signifikans-sannsynlighet
Konstantledd	2,107		
x_2	0,237	0,0395	0,0001
x_4	0,082	0,0413	0,0464
x_{10}	0,093	0,0420	0,0270
x_{11}	-0,0028	0,0015	0,0559
x_{14}	0,171	0,0842	0,0422
x_{15}	0,090	0,0459	0,0508
x_{17}	0,232	0,0494	0,0001
x_{18}	0,026	0,0115	0,0237
x_{23}	0,151	0,0553	0,0065
x_{25}	-0,202	0,0843	0,0170
x_{28}	-0,134	0,0549	0,0147
x_{29}	0,125	0,0432	0,0040
x_{33}	0,277	0,0572	0,0001
x_{34}	0,158	0,0452	0,0005
x_{35}	0,226	0,0583	0,0001
x_{121}	0,323	0,0671	0,0001
x_{501}	-0,275	0,0717	0,0001
x_{506}	0,082	0,0441	0,0620

$R^2 = 0,2818$

7. DIAGNOSTISKE MÅL

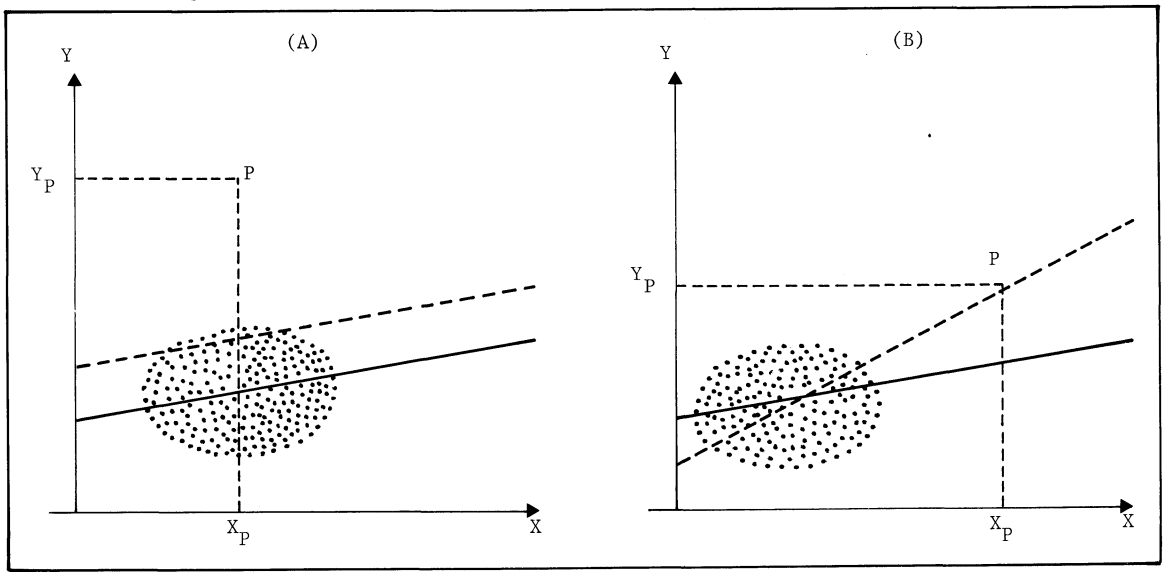
7.1. Innflytelsen til enkeltobservasjoner

I de fleste datamaterialer er det observasjoner som har en ekstrem verdi m.h.t. en eller flere av de variablene som inngår. Slike observasjoner kan, dersom de ikke er resultatet av en feilmåling, være verdifulle ved at de øker presisjonen ved estimeringen. På den annen side kan observasjoner som er ekstreme p.g.a. feil ha en ødeleggende effekt på estimeringen.

Vi skal se på to eksempler som viser hvordan én ekstrem observasjon i sterk grad kan påvirke resultatet av en regresjon. I eksemplene inngår én avhengig variabel, y, og én uavhengig variabel, x. Figur 8 (A) illustrerer hvordan én observasjon, P, kan føre til en vertikal forskyvning av regresjonslinja, mens figur 8 (B) illustrerer hvordan én observasjon i stor grad kan påvirke stigningskoeffisienten til regresjonslinja.

Figur 8. Regresjonslinja ved regresjon med én uavhengig variabel. Den stipla linja er regresjonslinja når observasjon P er inkludert i datamaterialet. Den heltrukne linja er regresjonslinja når P er fjernet fra datamaterialet.

Figur 8. Regresjonslinja ved regresjon med én uavhengig variabel. Den stipla linja er regresjonslinja når observasjon P er inkludert i datamaterialet. Den heltrukne linja er regresjonslinja når P er fjernet fra datamaterialet



I figur 8 (A) ser vi at observasjon P har stor residual, mens den har liten residual i figur 8 (B). Dette viser at det ikke er nok å studere residualene for å identifisere ekstreme observasjoner. Det er to grunner til at en bør se nærmere på ekstreme observasjoner:

(i) Observasjoner kan ha blitt ekstreme p.g.a. feilmålinger eller punchefeil o.l. I så fall bør de fjernes.

(ii) Observasjoner kan være tatt under så spesielle forhold at de ikke bør være med i analysen.

Når det gjelder pkt. (ii), vil vi advare mot å bruke dette som en unnskyldning for å fjerne observasjoner som påvirker analysen i en retning en ikke ønsker. Som nevnt tidligere kan ekstreme observasjoner gi verdifull informasjon, og en bør derfor ikke fjerne dem uten at en har gode grunner for det.

I en situasjon der det er bare én uavhengig variabel slik som i figur 8, er det lett å identifisere ekstreme observasjoner med stor innflytelse bare ved å se på et plott. Når en har mange forklaringsvariable, kan det derimot være svært komplisert å finne fram til ekstreme observasjoner. En observasjon kan nemlig være ekstrem m.h.t. alle variablene simultant uten at den har en ekstrem verdi m.h.t. noen av variablene separat.

I dette kapitlet skal vi se på diagnostiske metoder som kan brukes til å identifisere ekstreme observasjoner, og som også gir informasjon om graden av ekstremitet. Videre skal vi se på mål for innflytelsen til de enkelte observasjonene ved parameterestimeringen. De diagnostiske målene som er beskrevet i dette kapitlet, finnes alle i programpakken SAS (Statistical Analysis System). I Belsley, Kuh & Welsch (1980) er de forskjellige målene mer utførlig beskrevet enn her.

7.1.1. Definisjoner

I dette avsnitt defineres symboler som er brukt i kapittel 7.

n = antall observasjoner

p = antall uavhengige variable

Y = (nx1) - vektoren av den avhengige variabelen

X = (nx(p+1)) - matrisen av de uavhengige variable (antall kolonner er p+1 p.g.a. konstantleddet).

$$Y = \begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \text{----} & x_{1p} \\ & . & & . \\ & . & & . \\ 1 & x_{n1} & & x_{np} \end{bmatrix}$$

x_i = rad nr. i i X

Vi forutsetter at

$E(Y|X) = X\beta$, der β er ((p+1) x1) - vektoren av ukjente parametre.

$b = (X'X)^{-1} X'Y$ = minste kvadraters estimatorene for β

$\hat{Y} = Xb$

$\hat{y}_i = x_i b$

σ^2 = variansen til restleddene

$s^2 = \frac{1}{n-p-1} (Y-Xb)'(Y-Xb)$

s^2 er en estimator for σ^2

Y(i) = ((n-1)x1)-vektoren av den avhengige variabelen etter at observasjon nr. i er utelatt.

X(i) = ((n-1)x(p+1))-vektoren av de uavhengige variablene etter at observasjon nr. i er utelatt

b(i) = minste kvadraters estimatoren for β etter at observasjon nr. i er utelatt

7.1.2. Hattematrisen

Følgende matrise blir ofte kalt hattematrisen:

$$H = X(X'X)^{-1} X'$$

En har:

$$\hat{Y} = Xb = HY$$

Matrisen H er projeksjonsmatrisen for Y ned i prediktor-rommer utspent av kolonnene i matrisen X.

La $h_i = x_i (X'X)^{-1} x_i'$

betegne det i'te diagonalelementet i matrisen H. Elementet h_i kan betraktes som et mål for avstanden fra i'te observasjon til "tyngdepunktet" i X-rommet. Vi skal begrunne dette nærmere.

Anta først at vi har én uavhengig variabel slik at

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ : & : \\ : & : \\ 1 & x_n \end{bmatrix}$$

Det kan vises at h_i i dette tilfellet blir:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2},$$

der $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$. Altså jo mer x_i avviker fra gjennomsnittet desto større blir h_i. Merk også at

$$\sum_i h_i = 2 = p+1$$

I Belsley, Kuh & Welsch (1980) er det vist at

$$\sum_{i=1}^{n} h_i = p+1$$

generelt. Det er også vist at $0 \leq h_i \leq 1$.

Matrisen X kan skrives på følgende form uttrykt ved sine kolonnevektorer:

$$X = (1, X_1 \text{ ----}, X_p)$$

La $\tilde{X}$ betegne (nxp) - matrisen som en får ved "sentrering" av alle forklaringsvariable, dvs.

$\tilde{X} = (X_1 - \bar{X}_1, \text{ ----}, X_p - \bar{X}_p)$, der

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

La $\tilde{H}$ betegne hattematrisen etter "sentrering" av både Y og X. Vi har da:

$$Y - \bar{Y} = HY - \bar{Y} = \tilde{H}Y$$

Det kan da vises at

$$\tilde{h}_i = \tilde{x}_i (\tilde{X}'\tilde{X})^{-1} x_i = h_i' - \frac{1}{n}$$

Dersom $\tilde{X}$ er ortonormal, slik at $\tilde{X}'\tilde{X} = I$, blir

$\tilde{h}_i = \sum_{j=1}^{p} \tilde{x}_{ij}^2$, der $\tilde{x}_{ij}$ er element (i, j) i $\tilde{X}$.

I dette tilfellet er altså $\tilde{h}_i$ kvadratet av den "vanlige" Euklidske avstand fra 0. Generelt kan en si at $\tilde{h}_i$ er en veiet versjon av kvadratet av den Euklidske avstand.
Dersom vektoren av sentrerte forklaringsvariable, $\tilde{X}$, var multinormalt fordelt, ville $\tilde{X}$ ha sannsynlighetstetthet

$$k e^{-\frac{1}{2}\tilde{X}'\Sigma^{-1}\tilde{X}},$$

der k er en konstant og Σ er kovariansmatrisen til $\tilde{X}$.

Sannsynlighetskonturene med konstant tetthet vil da være ellipsoider rundt origo. I denne situasjonen kan en tolke $\tilde{h}_i$ slik at $\tilde{h}_i$ forteller ved hvilken sannsynlighetskontur observasjon nr. i befinner seg, dvs. at alle observasjoner som ligger på samme sannsynlighetskontur vil ha samme verdi m.h.p. $\tilde{h}$.

