Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 68/1

Oslo, 19. januar 1968

Orientering for bruk av SSB's regresjonsprogram. 45 variable, dobbel presisjon

a v

Ingar Holme, Ib Thomsen, Tor Halvorsen

I N N H O L D

1. Generell beskrivelse

a. Programmet er en modifikasjon og utvidelse av IBM's program: Multiple Linear Regression. Dette er dog bare en spesiell type regresjonsprogram. Det finnes andre varianter, hvor man f.eks. begrenser output.

b. Programmet estimerer koeffisientene i en likning av formen:

$$(1)\quad Y = A + B_1X_1 + B_2X_2 + \ldots\ldots\ldots B_nX_n + \varepsilon$$

for et gitt tallsett $(Y_i, X_{1i}, X_{2i} \ldots\ldots X_{ni})$

$i = 1,2 \ldots\ldots,p$, der p er antall observasjonssett.

(1) er lineær i $B_1, B_2 \ldots\ldots B_n$. $X_1, X_2 \ldots\ldots X_n$ kan være transformerte variable.

c. Ofte er det av interesse å utføre analyse på forskjellige sett av variable. Programmet gir mulighet for dette. Enhver variabel kan velges som den avhengige. Utskillingen av den avhengige variable og utelatelse av uavhengige variable må spesifiseres i parameterkortene (se Regresjonskortene).

d. Formatet av input skal spesifiseres i subrutine data.

Begrensninger

Maksimalt antall variable er 45.
Maksimalt antall observasjonssett er 99 999. Minimalt antall observasjonssett er 2 større enn antall variable som inngår i regresjonsstigningen.

2. Beskrivelse av programmet

Regresjonsprogrammet består av en hovedrutine ved navn REGRE, en spesiell subrutine DATA, samt fem subrutiner, CORRE, ORDER, MINV, MULTR, GRUPPE, som alle nå ligger i maskinens disk, slik at den eneste subrutine vi må forandre hver gang, er subrutine DATA, idet denne tilpasses hvert oppdrag. Dette gjelder for programmet i dets opprinnelige form. I tillegg til dette har man nå, hvis restleddene i regresjonen ønskes, muligheten til å kalle inn fra disk et program som tegner histogram over restleddene. (Nærmere om dette i Vedlegg C.)

Da dette programmet bør kunne brukes uten særlig kjennskap til programmering, skal her i første omgang gis en rettledning til programmering av subrutine DATA.

For bruk av regresjonsprogrammet forutsettes en viss kontakt med Systemkontoret. Det første man må få fatt i, er en filebeskrivelse. Dette gjelder uansett om data er gitt på bånd eller på hullkort. For å få denne filebeskrivelsen utlevert trengs følgende 3 opplysninger: 1) Statistikk-nummer, 2) Navn på JOB, 3) Prosjektnummer. Disse opplysninger fås hos forværelsesdamen på Systemkontoret. Filebeskrivelsen skal hjelpe oss til å finne hvor i båndet eller hullkortene hver opplysning finnes. En file består av records som består av felter som hvert igjen består av et bestemt antall posisjoner, f.eks. 8. Dette antall varierer som regel litt fra felt til felt. I filebeskrivelsen er feltene og posisjonen alle nummererte; posisjonene nummereres fortløpende f.eks. 1 ..... 8,9 .... 16,17 ..... 255.

I en regresjonsanalyse er vi interessert i kvantitative variable som f.eks. sysselsetting, investering, bruttoproduksjon, realkapital, formue etc., etc. Disse variable finnes på båndet (kortene) og opptar hver et bestemt felt. Vår oppgave er nå følgende: Plukk ut de variable som interesserer og finn ut nøyaktig hvilke posisjoner disse variable har; f.eks. finnes at sysselsetting har posisjonene 161-168. Ofte er en også interessert i visse grupperinger, f.eks. etter næring, kommune, geografiske kriterier etc., etc. Behandlingen av slike kvalitative variable som hver antar diskrete verdier er helt analog, dog vil de oppta et annet (som regel mindre) antall posisjoner. Her må man skaffe seg en kodeliste over grupperingene. For å få fatt i denne trengs samme opplysninger som når det gjaldt filebeskrivelsen. (Nærmere herom i Vedlegg B). I tillegg trengs opplysning om båndets blokk-faktor, som står bak filebeskrivelsen. Hvis denne faktor er 1, er alt vel og bra, men hvis den har et høyere tall skaper dette visse komplikasjoner som en i tilfelle bør få hjelp av Systemkontoret til å løse. Blokkfaktor 1 betyr at filen er ordnet i adskilte observasjonssett.