La oss se på et eksempel med to forklaringsvariable x og z som er sentrerte. $\tilde{X}$-matrisen har da følgende utseende:

$$\begin{bmatrix} x_1 & z_1 \\ x_2 & z_2 \\ : & : \\ : & : \\ x_n & z_n \end{bmatrix}$$

Litt algebra gir da at

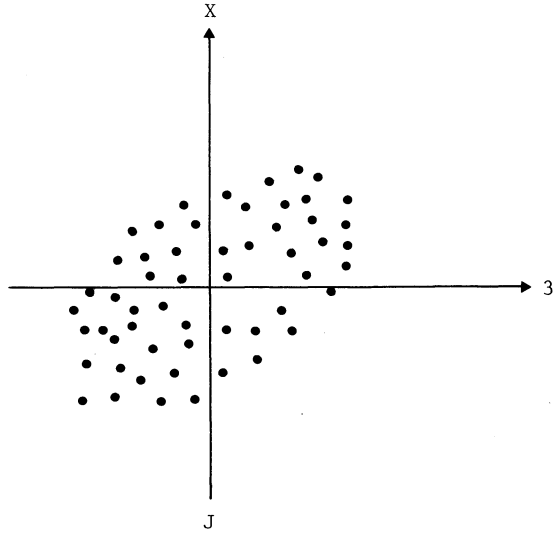
$$h_i = \frac{x_i^2 \sum z_j^2 - 2x_i z_i \sum x_j z_j + z_i^2 \sum x_j^2}{\sum z_j^2 \sum x_j^2 - (\sum x_j z_j)^2}$$

Uten tap av generalitet kan antas at x og z er skalert slik at $\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} z_j^2 = n$ og $\sum x_j z_j = rn$, der r er den empiriske korrelasjonskoeffisienten mellom x og z. Vi får da

$$\tilde{h}_i = \frac{x_i^2 - 2x_i z_i r + z_i^2}{n(1-r^2)}$$

For ortogonale variable, dvs. $r = 0$, er $\tilde{h}_i$ lik kvadratet av den vanlige Euklidske avstanden fra origo dividert på n. Når r, x_i og z_i er positive blir h_i mindre enn den "vanlige avstand" skulle tilsi. Når $r > 0$, $x_i > 0$ og $z_i < 0$, blir h_i større enn den "vanlige avstand" skulle tilsi. Dette har sammenheng med at når $r > 0$, er det mer sannsynlig at x_i og z_i har samme fortegn enn at de har motsatt fortegn. Situasjonen er illustrert i figur 9. Punkter som ligger på samme ellipse har samme $\tilde{h}_i$-verdi.

Figur 9. En grov illustrasjon av "$\tilde{h}$,-avstand"



I Belsley, Kuh & Welsch (1980) er det vist at under forutsetning om normalfordeling på x_i-ene vil

$$\frac{(n-p-1)(h_i - \frac{1}{n})}{p(1-h_i)}$$

være F-fordelt med p og n-p-1 frihetsgrader. For p > 10 og (n-p-1) > 50 er 95%-fraktilen for F mindre enn 2. Det er derfor naturlig å se nærmere på observasjoner som har en h_i slik at uttrykket over er større enn 2.

At

$$\frac{(n-p-1)(h_i - \frac{1}{n})}{p(1-h_i)} > 2$$

er ekvivalent med at

$$h_i > \frac{2p+1 - \frac{p+1}{n}}{n+p-1}$$

I situasjoner der n er mye større enn p gjelder at

$$\frac{2p+1-\frac{p+1}{n}}{n+p-1} \approx \frac{2(p+1)}{n}$$

Siden $\frac{2(p+1)}{n}$ er 2 ganger gjennomsnittsverdien for h_i-ene, er det en størrelse som er lett å huske. Derfor blir $\frac{2(p+1)}{n}$ foreslått som en grov grenseverdi for observasjoner en bør studere nærmere.

7.1.3. Standardiserte residualer

Residualen til observasjon nr. i er definert som

$$e_i = y_i - y_i = y_i - x_i b$$

Under forutsetning av at Var $(y_i) = \sigma^2$ for alle i, der i = 0,1, ...n, kan variansen til y_i og e_i uttrykkes ved σ^2 og det i'te diagonalelementet i hattemattematrisen

$$\text{Var}(\hat{y}_i) = h_i\sigma^2$$

og Var $(e_i) = (1-h_i)\sigma^2$

Var (e_i) og Var $(\hat{y}_i)$ er altså ikke konstante. En stor h_i medfører at Var (e_i) blir liten og Var $(\hat{y}_i)$ stor.

I en analyse av residualene er det mer korrekt å bruke standardiserte residulaer (som har samme varians) enn ustandardiserte residualer. I SAS finnes to former for standardiserte residualer:

(i) STUDENT $= \dfrac{e_i}{s\sqrt{1 - h_i}}$

(ii) RSTUDENT $= \dfrac{e_i}{s(i)\sqrt{1 - h_i}}$

Forskjellen mellom (i) og (ii) er at observasjon nr. i er utelatt ved estimering av σ i det siste tilfellet. Fordelen med å utelate observasjon nr. i er at teller og nevner da blir stokastisk uavhengige under forutsetning av at y_i- ene er normalfordelt. Under nevnte forutsetning er altså RSTUDENT t-fordelt med n-p-2 frihetsgrader.

Fra (i) og (ii) ser vi at observasjoner med stor h_i vil få en større standardisert residual enn ustandardisert residual relativt til andre observasjoner. Dersom en studerer residualer for å identifisere ekstreme observasjoner, vil altså ekstreme observasjoner i x-rommet tiltrekke seg mer oppmerksomhet når en bruker standardiserte residualer enn når en bruker ustandardiserte residualer. Som illustrert i figur 8 (B) må en være oppmerksom på at en ekstrem observasjon likevel kan ha en liten standardisert residual, siden observasjonen i forholdsvis stor grad vil "trekke til seg" regresjonslinja.

7.1.4. COOKS D

Cook (1977) har foreslått følgende mål for innflytelsen til i'te observasjon:

$$D_i = \frac{(b - b(i))'x'x(b - b(i))}{(p + 1)\, s^2},$$

der b som tidligere betegner ((p+1) x1) - vektoren av minste - kvadraters - estimatorene for regresjonskoeffisientene og b(i) den tilsvarende vektoren etter at observasjon nr. i er fjernet.

En (1 - α) x 100% konfidensellipsoide for den ukjente parametervektoren, β, er gitt ved mengden av vektoren β^* som tilfredstiller

$$\frac{(\beta^* - b)'x'x\,(\beta^* - b)}{(p + 1)\, s^2} \leq F\,(p+1,\ n-p-1,\ 1-\alpha),$$

der F (p+1, n-p-1, 1-α) er (1-α) - fraktilen i F-fordelingen med p+1 og (n-p-1) frihetsgrader. Det er derfor naturlig å sammenligne D_i med fraktilene i nevnte F-fordeling. Dersom $D_i \approx F\,(p-1,\ n-p-1,\ \alpha)$, betyr det at fjerningen av i'te observasjon medfører at b blir forskjøvet mot randen av en α . 100% - konfidensellipsoide for β basert på b. Stor D_i tyder derfor på stor innflytelse av i'te observasjon.

Da størrelsen D_i ikke selv har en F-fordeling, har vi ikke sammenlignet D_i-ene med fraktilene i F-fordelingen. Vi har derimot sett nærmere på observasjoner som har stor D i forhold til andre observasjoner.

COOKS D er nærmere beskrevet i Cook (1977).

7.1.5. DFFITS

Et tilsvarende mål til Cooks D er:

$$DFFITS_i = (\hat{y}_i - \hat{y}_i(i))/\, s(i)\, \sqrt{h_i},$$

der $\hat{y}_i = x_i b$ og $\hat{y}_i(i) = x_i b(i)$

Som nevnt i avsnitt 7.1.3. er $\mathrm{Var}\,(\hat{y}_i) = h_i\sigma^2$. En estimator for standardavviket til y_i er derfor $s(i)\,\sqrt{h_i}$.

$DFFITS_i$ er et skalert mål for endringen i $\hat{y}_i$ når observasjon nr. i utelates. En stor absoluttverdi for $DFFITS_i$ tyder på en observasjon med stor innflytelse. $DFFITS_i$ har den tilsvarende relasjon til t-fordelingen som Cooks D har til F-fordelingen. Det er imidlertid ikke særlig interessant å sammenligne med fraktilen i t-fordelingen siden størrelsen til $DFFITS_i$ vil være avhengig av antall observasjoner. $DFFITS_i$ kan skrives på følgende form (Belsley, Kuh & Welsch (1980)):

$$DFFITS_i = \sqrt{\frac{h_i}{1-h_i}}\ \frac{e_i}{s(i)\,\sqrt{1-h_i}} = \sqrt{\frac{h_i}{1-h_i}}\, e_i^*,$$

$$\text{der } e_i^* = \frac{e_i}{s(i)\sqrt{1-h_i}} = \text{RSTUDENT (jfr. avsn. 7.1.3.)}$$

Anta at vi har en "perfekt balansert design-matrise" X slik at $h_i = \frac{p+1}{n}$ for alle i. Da blir:

$$DFFITS_i = \sqrt{\frac{p+1}{n-p-1}} \cdot e_i^*$$

Under forutsetning av at y_i-ene er normalfordelt, vil e_i^* være t- fordelt. Når n er stor er 0,975-fraktilen i t-fordelingen tilnærmet lik 2. Grenseverdien $|e_i^*| = 2$ vil for $h_i = \frac{p+1}{n}$ og store n svare til:

$$|DFFITS_i| \approx 2\sqrt{\frac{p+1}{n}}$$

Belsley, Kuh & Welsch (1980) foreslår $2\sqrt{\frac{p+1}{n}}$ som en "grov" grenseverdi for observasjoner en bør se nærmere på. I praksis ser en nærmere på observasjoner som har stor $|DFFITS_i|$ relativt til andre observasjoner. I små datamaterialer ser en lett hvilke observasjoner dette gjelder. "Grenseverdien" $2\sqrt{\frac{p+1}{n}}$ har først og fremst praktisk verdi i store datamaterialer.

$DFFITS_i$ er et skalert mål for endringen i $\hat{y}_i$ når observasjon nr. i utelates. Tilsvarende kan en lage et skalert mål $DFFITS_{ik}$ for endringen i Y_k når observasjon nr. i utelates.

$$DFFITS_{ik} = \frac{x_k(b-b(i))}{s(i)\sqrt{h_k}}$$

Det kan vises at $|DFFITS_{ik}| \leq |DFFITS_i|$ for alle $k \neq i$. Det er derfor ikke interessant å studere $DFFITS_{ik}$ når $|DFFITS_i|$ er liten.

7.1.6. DFBETAS

Et skalert mål for endringen i koeffisient nr. j ved utelatelse av observasjon nr. i er

$$DFBETAS_{ij} = (b_j - b_j(i))/s(i)\sqrt{(X'X)^{-1}_{jj}},$$

der $(X'X)^{-1}_{jj}$ er element (j,j) i matrisen $(X'X)^{-1}$. Variansen til b_j er $\sigma^2(X'X)^{-1}_{jj}$. En estimator for standardavviket til b_j er derfor $s(i)\sqrt{(X'X)^{-1}_{jj}}$. Grunnen til at en benytter s(i) istedet for s, er at teller og nevner da blir stokastisk uavhengige når y_i-ene er normalfordelte.

En stor verdi av $|DFBETAS_{ij}|$ indikerer at observasjon nr. i har stor innflytelse ved estimering av koeffisient nr. j, β_j. Belsley, Kuh & Welsch (1980) har foreslått at $|DFBETAS_{ij}|$ betraktes som "stor" når $|DFBETAS_{ij}| \geq \frac{2}{\sqrt{n}}$.

7.1.7. COVRATIO

$COVRATIO_i$ måler endringen i determinanten når en utelater observasjon nr. i.

$$COVRATIO_i = \det\left[s^2(i)(X(i)'X(i))^{-1}\right] / \det\left[s^2(X'X)^{-1}\right]$$

$COVRATIO_i$ fokuserer ikke bare på endringer i $(X'X)^{-1}$ når obsrvasjon nr. i blir utelatt, men også på endringen i $\hat{y}$ ved at σ^2 blir estimert ved henholdsvis s^2 og $s^2(i)$. Dersom $COVRATIO_i \approx 1$, indikerer dette at observasjon nr. i har liten innflytelse på estimeringen. Belsley, Kuh & Welsch (1980) har vist at $COVRATIO_i$ kan skrives på følgende form:

$$COVRATIO_i = \frac{1}{\left[\frac{n-p}{n-p-1} + \frac{e_i^{*2}}{n-p-1}\right]^{p+1}(1-h_i)}$$

Vi ser at $COVRATIO_i$ vil ha en tendens til å være stor når h_i er stor, og liten når e_i^* er stor.

Belsley, Kuh & Welsch har foreslått at en ser nærmere på observasjoner med

$$\left|COVRATIO - 1\right| \geq \frac{3(p+1)}{n}$$

siden slike observasjoner kan ha stor innflytelse.

7.2. Mål for kolinearitet

To variable sies å være kolineære hvis datavektorene som representerer dem ligger på samme linje. Mer generelt er k variable kolineære hvis datavektorene som representerer dem, ligger i et delrom med dimensjon mindre enn k, dvs. hvis en eller flere av vektorene er en lineærkombinasjon av andre vektorer.

Eksakt kolinearitet forekommer sjeldent i praksis, likevel har en ofte problemer i regresjonsanalyse med at variable er "nesten" kolineære. Innen regresjonsanalyse brukes en "løsere" definisjon av begrepet kolinearitet enn den som er presentert over. To variable sies å være kolineære dersom datavektorene deres nesten ligger på samme linje, dvs. hvis vinkelen mellom de to vektorene er liten. Dette er ekvivalent med at korrelasjonen mellom de to variablene er høy.

Flere enn to variable defineres tilsvarende til å være kolineære dersom den multiple korrelasjonskoeffisienten for den ene variabelen m.h.p. de andre variablene er stor.

Dersom vi i regresjonsmodellen $y = X\beta + \varepsilon$, bringer inn en ny forklaringsvariabel som er kolineær med forklaringsvariable som allerede er i modellen, vil den nye variabelen gi liten eller ingen informasjon i tillegg til den informasjonen som de andre variablene inneholder. Den nye variabelen vil altså gi lite eller ikke noe bidrag til forklaringen av y.

Dersom det er kolinearitet blant forklaringsvariablene i en regresjon, vil dette ha en ødeleggende effekt på estimeringen. Varians-kovarians-matrisen til estimatorene for regresjonskoeffisientene er gitt ved:

$$Var(b) = \sigma^2(X'X)^{-1},$$

der $(X'X)$ er en $((p+1)x(p+1))$-matrise. Dersom vi har eksakt kolinearitet, vil rang $(X'X)$ være mindre enn p+1, og $(X'X)^{-1}$ og $Var(b)$ vil ikke eksistere. Dersom vi har nesten eksakt kolinearitet, vil dette medføre at variansen til en eller flere (eller muligens alle) av b_j-ene blir stor. I Belsley, Kuh & Welsch (1980) er det mer om dette temaet.

I dette kapitlet skal vi se på metoder for å

(i) oppdage kolineære relasjoner blant forklaringsvariablene

(ii) identifisere hvilke forklaringsvariable som er involvert i hver kolineær relasjon.

Metodene som er beskrevet her, er inneholdt i programpakken SAS.

7.2.1. TOLERANCE OG VARIANCE INFLATION

Som tidligere anta at vi har p forklaringsvariable, $x_1,\ldots, x_p$ i modellen. Et mål for korrelasjonen til forklaringsvariabel x_j med de andre forklaringsvariablene er "$TOLERANCE_j$"

$$TOLERANCE_j = 1 - R_j^2,$$

der R_j^2 er den "vanlige R^2" i en regresjon med x_j som avhengig variabel og de andre forklaringsvariablene som uavhengige variable.

En tar altså utgangspunkt i modellen

$$x_j = \beta_0 + \sum_{i \neq j} \beta_i x_i + \varepsilon ,$$

og estimerer (nx1)-vektoren $X_j = (x_{1j},\ldots, x_{nj})'$ ved vanlig regresjon. Størrelsen R_j er lik den empiriske korrelasjonskoeffisienten mellom estimerte og observerte x_j-er.

Målet TOLERANCE er stort når R_j^2 er liten, og lite når R_j^2 er stor. En alternativ form av dette målet er "$VARIANCE\ INFLATION_j$" eller "VIF_j", som er definert ved

$$VIF_j = 1/TOLERANCE_j$$

Når variabelen x_j er sterkt korrelert med andre variable, er VIF_j stor, og når x_j er svakt korrelert med andre variable, er VIF_j liten.

7.2.2. Kondisjoneringsindeks og variansdekomponering

Målene VIF og TOLERANCE måler i hvilken grad hver variabel er korrelert med andre variable. De gir derimot liten informasjon om hvordan variablene er korrelerte med hverandre.

En kan studere sammenhengen mellom to og to variable ved å se på korrelasjonsmatrisen til forklaringsvariablene, X. Dette har den svakhet at hvis en har en større gruppe av variable som er lineært avhengige av hverandre, er det usikkert om dette vil bli oppdaget siden korrelasjonen mellom to og to variable kan være liten. For å studere sammenhengen mellom flere variable samtidig finnes bedre metoder. En av disse, som finnes i SAS, bygger på de såkalte kondisjoneringsindekser og dekomponering av matrisen X. Det blir nå gitt en kort beskrivelse av metoden.

Matrisen X'X blir skalert slik at den får 1-ere på diagonalen. Egenverdiene til X'X beregnes og listes ut. Disse er kvadratene til singulærverdiene til X. Videre beregnes kondisjoneringsindeksene som er kvadratrøttene til den største egenverdien dividert på de ulike egenverdiene.

Stor variasjon i egenverdiene tyder på at det er kolinearitet i dataene. Dersom en av variablene er en eksakt lineærkombinasjon av andre variable, vil en av egenverdiene være 0. Utfra empiriske eksperimenter har Belsley, Kuh & Welsch (1980) funnet at kondisjoneringsindekser av størrelse 5-10 indikerer svak kolinearitet i dataene, mens kondisjoneringsindekser i området 30-100 indikerer sterk kolinearitet i dataene.

Matrisen X'X kan dekomponeres slik at $X'X = VD^2V'$, der V er ortogonal og D^2 er en diagonalmatrise. Elementene på diagonalen er egenverdiene til X'X. Varians-kovarians-matrisen til b kan skrives på følgende form:

$$Var(b) = \sigma^2(X'X)^{-1} = \sigma^2 VD^{-2}V'$$

Videre kan variansen til koeffisient nr. j, b_j, skrives:

$$(4) \quad Var(b_j) = \sigma^2 \sum_k \frac{v_{jk}^2}{\lambda_k},$$

der v_{jk} er element (j,k) i matrisen V og λ_k er egenverdi nr. k til X'X, dvs. element (k,k) i D^2.

Vi ser at (4) er en dekomponering av variansen slik at en får en komponent knyttet til hver egenverdi eller egenvektor.

SAS lister ut andelen av variansen som er "forklart" ved hver egenvektor. For variable som er korrelerte med hverandre, vil det være en tendens til at variansen har størst komponenter m.h.t. de samme egenverdiene. Vi har et kolineært problem når samme komponent bidrar sterkt til variansen til to eller flere variable, samtidig som egenvektoren har en stor kondisjoneringsindeks.

7.2.3. Et eksempel

Vedlegg 3 er et eksempel på en SAS-utskrift med utlisting av egenverdier, kondisjoneringsindekser, andelen av variansen knyttet til hver egenvektor, TOLERANCE og VIF. Variablene x_2, K1, K2 og K4 har alle TOLERANCE mindre enn 0,03. Det viser at disse 4 variablene er sterkt korrelert med andre variable, og at det følgelig er høy kolinearitet i dataene.

Kondisjoneringsindeksene varierer fra 1 til 41,8, og det indikerer også høy kolinearitet i dataene. Komponenten knyttet til egenvektor nr. 8 bidrar med mer enn 80 prosent av variansen til variablene x_2, K1, K2 og K4. Det indikerer en sterk kolineær relasjon mellom de 4 variablene. Kondisjoneringsindeks nr. 7 er 21,5. Komponenten knyttet til egenvektor nr. 7 bidrar med over 45 prosent av variansen til de 3 variablene x_3, x_{33} og K3. Dette tolker vi slik at det også er en kolineær relasjon mellom disse 3 variablene, men at denne relasjonen er adskillig svakere enn den første.

Komponenten knyttet til egenvektor nr. 8 bidrar med over 35 prosent av variansen til samtlige variable. Det indikerer at alle variable er mer eller mindre korrelerte med hverandre.

8. INNFØRING AV KRYSSPRODUKTLEDD I MODELLEN

8.1. Innledning

I dette kapitlet utvides modellen (3) i kapittel 6 til også å omfatte kryssproduktledd, altså en modell av formen

$$(5)\quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \beta_{12} x_1 x_2 + \ldots + \beta_{p-1,p} x_{p-1} x_p + \text{feil}.$$

Vi vil også av og til omtale kryssproduktleddene som samspillsledd. Vanligvis ville en også ha tatt med rene annengradsledd i (5), slik at modellen var et fullstendig annengradspolynom i forklaringsvariablene. Grunnen til at det ikke er gjort her, er at i Eksempelet er det bare dikotome vaiable. Da blir annengradsleddet lik førstegradsleddet.