Nummerer nå de variable, D(1), D(2) ..... D(N) der N er tallet på variable som tas ut av filebeskrivelsen. Disse kalles originale variable. Når dette er gjort, går man til Systemkontoret og ber om å få skrevet et såkalt FORMAT-kort, ved hjelp av de opplysninger som er tatt ut av filebeskrivelsen. Man bruker samme FORMAT-kort så lenge man opererer med samme data og samme originalvariable. En trenger altså ikke nødvendigvis skifte dette kortet ut for hver gang man kjører regresjon. I forbindelse med FORMAT-kortet bes også om å få skrevet det tilhørende READ-kortet.

Videre vil det ofte være av interesse å transformere de kontinuerlige variable til nye variable, eksempelvis log D(1), exp D(2), $D(3)^2$, $\sqrt{D(4)}$ etc. etc. (Se Vedlegg B om programmering av transformasjoner). Disse nye

transformerte variable kan nummereres fra D(N + 1) .....D(N + 2) ..... etc., men det er av og til slik at vi bare har bruk for de p første (p < N) originale variable i regresjonen, mens vi bare har bruk for transformasjoner av de N - p resterende variable. I så fall kan vi begynne nummereringen av de transformerte variable allerede på nr. p + 1, f.eks. D(1)(p + 1) = D LOG D(p). Dette gjelder helt analogt også for sammensatte transformasjoner. Det er viktig å være klar over dette, da vi har den relativt strenge innskrenkning at vi ikke må overstige 45 definerte variable. Når nå alle variable er definert (kontinuerlige som diskrete), er subrutine DATA ferdigprogrammert. (Se Vedlegg A).

3. Input

a) Kontrollkort (Utfylles av Systemkontoret). Vedlegg A.

b) Parameterkort

i) Problemkort

Ethvert problem må begynne med et problemkort, som leses av subrutine REGRE. Kortet må fylles ut på følgende måte:

Kol.	1 - 4	Problemnavn (kan være bokstaver eller tall),
Kol.	5 - 6	Problem nr. (må være tall. For bruk av nr. 00 se *).
Kol.	7 - 11	Antall observasjoner (sett)
Kol.	12 - 13	Antall variable
Kol.	14 - 15	Antall regresjonskort (se nedenfor)
Kol.	16	1 Hvis gjennomsnittene ønskes i output 0 " " ikke ønskes som output
Kol.	17	Som kol. 16 for standardavvik for hver enkelt variabel
Kol.	18	Som kol. 16 for sentrale momentmatriser
Kol.	19	Som kol. 16 for simple korrelasjonskoeffisienter mellom de variable
Kol.	20	Som kol. 16 for summatrise
Kol.	21	" " 16 for sum av kryssprodukt
Kol.	22	" " 16 for invertert korrelasjonsmatrise

ii) Datakort (Hvis altså data befinner seg på kort)

iii) Regresjonskort

På regresjonskortet skal angis en avhengig variabel samt alle uavhengige variable i en likning. Enhver innlest variabel kan

---

* Hvis en deler opp data i flere grupper og ønsker regresjon på hver gruppe samt totalen i en kjøring, brukes nr. 00 på totalen. For øvrig brukes nr. 00 ikke. OBS: Totalen må komme sist! (Nr. 00 kan bare brukes når data er på tape).

velges som den avhengige. Ved å lage flere regresjonskort for hvert problemkort kan brukeren estimere flere likninger for samme originale data. Kortet utfylles på følgende måte:

Kol.	1 - 2	01 Hvis restledd ønskes som output 00 " " ikke ønskes som output
Kol.	3 - 4	Nummer på den avhengige variable*)
Kol.	5 - 6	Antall uavhengige variable
Kol.	7 - 8	Nummer på første uavhengige variable
Kol.	9 - 10	Nummer på annen uavhengige variable
		osv. Se Vedlegg A.

iv) Subrutine DATA

4. Output

Programmet kan gi følgende output:

1. Sum av kryssprodukt
2. Sum-matrise
3. Gjennomsnittene
4. Standard avvik
5. Sum av kryssprodukt fra middeltallene
6. Korrelasjonsmatrisen
7. Invertert korrelasjonsmatrise
8. Korrelasjonskoeffisienten mellom den avhengige og hver enkelt uavhengig variabel
9. Regresjonskoeffisientene
10. Standardavviket for regresjonskoeffisientene
11. t-verdien
12. Konstantleddet (intercept)
13. Multippel korrelasjonskoeffisient
14. Standardavvik på estimatet (residualspredning)
15. F-verdien med de tilhørende kvadratiske former
16. Restledd (Hvis ønsket på regresjonskortet)
17. Durbin-Watson's observator (- " -)
18. X^2-test på normalitet (XJIKVA) (Hvis ønsket på regresjonskortet)

---

* Variablene nummereres i den rekkefølge de står i på kortet eller magnetbåndet fra venstre mot høyre.

Som nevnt ovenfor, er det her mulig å få tegnet et histogram over restleddene, hvis disse ønskes. Dette gjøres ved hjelp av et spesielt program. Her må man selv fylle ut en del parameterkort. (Se nærmere herom i Vedlegg C).

5. Feilutskrifter

Visse feil i parameterkort og data gir feilutskrifter, så brukeren kan foreta de nødvendige rettelser.

(1) Hvis antall regresjonskort ikke er spesifisert på problemkortet, fås følgende utskrift:
NUMBER OF SELECTIONS NOT SPECIFIED. JOB TERMINATED

(2) Hvis korrelasjonsmatrisen er singulær, fås følgende utskrifter:
THE MATRIX IS SINGULAR. THIS SELECTION IS SKIPPED
Den første feiltype resulterer i at programmet går til STOP, mens den annen feiltype resulterer i at beregningene fortsetter for neste regresjonskort.

6. Logiske enheter (For Systemkontoret)

a) Logisk enhetsnr. 1 brukes for parameterkort (Automatisk assign til kortleseren)
b) Logisk enhetsnr. 3 brukes for output (Automatisk assign til skriveren)
c) Logisk enhetsnr. 13 brukes for MELIOMLAGER
d) Logisk enhetsnr. 5 brukes for INPUT-DATA
e) Logisk enhetsnr. 7 brukes for RESTLEDD
f) Logisk enhetsnr. 6 brukes for restledd til HIST og RAM-programmet.

7. Et eksempel

a) Problem

Input-data finnes på kort.
Antall observasjoner er 30.
Antall variable er 6.
Vi ønsker å estimere koeffisientene i en likning av formen

$$V(6) = A.V(1) + B.V(2) + C.V(3) + D.V(4) + E.V(5)$$

og en annen likning av formen

$$V(6) = A.V(2) + B.V(3) + C.V(5)$$

hvor V(i) betegner variabel nummer i.
Vi ønsker restleddene som output.

b) Utfylling av parameterkort:

1) Problemkort:

Kol. 1 - 6	Eks. 01
Kol. 7 - 11	00030
Kol. 12 - 13	06
Kol. 14 - 15	02

2) Regresjonskort 1

Kol. 1 - 2	01
Kol. 3 - 4	06
Kol. 5 - 6	05
Kol. 7 - 8	01
Kol. 9 - 10	02
Kol. 11 - 12	03
Kol. 13 - 14	04
Kol. 15 - 16	05

3) Regresjonskort 2

Kol. 1 - 2	01
Kol. 3 - 4	06
Kol. 5 - 6	03
Kol. 7 - 8	02
Kol. 9 - 10	03
Kol. 11 - 12	05

8. Noen sluttbemerkninger

Når man skal programmere som vist i Vedlegg B, bør man alltid påse at det bare står ett symbol i hver rubrikk (kolonne), for ellers kan dette lett forvirre punchedamene. Av samme grunn bør man også skrive tydelig, slik at man unngår feilpunch. Hvis slike forekommer, bør en selv punche om de kortene som er feil. Hvis man ikke vet hvordan man puncher, tar dette ca. 1/2 time å lære, noe enhver som driver regresjonskjøring bør gjøre.

KONTROLLKORT FOR REGRESJONSPROGRAMMET (fylles ut av Systemkontoret)

1. JOB-kort

Kol.	1,2	//
"	3	Blank
"	4-6	JOB
"	7	Blank
"	8-15	Valgfritt navn

2. OPTION-kort

Kol.	1,2	//
"	3	Blank
"	4-9	OPTION
"	10	Blank
"	11-14	LINK

3. INCLUDE-kort

Kol.	1,3	Blank
"	4-10	INCLUDE
"	11	Blank
"	12-16	REGRE

4. EXEC FORTRAN-kort

Kol.	1,2	//
"	3	Blank
"	4-7	EXEC
"	8	Blank
"	9-15	FORTRAN

5. Subrutine DATA

Kort nr.:

1. SUBROUTINE DATA (M, D, IPR)
2. DIMENSION D(1)

Etter disse to kortene kommer kort for innlesning og transformasjoner av data, samt FORMAT kort og READ kort. (Se ovenfor og Vedlegg B).