Som for førstegradsleddene regner vi med at bare et fåtall av samspillsleddene gir signifikante bidrag. Problemet er å anslå hvor mange og hvilke. For å gjøre dette vil vi bruke tilsvarende metoder som i kapittel 6. Den viktigste forskjellen er at nå er det enda flere mulige effekter å handskes med. Med utgangspunkt i p hovedeffekter, blir det $\frac{1}{2}p(p-1)$ samspillseffekter. For Eksempelet med p = 18 funnet i kapittel 6, er det altså 153 mulige samspillseffekter.

8.2. Antall betydningsfulle kryssproduktledd

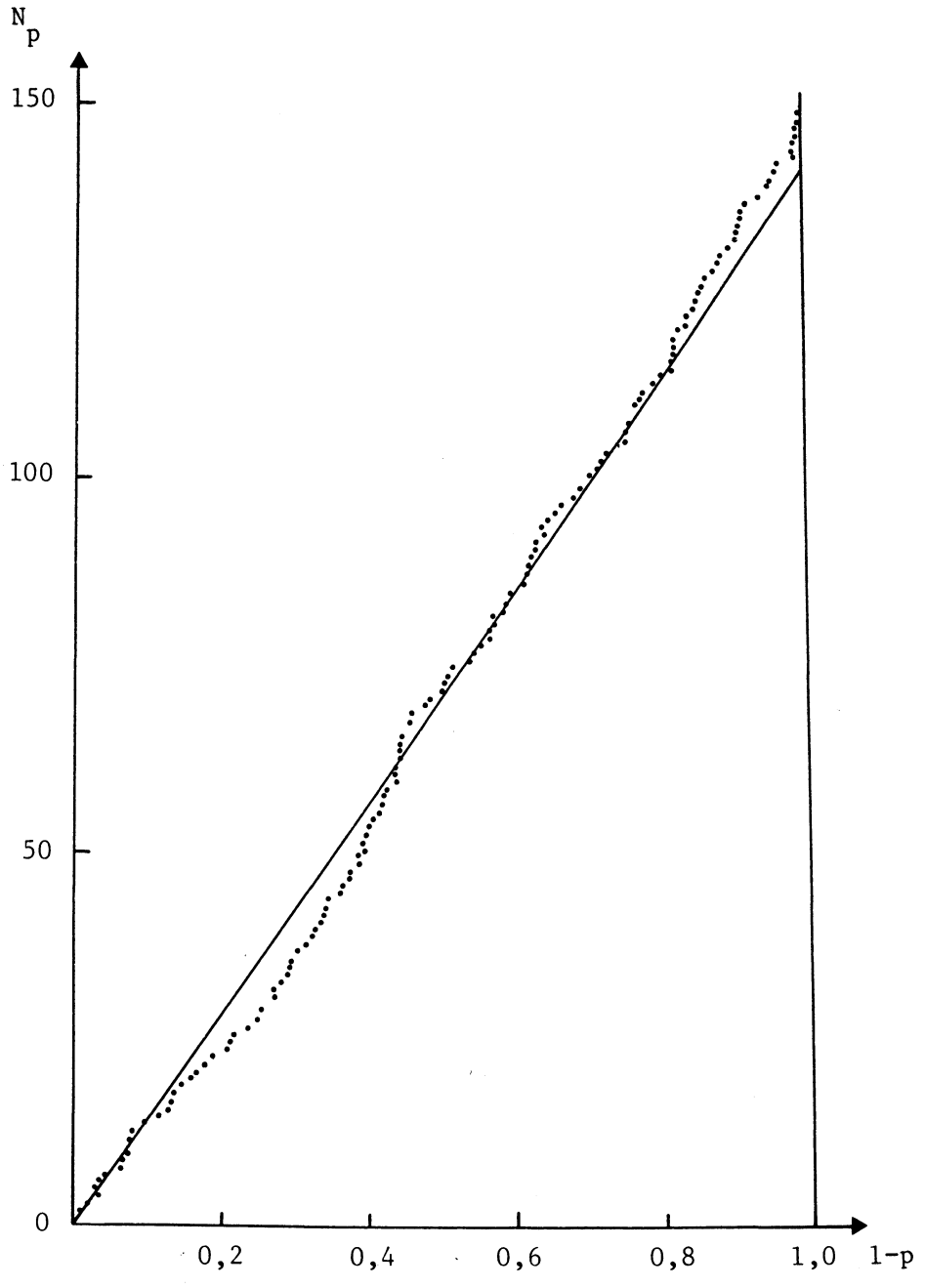
Den generelle framgangsmåte er å kjøre den store regresjonsmodellen (5) med hovedeffektene funnet som beskrevet i kapittel 6 og med alle deres kryssproduktledd. Ved et P-plott kan en da anslå antall betydningsfulle kryssproduktledd.

En støter imidlertid på den praktiske vanskelighet at regresjonsmodellen kan inneholde for mange variable. I Eksempelet blir det totale antall uavhengie variable 18 + 153 = 171. Det er mer enn regresjonsprogrammene i SAS kan ta.

For å omgå dette problemet måtte vi modifisere den generelle fremgangsmåte noe. Vi delte opp samspillene i 3 grupper slik at disse tilsammen utgjorde alle samspill. Gjennomgangsstørrelsen på gruppene er da 51. Hver gruppe ble kjørt i en regresjonsanalyse sammen med de 18 hovedeffektene. På den måten får en P-verdier for alle samspillene. Det er riktignok ikke P-verdiene relativt til den fulle modellen.

Figur 10 viser et P-plott av signifikanssannsynlighetene for de 149 testene for at samspillseffektene er 0 i Eksempelet. Ved å gå inn plottet for p = 0,3 får en estimert antall sanne nullhypoteser til å være 99/0,7 = 141,4. Da det totalt inngår 149 samspill, vil estimatet for antall betydningsfulle samspill være 7 eller 8. (Merk at her inngår 149 samspill og ikke 153. Grunnen er at regresjonsberegningene viste at 4 av dem var lineært avhengige av de andre). Estimatet er relativt usikkert. Det omtrentlige standardavviket er $\sqrt{141,4 \cdot 0,3/0,7} = 7,8$.

Figur 10. P-plott for 149 kryssproduktledd



8.3. Utvelgelse av kryssproduktledd

I foregående avsnitt ble antall betydningsfulle kryssproduktledd anslått ut fra et P-plott. For å finne frem til hvilke ledd som er betydningsfulle kan en i prinsippet gå frem som i kapittel 6. Der ble R^2-plottet brukt sammen med estimatet for antall effekter ut fra P-plottet. Et knekkpunkt eller utflating av R^2-plottet i nærheten av det estimerte antall effekter gir grunnlag til å stoppe med de variable en har i det punktet.

Det er imidlertid en forskjell fra situasjonen i kapittel 6. Der var vi interessert i å ta med oss i meste laget av effekter, med tanke på at vi ikke skulle miste interessante samspill på annet trinn i analysen. Men nå gjelder det å begrense seg til de effekter som er mest relevante. En må altså være mer omhyggelig i valget en gjør.

I Eksempelet er det ennå en ekstra vanskelighet. Etter som det ikke er mulig å kjøre regresjon med alle samspillseffekter i modellen, må en først fjerne en del av disse. Nemlig helst de med liten betydning. Som grunnlag for å fjerne variable brukte vi P-verdien og målet TOLERANCE.

Vi fjerner variable med lav "TOLERANCE" og stor P-verdi for å oppnå økt presisjon i estimeringen. Dersom "TOLERANCE" er liten for flere av variablene, betyr det at en har problemer med kolinearitet. Det er ønskelig med verdier nær 1 for alle variablene. Når en fjerner flere variable med liten "TOLERANCE", kan det være fare for at en fjerner en gruppe av variable som alle er innbyrdes korrelerte og ukorrelerte med variablene som blir igjen i modellen. For å unngå dette problemet med å fjerne for mange variable kan en i SAS gjøre bruk av "OPTION COLLIN" som består av en diagnostisk metode som ikke bare viser hvor mye variablene er korrelerte med hverandre, men også hvordan (jfr. avsnitt 7.2.2 og 7.2.3).

Variable med "TOLERANCE" nær 1 er nesten ukorrelerte med andre variable. Parameterestimat for slike variable påvirkes lite av fjerning av andre variable. En risikerer derfor ikke at fjerning av variable fører til at en variabel med stor "TOLERANCE" og opprinnelig stor P-verdi, blir signifikant. Følgelig fjernet vi variable med stor "TOLERANCE" og stor P-verdi med liten risiko for å gjøre noe galt.

8.4. Eksemplet

Ved å bruke den noe skjønnsmessige fremgangsmåten beskrevet i foregående avsnitt, reduserte vi i Eksemplet antall samspill ned til 53. På disse pluss de 18 hovedeffektene kjørte vi de trinnvise prosedyrene BACKWARDS (med nivå 0,03) og MAXR. I figur 11 er gjengitt R^2-plottet basert på MAXR. En ser at det har en liten knekk ved 8 samspillsledd. De to trinnvise prosedyrene ga også de samme 8 variable på dette trinnet. Dette sammen med estimatet 7-8 for antall betydningsfulle samspill fra avsnitt 8.2, gjorde det lett å bestemme seg for de 8 "beste" samspillsvariable en fant ved de trinnvise metodene. Dette var samspillene

$K4 = x_2 \cdot x_{33}$ $\qquad$ $K121 = x_{15} \cdot x_{33}$

$K12 = x_{19} \cdot x_{34}$ $\qquad$ $K132 = x_{35} \cdot x_{34}$

$K85 = x_2 \cdot x_{15}$ $\qquad$ $K151 = x_{28} \cdot x_{121}$

$K114 = x_{25} \cdot x_{29}$ $\qquad$ $K158 = x_{18} \cdot x_{501}$

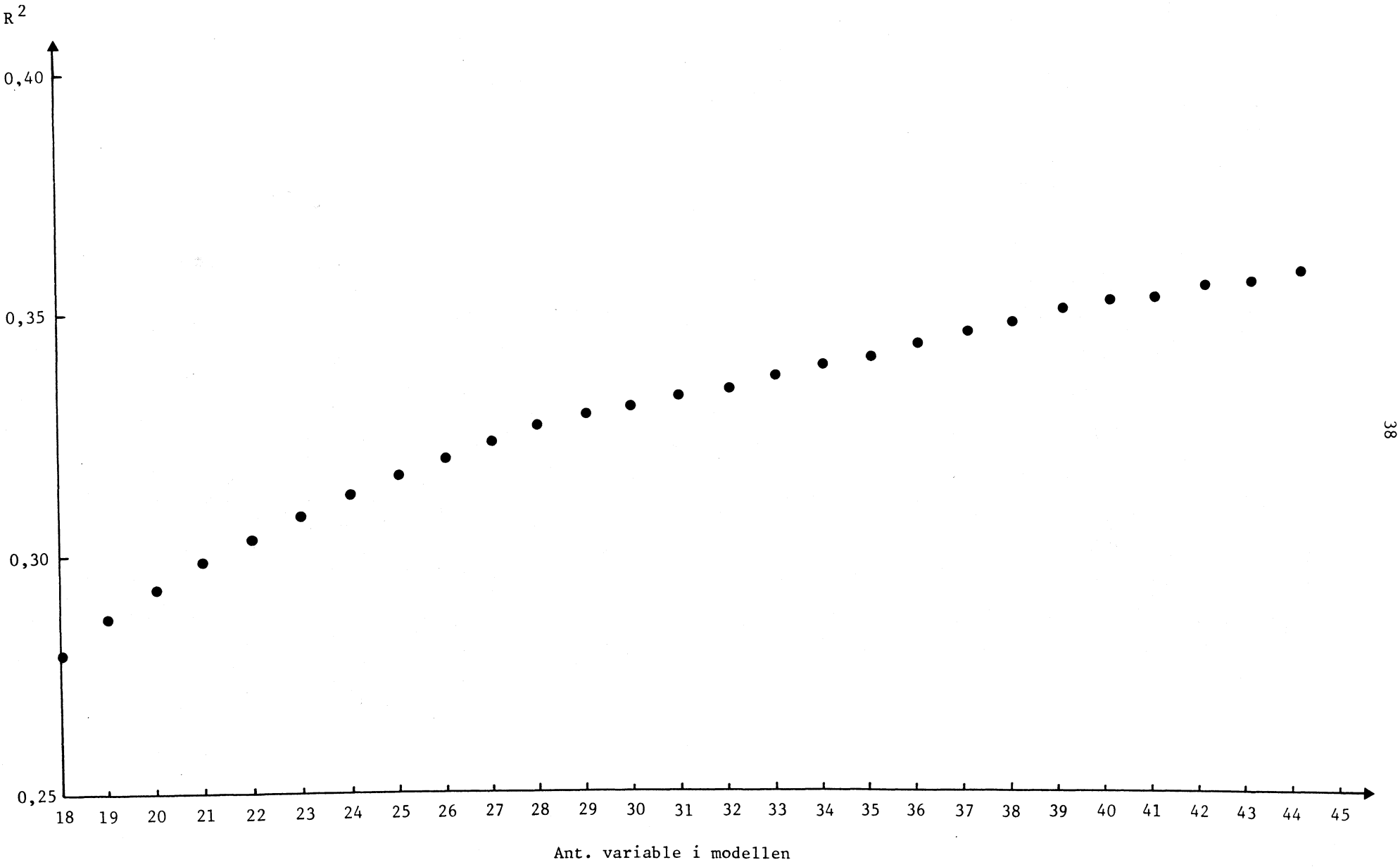
En indikasjon på at disse samspillene er relevante er at alle unntat K114 består av produkt av de effekter som i avsnitt 6.5 ble funnet som "sikkert" signifikante på 5 prosent nivå. Det var variablene

$$x_2, x_{17}, x_{33}, x_{34}, x_{35}, x_{121}, x_{501}$$

Alle disse, unntatt x_{35}, er med i de 8 samspillene.

Men samspillene er ikke signifikante på for eksempel 5 prosent nivå når en tar hensyn til at de er plukket ut blant 149. Antall samspill med forventning 0 ble estimert til 141. Dermed burde nivået .05/141 = .0004 brukes på de enkelte tester. Ingen av de 8 samspillene har såpass lav P-verdi. Dette fremgår av vedlegg 4 med regresjonsmodellen for 18 hovedeffekter og 8 samspill.

Figur 11. Trinnvis regresjon (MAXR) på samspillsledd. De 18 hovedeffektene er fast i modellen. R^2 som funksjon av antall variable i modellen



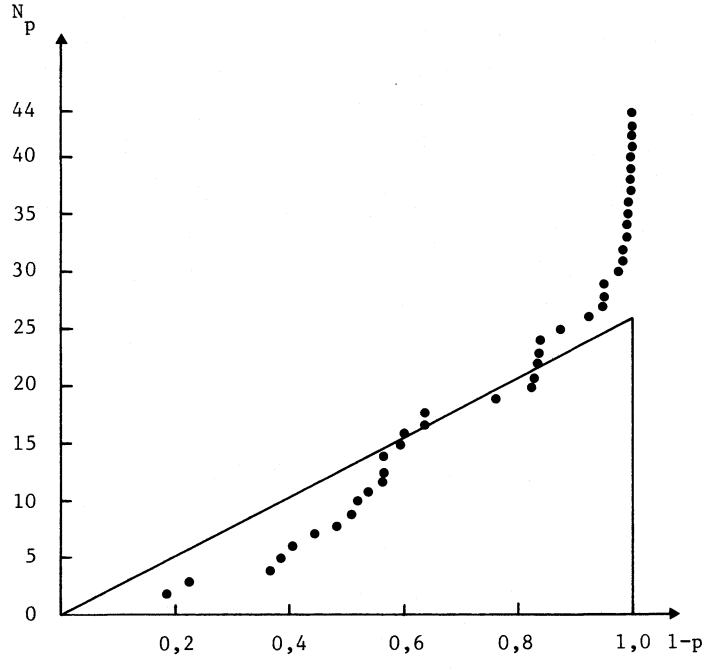
9. KOMBINERING AV FØRSTEGRADSLEDD OG KRYSSPRODUKTLEDD

9.1. Antall betydningsfulle ledd

På dette trinn i analysen står en igjen med førstegradsleddene, funnet som beskrevet i kapittel 6, og kryssproduktleddene fra kapittel 8. Det kan hende at noen av førstegradsleddene kunne ha blitt med på grunn av sin virkning gjennom kryssproduktleddene. Dermed kan de miste sin betydning i den større modellen i kapittel 8. For å anslå antall betydningsfulle ledd kan igjen et P-plott brukes. Ved vurderingen av P-plottet må en nå ta hensyn til at samspillseffektene ble valgt ut fordi de hadde små P-verdier. Dermed vil en uten videre ha et antall signifikante effekter som minst er lik antall samspillseffekter. Noe av det samme kan gjøre seg gjeldende for førstegradsleddene. De var også opprinnelig valgt ut på grunn av sin betydning. For å eliminere den siste effekten, lagde vi P-plott på grunnlag av alle opprinnelige forklaringsvariable (ikke bare de som ble tatt med for å studere samspill) sammen med de utvalgte samspillene. Dermed er det ingen utvelgelseseffekt for førstegradsleddene. Dette har også den fordel at en har mulighet til å oppdage variable som blir betydningsfulle først etter at samspillene har blitt med i modellen.

I figur 12 er P-plottet gitt for de 36 + 8 = 44 variable i Eksempelet. Anslaget for antall betydningsløse effekter ut fra den inntegnete linjen er 26. Totalt skulle det dermed være ca. 18 betydningsfulle variable. Det grovt anslåtte standardavviket ut fra p = 0.3 er 2.8.

Figur 12. P-plott for 36 hovedaffekter og 8 samspillsledd



9.2. Endelig modell

Ut fra anslaget av antall betydningsfulle ledd, trinnvise metoder og et R^2-plott kan en som tidligere komme frem til et forslag til endelig modell.

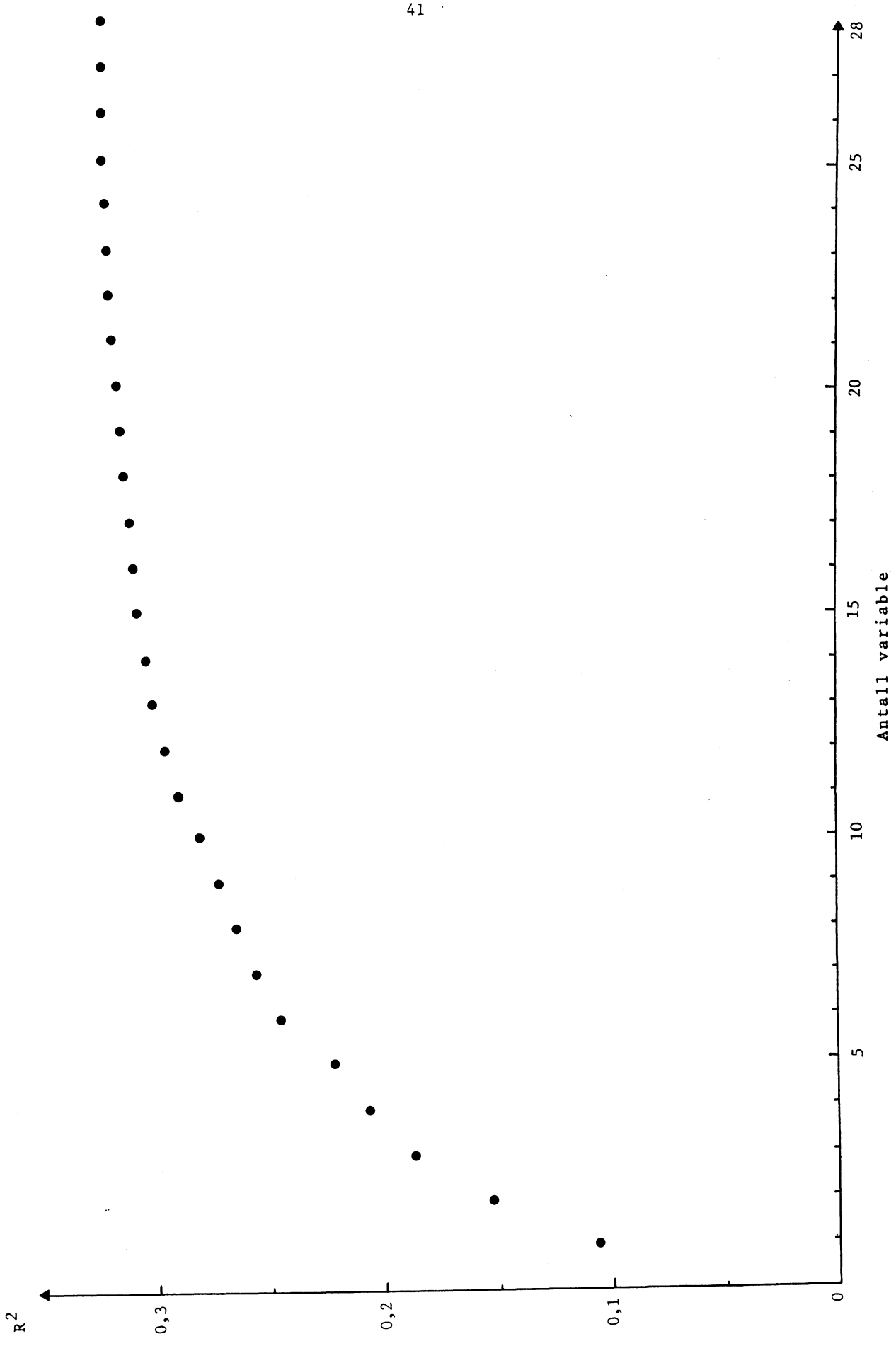
I figur 13 er R^2-plottet gitt for Eksemplet. Det ser ut til å endre stigningstakt ved 7, 13 og 21 variable. I betraktning av den relativt lille endringen i R^2 fra 13 variable og ut, vil vi foreslå å ta utgangspunkt i modellen med 13 variable. Det er riktignok et litt lite antall i forhold til estimatet 18 av antall signifikante variable, men likevel ikke urimelig. En bør også redusere antall noe på grunn av mulig seleksjonseffekt når samspillsvariablene ble valgt ut.

Begge de trinnvise prosedyrene BACKWARDS (med nivå 0,02 og MAXR ga de samme variable når 13 variable ble valgt. Estimatene for modellen er gitt i tabell 5.

Tabell 5. Regresjon på de 13 variable i modellen

Variabel	Regresjons-koeffisient	Standardavvik	Signifikans-sannsynlighet
Konstantledd	2,230		
x_{17}	0,348	0.0584	0.0001
x_{18}	0,046	0,0106	0,0001
x_{23}	0,175	0,0532	0,0011
x_{34}	0,210	0,0464	0,0001
x_{35}	0,202	0,0557	0,0003
x_{121}	0,826	0,1502	0,0001
K4 = $x_2 \cdot x_{33}$	0,364	0,0402	0,0001
K12 = $x_{17} \cdot x_{34}$	-0,270	0,0960	0,0050
K85 = $x_2 \cdot x_{15}$	-0,290	0,0808	0,0004
K121 = $x_{15} \cdot x_{33}$	0,292	0,0598	0,0001
K132 = $x_{25} \cdot x_{34}$	-0,540	0,1320	0,0001
K151 = $x_{28} \cdot x_{121}$	-0,555	0,1606	0,0006
K158 = $x_{18} \cdot x_{501}$	-0,064	0,0138	0,0001

$R^2 = 0,2993$



Figur 13. Plott av R^2 mot antall variable i modellen. 36 hovedeffekter og 8 samspill

9.3. Generelle kommentarer til modellen i Eksempelet

9.3.1. Skal en bruke hierarkisk modell?

Modellen i foregående avsnitt og estimert i tabell 5 er ikke hierarkisk i den forstand at den ikke inneholder førstegradsledd for alle de variablene som inngår i kryssproduktleddene. Variablene x_2, x_{15}, x_{25}, x_{28}, x_{33} og x_{501} inngår bare i samspillene. I modeller med kontinuerlige forklaringsvariable bruker en ofte automatisk å ta med førstegradsledd når ledd av høyere orden er med. Vi finner det ikke naturlig å gjøre dette i vår situasjon hvor variablene er dikotome.

9.3.2. Høyere ordens ledd

Hvis vi sr på samspillsleddene, ser vi at variablene x_2, x_{22}, x_{15}, x_{17}, x_{34} og x_{25} er relatert til hverandre på følgende måte (der forbindelsene med strek betyr samspill):

De andre variablene som er med i samspill er bare involvert i ett samspill.

Variablene x_2, x_{15} og x_{33} ser ut til å henge sammen på en komplisert måte. Det er derfor naturlig å prøve å ta inn trefaktorleddet $x_2 \cdot x_{15} \cdot x_{33}$. På tilsvarende måte er det naturlig å ta inn trefaktorleddet $x_{17} \cdot x_{25} \cdot x_{34}$. I regresjonen tok vi også inn hovedeffektene til de impliserte variablene. SAS avviste trefaktorleddet $x_{17} \cdot x_{25} \cdot x_{34}$ med den begrunnelse at det var en lineærkombinasjon av andre variable. Det førstnevnte trefaktorleddet hadde en signifikanssannsynlighet på ca. 0,07.

Trinnvis regresjon ga som resultat modellen i tabell 5, men med K85 = $x_2 \cdot x_{15}$ erstattet av $x_2 \cdot x_{15} \cdot x_{33}$. I den "nye" modellen var R^2 noe større enn R^2 i den "gamle" modellen, men forskjellen var ubetydelig. Vi valgte derfor å holde fast ved den "gamle" modellen siden den var enklere.

9.3.3. Intern estimering av feil

Som beskrevet i kapittel 3 kan vi sammenligne restvariansen i den endelige modellen med et internt estimat av variansen. Det interne estimatet baserte vi på en oppdeling etter dikotome variable (se kapittel 4.2) med hjelpevariabelen

$$HJELP = 16\, x_{17} + 8\, x_{23} + 4\, x_{34} + 2\, x_{35} + x_{121}$$

hvor de viktigste signifikante førstegradsledd inngår.

Innen de forskjellige gruppene for HJELP kjørte vi regresjon m.h.p. resten av de uavhengige variablene i den endelige modellen. Tabell 6 viser antall frihetsgrader og estimert varians i hver enkelt gruppe.

Tabell 6. Estimert varians innen forskjellige delgrupper

HJELP	Antall frihetsgrader	Estimert varians
0	295	0,3434
1	14	0,2929
2	65	0,3844
4	213	0,2743
10	67	0,2629
12	27	0,2761
16	47	0,2301
20	40	0,1783
26	20	0,2379

Det veiete gjennomsnitt av de estimerte variansene i gruppene i tabell 6 er 0,3002. Den endelige modellen (tabell 5) hadde estimert restvarians 0,3040. Forskjellen er liten. Det tyder på at det ikke er mulighet til vesentlig modellforbedring.

9.3.4. Gevinst i modelltilpasning ved å ta inn samspill

I den endelige modellen (tabell 5) inngår totalt 12 av forklaringsvariablene. Det er

x_2, x_{15}, x_{17}, x_{18}, x_{23}, x_{25}, x_{28}, x_{33}, x_{34}, x_{35}, x_{121}, x_{501}

Multippel korrelasjonskoeffisient er R^2 = .299. Den beste modellen med 12 førstegradsledd hadde R^2 = 0,265 (figur 5).

Dette er vesentlig forbedring, selv om den absolutte forbedring ikke er så stor. Det kan en ikke vente i vårt eksempel med relativt svake direkte sammenhenger mellom den avhengige og de uavhengige variable. I tillegg kommer at samspill i modellen burde gi bedre uttrykk for de faktiske sammenhenger.

10. FJERNING AV OBSERVASJONER MED STOR INNFLYTELSE OG REESTIMERING I EKSEMPELET

I eksempelet kjørte vi regresjon m.h.p. de 36 hovedeffektene, og listet ut Y, RSTUDENT, Cook's D og h_i (se kap. 7). Observasjoner med stor verdi på minst én av de 4 nevnte størrelsene ble undersøkt nærmere. Formålet med dette var å finne frem til observasjoner med stor innflytelse. Vi valgte å fjerne 3 observasjoner som vi fant "mistenkelige" eller svært avvikende.

Den første observasjonen vi fjernet fra utvalget, gjaldt en mann på 68 år med 124 ferieturdager. Et så stort antall feriedager var svært utypisk i og med at mannen var yrkesaktiv med vanlig virkedagsarbeid minst 5 dager i uka.

Den andre observasjonen ble fjernet fordi den var en dublett, dvs. at en og samme person forekom i datamaterialet 2 ganger.

Den tredje observasjonen som ble fjernet, gjaldt en enslig kvinne på 25 år med 83 ferieturdager. Igjen var det utypisk med så mange ferieturdager siden hun hadde fast arbeid med forholdsvis lav inntekt.

Dersom vi primært skulle analysere folks ferievaner, kunne det diskuteres om grunnene til å fjerne de 3 observasjonene var gode nok. Vårt utgangspunkt var imidlertid å studere hvorledes avvikende observasjoner kunne påvirke resultatene.

Vi gjentok analysen som er beskrevet tidligere i dette notatet. Først ble kjørt regresjon på de 36 hovedeffektene i materialet, og laget et P-plott for t-observatorene for regresjonskoeffisientene. I dette tilfellet ble antall signifikante hovedeffekter estimert til å være ca. 20. Videre kjørte vi trinnvis regresjon på de 36 variablene, og lagde et R^2 plott. Dette hadde et "knekkpunkt" ved 20 variable, og vi valgte da å beholde 20 variable. De 18 variablene som vi "valgte ut" i avsnitt 6.6, var alle blant disse 20. De to nye variable var x_{126} og x_{371}.

Neste trinn i analysen var å danne alle mulige kryssproduktledd av de 20 variablene. Disse ble så delt opp i 4 grupper. For hver gruppe ble kjørt regresjon med kryssproduktledd og de 20 hovedeffektene. For regresjonskoeffisientene lagde vi deretter et P-plott. Antall signifikante kryssproduktledd ble estimert til 3. Dette estimatet har imidlertid stor usikkerhet. Det omtrentlige standardavviket er 8,8 (jfr. avsnitt 8.2).

Deretter ble utført trinnvis regresjon m.h.p. kryssproduktleddene i hver av de 4 gruppene. De 20 hovedeffektene ble hele tiden "holdt fast" i modellene. De mest betydningsfulle kryssproduktledd ble så plukket ut fra de 4 gruppene. På de ca. 30 gjenværende kryssproduktledd og 20 hovedeffekter kjørte vi to trinnvise regresjoner (MAXR). I den ene regresjonen holdt vi de 20 hovedeffektene fast i modellen, og i den andre regresjonen lot vi hovedeffektene "konkurere" på like fot med kryssproduktleddene. Etter å ha studert plott av R^2 fra disse to regresjonene valgte vi å beholde 8 kryssproduktledd. Det var de 8 kryssproduktleddene som ble tatt inn først i den sistnevnte regresjonen.

De 8 kryssproduktleddene ble så kjørt i en regresjon sammen med de 36 hovedeffektene. Ved hjelp av P-plott ble antall signifikante effekter estimert til å være 18. Fra den trinnvise regresjonen på hovedeffekter og samspillsledd ble lagd et R^2-plott. Ut fra dette plottet valgte vi å ta med 4 hovedeffekter og 8 kryssproduktledd. Dette til tross for at P-plottet estimerte antall signifikante effekter til å være 18. De resterende effektene så nemlig ut til å ha liten betydning. Tabell 7 viser resultatet av regresjonsberegningene på den endelige modell.

Tabell 7. Parameterestimat, standardavvik og signifikanssannsynligheter i den "endelige modell"

Variabel	Koeffisient-estimat	Standardavvik	Signifikans-sannsynlighet
Konstantledd	2,250		
x_{17}	0,257	0,047	0,0001
x_{18}	0,046	0,010	0,0001
x_{34}	0,145	0,040	0,0003
x_{121}	0,571	0,090	0,0001
K4 = $x_2 \cdot x_{33}$	0,377	0,040	0,0001
K19 = $x_{23} \cdot x_{35}$	0,359	0,060	0,0001
K51 = $x_{10} \cdot x_{506}$	0,097	0,032	0,0028
K85 = $x_2 \cdot x_{15}$	-0,296	0,081	0,0003
K121 = $x_{15} \cdot x_{33}$	0,273	0,060	0,0001
K132 = $x_{25} \cdot x_{34}$	-0,490	0,130	0,0002
K152 = x_{121} x_{371}	-0,414	0,116	0,0004
K217 = $x_{11} \cdot x_{501}$	-0,013	0,003	0,0001

Ved denne modellen er R^2 = 0,2988 og av samme størrelsesorden som R^2 i modellen presentert i avsnitt 9.2. Alle de 4 hovedeffektene i tabell 7 er også hovedeffekter i modellen presentert i avsnitt 9.2. De to hovedeffektene i modellen i avsnitt 9.2 som ikke er med som hovedeffekter i den nye modellen, er begge involvert i samspillsledd i den nye modellen. I alt 4 av samspillsleddene er felles i de to modellene.

Selv om regresjonsfunksjonene har noe forskjellig form, er det likevel stort sett de samme variable som inngår i de to modellene. Den eneste variable i den "gamle" modellen som ikke inngår i den nye er X_{28} (virkedagsarbeid). Nye variable er X_{10} (kjønn), X_{11} (fødselsår) og X_{371} (innadvendte sosiale interesser). Når det gjelder x_{10} og x_{11} ville disse ha blitt tatt inn på henholdsvis trinn 18 og 16 i den trinnvise prosedyren anvendt på alle dataene. De var altså nær ved å være "signifikant". Det er sannsynlig at den første og tredje observasjonen som ble tatt ut, har bidratt til å øke betydningen av x_{10} og x_{11}. De representerte to ekstreme observasjoner for hvert kjønn, samt en ekstrem observasjon knyttet til alder.

Den samlete konklusjonen er likevel at de tre observasjonene later til å ha relativt liten innflytelse på resultatet av regresjonen.

11. REFERANSER

Belsley, Kuh and Welsch (1980): Regression Diagnostics. John Wiley & sons.

Cook, R.D. (1977): Detection of influential observations in linear regression. Technometrics, 19, 1977, s. 15-18.

Daniel, C. (1959): Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments. Technometrics, 1, No. 4, 1959, s. 311-341.

Daniel, C. and Wood, F.S. (1980): Fitting Equations to Data. Second edition. John Wiley & sons.

Haldorsen, T. (1981): Norske Ferieformer. Statistisk Sentralbyrå. Rapporter 81/25.

Mordal, T.L. (1979): Nordmenns feriereiser. Statistisk Sentralbyrå. Samfunnsøkonomiske studier nr. 41.

Schweder, T. og Spjøtvoll, E. (1982): Plots of P-values to evaluate many tests simultaneously, Biometrika, 1982, 69, 3, s. 493-502.

Statistisk Sentralbyrå (1975): Ferieundersøkelsen 1974. Norges offisielle statistikk A732.

Zirphile, J. (1975): Letter to the Editor. Technometrics, 1975, 17, No. 1, s. 145.

Vedlegg 1

LISTE OVER UAVHENGIGE VARIABLE I GJENNOMGANGSEKSEMPELET

x_2 : 1 dersom personen eide/disponerte fritidshus
0 ellers

x_3 : 1 dersom personen eide/disponerte personbil
0 ellers

x_4 : 1 dersom personen eide/disponerte båt
0 ellers

x_5 : Tallet på ganger (dager) personen utøvde friluftsaktiviteter i løpet av året

x_6 : Tallet på ganger (dager) personen utøvde idretts- og mosjonsaktiviteter i løpet av året

x_7 : Indeks for fysisk rørlighet (verdier: 1, 2, 3, 4)

x_{10}: 1 dersom personen er mann
2 dersom personen er kvinne

x_{11}: Fødselsår

x_{12}: Husholdningsstørrelse (verdier: 1-6)

x_{13} : 1 dersom personen er bosatt i Nord-Norge
0 ellers

x_{14} : 1 dersom personen er bosatt i Trøndelag
0 ellers

x_{15} : 1 dersom personen er bosatt på Vestlandet
0 ellers

x_{16} : 1 dersom personen er bosatt på Sørlandet
0 ellers

x_{17} : 1 dersom personen bor i hus med flere enn 4 leiligheter (blokk, leiegård e.l.)
0 ellers

x_{18} : Husholdningsinntekt i 10 000 kroners intervall (verdier: 1, 2, ..., 8)

x_{20} : 1 dersom personen var ansatt
0 ellers

x_{21} : 1 dersom personen var selvstendig uten leid hjelp
0 ellers

x_{22} : 1 dersom personen var selvstendig med leid hjelp
0 ellers

x_{23}: 1 dersom personen hadde teknisk, vitenskapelig, humanistisk arbeid, administrasjon som yrke
0 ellers

x_{24}: 1 dersom personen hadde kontor- eller handelsarbeid som yrke
0 ellers

x_{25}: 1 dersom personen hadde jordbruk, skogbruk, eller fiske som yrke
0 ellers

x_{26}: 1 dersom personen hadde industri, bygge- og anleggsarbeid som yrke
0 ellers

x_{27}: 1 drsom personen hadde transportarbeid som yrke
0 ellers

x_{28}: 1 dersom personen hadde vanlig virkedagsarbeid
0 ellers

x_{29}: 1 dersom personen hadde fri hver lørdag
0 ellers

x_{30}: 1 dersom personen arbeidde minst 5 dager i uken og mindre enn 9 timer[1] hver dag
0 ellers

x_{31}: 1 dersom personen arbeidde minst 5 dager i uken og 9 timer[1] eller mer hver dag
0 ellers

x_{32} : 1 dersom personen arbeidde mindre enn 5 dager i uken
0 ellers

x_{33} : 1 dersom personen bodde i tettbygd strøk
0 ellers

x_{34} : 1 dersom personen hadde utdanning på gymnasnivå
0 ellers

x_{35} : 1 dersom personen hadde utdanning på universitets- eller høgskolenivå
0 ellers

x_{36} : Indikator for utadvendte sosiale interesser for sommerferien (verdier: 4, 5, ..., 12)

x_{37} : Indikator for innadvendte sosiale interesser for sommerferien (verdier: 2, 3, ..., 6)

x_{38} : Indikator for interesse i friluftsaktiviteter for sommerferien (verdier: 4, 5, ..., 12)

---

[1] Arbeidstid + reisetid til/fra arbeidssted.

Vedlegg 2

REGRESJON MED ALLE 36 VARIABLE. UTSKRIFT FRA SAS.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	36	116.586	3.238497	10.249	0.0001
ERROR	887	280.276	0.315981		
C TOTAL	923	396.861			

ROOT MSE	0.562122	R-SQUARE	0.2938
DEP MEAN	2.848848	ADJ R-SQ	0.2651
C.V.	19.73157		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > !T!
INTERCEP	1	1.931073	0.218646	8.832	0.0001
X2	1	0.235006	0.040248	5.839	0.0001
X3	1	-0.016080	0.051412	-0.313	0.7545
X4	1	0.085044	0.042057	2.022	0.0435
X7	1	-0.030379	0.031111	-0.976	0.3291
X10	1	0.124669	0.049077	2.540	0.0112
X11	1	-0.00291466	0.001599078	-1.823	0.0687
X13	1	0.079964	0.070995	1.126	0.2603
X14	1	0.184027	0.086154	2.136	0.0330
X15	1	0.103948	0.048798	2.130	0.0334
X16	1	-0.042301	0.080883	-0.523	0.6011
X17	1	0.219919	0.051663	4.257	0.0001
X18	1	0.028417	0.012325	2.306	0.0214
X20	1	0.244596	0.131449	1.861	0.0631
X21	1	0.241354	0.140792	1.714	0.0868
X22	1	0.221391	0.149753	1.478	0.1397
X23	1	0.164903	0.071662	2.301	0.0216
X24	1	0.037510	0.061611	0.609	0.5428
X25	1	-0.139294	0.106149	-1.312	0.1898
X26	1	0.052699	0.065201	0.808	0.4192
X27	1	-0.020506	0.106550	-0.192	0.8474
X28	1	-0.119482	0.059752	-2.000	0.0458
X29	1	0.106250	0.047028	2.259	0.0241
X30	1	-0.044493	0.079095	-0.563	0.5739

S T A T I S T I C A L A N A L Y S I S S Y S T E M

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > !T!
X31	1	-0.044303	0.073794	-0.600	0.5484
X32	1	-0.106701	0.123060	-0.867	0.3861
X33	1	0.260545	0.058944	4.420	0.0001
X34	1	0.165806	0.047153	3.516	0.0005
X35	1	0.241047	0.061474	3.921	0.0001
X38	1	-0.00939949	0.010192	-0.922	0.3567
X601	1	0.018487	0.043859	0.422	0.6735
X121	1	0.310531	0.069439	4.472	0.0001
X126	1	-0.096445	0.080259	-1.202	0.2298
X361	1	0.058199	0.064017	0.909	0.3635
X371	1	-0.062254	0.042368	-1.469	0.1421
X501	1	-0.241294	0.074867	-3.223	0.0013
X506	1	0.069191	0.046016	1.504	0.1330

ET EKSEMPEL PÅ EN SAS-UTSKRIFT MED UTLISTING AV EGENVERDIER, KONDISJONERINGSINDEKSER, ANDELEN AV VARIANSEN KNYTTET TIL HVER EGENVEKTOR, TOLERANCE OG VIF.

DEP VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	7	56.663255	8.094751	21.648	0.0001
ERROR	972	363.460	0.373930		
C TOTAL	979	420.124			

ROOT MSE	0.611490	R-SQUARE	0.1349
DEP MEAN	2.851403	ADJ R-SQ	0.1286
C.V.	21.44553		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|	TOLERANCE	VARIANCE INFLATION
INTERCEP	1	2.475038	0.101916	24.286	0.0001	.	0.000000
X2	1	-0.341903	0.227892	-1.500	0.1339	0.029432	33.976616
X3	1	-0.056958	0.128915	-0.442	0.6587	0.131454	7.607211
X33	1	0.335440	0.118102	2.840	0.0045	0.200246	4.993845
K1	1	0.422226	0.254884	1.657	0.0979	0.024493	40.828106
K2	1	0.597299	0.246501	2.423	0.0156	0.025860	38.669526
K3	1	-0.00639263	0.146268	-0.044	0.9651	0.079157	12.633074
K4	1	-0.318158	0.275899	-1.153	0.2491	0.022248	44.947954

COLLINEARITY DIAGNOSTICS — VARIANCE PROPORTIONS

NUMBER	EIGENVALUE	CONDITION INDEX	PORTION INTERCEP	PORTION X2	PORTION X3	PORTION X33	PORTION K1	PORTION K2	PORTION K3	PORTION K4
1	6.260	1.000	0.0007	0.0003	0.0006	0.0006	0.0003	0.0003	0.0005	0.0003
2	1.037	2.457	0.0050	0.0018	0.0034	0.0042	0.0018	0.0018	0.0030	0.0018
3	0.323934	4.396	0.0164	0.0050	0.0140	0.0140	0.0044	0.0053	0.0128	0.0043
4	0.243155	5.074	0.0206	0.0071	0.0156	0.0179	0.0075	0.0071	0.0153	0.0069
5	0.071791	9.338	0.0669	0.0341	0.0457	0.0319	0.0259	0.0259	0.0374	0.0367
6	0.046927	11.550	0.1033	0.0345	0.0612	0.1229	0.0489	0.0472	0.0388	0.0371
7	0.013487	21.545	0.4352	0.0785	0.4307	0.4514	0.0369	0.0573	0.5128	0.0236
8	0.003590	41.760	0.3518	0.8387	0.3688	0.3571	0.8743	0.8551	0.3795	0.8894

Vedlegg 4

REGRESJON PÅ 18 HOVEDEFFEKTER OG 8 SAMSPILL. UTSKRIFT FRA SAS.

DEP VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	26	130.184	5.007069	16.747	0.0001
ERROR	923	275.956	0.298977		
C TOTAL	949	406.140			

ROOT MSE	0.546788	R-SQUARE	0.3205
DEP MEAN	2.847062	ADJ R-SQ	0.3014
C.V.	19.20535		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > !T!
INTERCEP	1	2.213256	0.121885	18.159	0.0001
X2	1	0.001828433	0.093964	0.019	0.9845
X4	1	0.070644	0.040377	1.750	0.0805
X10	1	0.091542	0.040512	2.260	0.0241
X11	1	-0.00312056	0.0014045	-2.222	0.0265
X14	1	0.105795	0.081757	1.294	0.1960
X15	1	-0.061948	0.109973	-0.563	0.5734
X17	1	0.327023	0.059181	5.526	0.0001
X18	1	0.036812	0.011515	3.197	0.0014
X23	1	0.155804	0.053317	2.922	0.0036
X25	1	-0.184668	0.105672	-1.748	0.0809
X28	1	-0.099980	0.055271	-1.809	0.0708
X29	1	0.094905	0.042381	2.239	0.0254
X33	1	0.038503	0.079250	0.486	0.6272
X34	1	0.223049	0.048412	4.607	0.0001
X35	1	0.217579	0.056326	3.863	0.0001
X121	1	0.716997	0.156958	4.568	0.0001
X501	1	0.102720	0.178486	0.576	0.5651
X506	1	0.072346	0.042883	1.687	0.0919
K4	1	0.327718	0.100926	3.247	0.0012
K12	1	-0.249240	0.095729	-2.604	0.0094
K85	1	-0.262555	0.087694	-2.994	0.0028
K114	1	0.513919	0.205044	2.506	0.0124
K121	1	0.344392	0.116991	2.944	0.0033

S T A T I S T I C A L A N A L Y S I S S Y S T E M

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > !T!
K132	1	-0.364350	0.161869	-2.251	0.0246
K151	1	-0.449021	0.168580	-2.664	0.0079
K158	1	-0.086400	0.036922	-2.340	0.0195

Utkommet i serien Rapporter fra Statistisk Sentralbyrå (RAPP) - ISSN 0332-8422

Trykt 1984

Nr. 84/1 Naturressurser og miljø 1983 Foreløpige nøkkeltall fra ressursregnskapene for energi, mineraler, skog, fisk og areal Sidetall 100 Pris kr 18,00 ISBN 82-537-1993-0

\- 84/2 Torstein Bye: Energisubstitusjon i næringssektorene i en makromodell Sidetall 47 Pris kr 12,00 ISBN 82-537-2042-4

\- 84/4 Jon Åge Vestøl: Kommunale avfallsbehandlingsanlegg Miljøstandard Sidetall 78 Pris kr 18,00 ISBN 82-537-2062-9

\- 84/5 Bjørg Moen: Bibliography of Population Studies in Norway Bibliografi over befolkningsstudier i Norge Sidetall 114 Pris kr 18,00 ISBN 82-537-2045-9

84/6 Grete Dahl: Folketrygden. Korttidsytelser og stønad ved yrkesskade Sidetall 26 Pris kr 12,00 ISBN 82-537-2069-6

\- 84/7 Tiril Vogt: Social Indicators and Environmental Dimensions Sidetall 33 Pris kr 12,00 ISNB 82-537-2060-2

\- 84/8 Otto Carlsen: Pasientstatistikk 1982 Statistikk fra Det økonomiske og medisinske informasjonssystem Sidetall 61 Pris kr 18,00 ISBN 82-537-2066-1

\- 84/9 Herdis Thorén Amundsen: Statistiske metoder for analyse av samvariasjon i kategoriske data Sidetall 228 Pris kr 24,00 ISBN 82-537-2074-2

\- 84/10 Audun Rosland: Vannkraftutbygging - Reguleringsinngrep - Virkninger på fisk Sidetall 127 Pris kr 18,00 ISBN 82-537-2102-1

\- 84/11 Skatter og overføringer til private Historisk oversikt over satser mv. Årene 1970 - 1984 Sidetall 75 Pris kr 18,00 ISBN 82-537-2081-5

\- 84/12 Arne Faye og Helge Herigstad: Friluftsliv i Norge 1970 - 1982 Sidetall 77 Pris kr 18,00 ISBN 82-537-2092-0

\- 84/13 Jon Paschen Knudsen: Boligstandard Variasjoner innen og mellom byer Sidetall 66 Pris kr 18,00 ISBN 82-537-2088-2

\- 84/14 Erling Siring og Emil Spjøtvoll: Regresjonsanalyse med et stort antall variable Sidetall 55 Pris kr 18,00 ISBN 82-537-2122-6

\- 84/15 Sindre Børke: Folke- og boligtelling 1980 Dokumentasjon Sidetall 211 Pris kr 24,00 ISBN 82-537-2112-9

\- 84/16 Stein Opdahl: Aleneforeldres levekår og tidsbruk Sidetall 188 Pris kr 18,00 ISBN 82-537-2127-7

\- 84/17 Alette Schreiner og Tor Skoglund: Virkninger av oljevirksomhet i Nord-Norge Sidetall 43 Pris kr 18,00 ISBN 82-537-2118-8

\- 84/18 Morten Reymert: Import- og eksportlikninger i KVARTS Utledning, estimering og simulering med likninger for utenrikshandelen Sidetall 83 Pris kr 18,00 ISBN 82-537-2123-4

\- 84/20 Arne Ljones: Energiundersøkelsen 1983 Om energibruk og energiøkonomisering i private husholdninger Sidetall 62 Pris kr 18,00 ISBN 82-537-2130-7

\- 84/21 Johan Heldal: Kvalitetskontrollundersøkelsen for Folke- og boligtellingen 1980 Sidetall 115 Pris kr 18,00 ISBN 82-537-2140-4

\- 84/22 Sindre Børke: Tilleggsundersøkelsen til Folke- og boligtelling 1980 Om muligheter for å erstatte skjema med registeropplysninger i senere folke- og boligtellinger Sidetall 61 Pris kr 18,00 ISBN 82-537-2136-6

\- 84/23 Roar Bergan: MINK En finansiell ettermodell til MSG En MSG-rapport Sidetall 71 Pris kr 18,00 ISBN 82-537-2138-2

\- 84/24 Yngvar Holm: Engrosomsetningsindeks Sidetall 19 Pris kr 12,00 ISBN 82-537-2141-2

\- 84/25 Morten Jensen og Morten Reymert: Kvartalsmodellen KVARTS-modellbeskrivelse og teknisk dokumentasjon Sidetall 87 Pris kr 18,00 ISBN 82-537-2139-0

Utkommet i serien Rapport fra Statistisk Sentralbyrå (RAPP) - ISSN 0332-8422 (forts.)

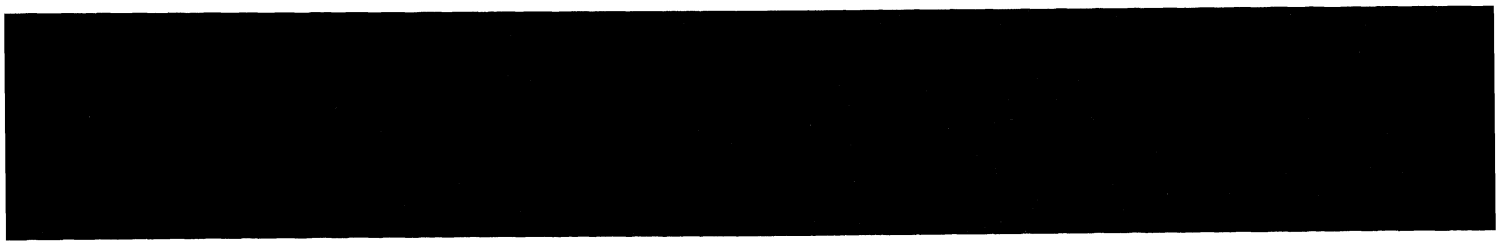
Trykt 1985

- 85/1 Naturressurser og miljø 1984 Foreløpige nøkkeltall fra ressursregnskapene for miljø, energi, mineraler, skog, fisk og areal Sidetall 94 Pris kr 30,00 ISBN 82-537-2133-1

- 85/2 Aktuelle skattetall 1984 Current Tax Data Sidetall 44 Pris kr 20,00 ISBN 82-537-2142-0

- 85/4 Lorents Lorentsen og Kjell Roland: Marked for råolje. Historisk utvikling. Teorier og modeller. Prisprognoser Sidetall 50 Pris kr 20,00 ISBN 82-537-2145-5

- 85/6 Referansearkiv for naturressurs- og forurensningsdata: Emnekatalog for ferskvann Sidetall 313 Pris kr 50,00 ISBN 82-537-2159-5



Pris kr. 18,00

Publikasjonen utgis i kommisjon hos H. Aschehoug & Co. og Universitetsforlaget, Oslo, og er til salgs hos alle bokhandlere.

ISBN 82-537-2122-6
ISSN 0332-8422