3. Siste kort (bare nødvendig ved restleddbehandling):

   WRITE (13) (D(I), I = 1, M)

   Nest siste kort:

   RETURN, og til slutt

   END.

6. End of data - kort:

Kol.	1	/
"	2	*

7. LNKEDT-kort

Kol.	1,2	//
"	3	Blank
"	4-7	EXEC
"	8	Blank
"	9-14	LNKEDT

8. ASSGN-kort

Formålet med disse kort er å opplyse maskinen om hvor innlesning og utlesning foregår.

Kortene fylles ut etter de til enhver tid gjeldende regler.

Eventuelt: kontakt Driftskontoret.

Hvis restledd ønskes, fås følgende ASSGN-kort:

```
// ASSGN        SYS003, X'184'
//   "          SYS010, X'183'
//   "          SYS004, X'NNN'
//   "          SYS002, X'MMM'
```

Hvis restledd ønskes på listen: NNN = 00E

" " " " bånd: NNN = nr. på båndstasjon (f.eks. 185)

Hvis data er på kort: MMM = 014

" " " " bånd: MMM = nr. på båndstasjon (f.eks. 182)

9. EXEC-kort

Kol.	1,2	//
"	3	Blank
"	4-7	EXEC

10. Sannsynlighetskort

Dette kort angir de teoretiske sannsynligheter for å falle i de bestemte grupper under X^2-testing av normalitet.

Kol.	1	Blank
"	2-10	0.0013499
"	11	Blank
"	12-20	0.0048598
"	21	Blank
"	22-28	0.01654

Kol.	29-31	Blank
"	32-39	0.044057
"	40-41	Blank
"	42-49	0.091253
"	50-51	Blank
"	52-58	0.14998
"	59-61	Blank
"	62-68	0.19146

11. Parameterkort og DATA-kort

OBS! Hvis antall variable blir så høyt i et parameterkort at man ikke får plass til dem på ett kort, så slutter man på Kol. 72 og fortsetter på et nytt kort.

12. 3 avslutningskort

i)	Kol.	1-4	Blank
	"	5-6	99
ii)	Kol.	1-2	/ ¥
iii)	Kol.	1-2	/ & (Rødt kort)

1. Eksempler på programmering av en del 1-1-tydige transformasjoner

Addisjon: $D(K) = D(S_1) + D(S_2)$

Subtraksjon: $D(K) = D(S_1) - D(S_2)$

Multiplikasjon: $D(K) = D(S_1) * D(S_2)$

Divisjon: $D(K) = D(S_1) / D(S_2)$

Eksponentiell: $D(K) = DEXP(D(S))$

Logaritmisk: $D(K) = DLOG(D(S))$ $D(S) > 0$ grunntall: e

Potensering: $D(K) = D(S) ** N$ der N er naturlig tall

Kvadratrot: $D(K) = DSQRT(D(S))$. $D(S) \geq 0$

(K kan her settes lik S, S_1 eller S_2)

2. Programmering av Dummy-gruppering

(Enveisgruppering)

1) Inndeling av kontinuerlig variabel i grupper

Eks.: Anta at inntekt skal inndeles i 5 grupper: 0-10, 10-50, 50-100, 100-1000, 1000 ———>, alt i f.eks. 1 000 kroner og at 1000 ———> er vår basisgruppe. Anta så at inntekt er originalvariabel nr. 5 (D(S)). Anta videre at vi hittil har definert 10 variable (originalvariable + transformerte variable).

```
   DO 2 i = 11,14
 2 D(i) = 0.0
   IF (D(5) - 10.0)  3,3,4
 3 D (11) = 1.0
   GO TO 14
 4 IF (D(5) - 50.0)  5,5,6
 5 D(12) = 1.0
   GO TO 14
 6 IF (D(5) - 100.0)  7,7,8
 7 D(13) = 1.0
   GO TO 14
 8 IF (D(5) - 1000.0) 9,9,14
 9 D(14) = 1.0
   GO TO 14
14 CONTINUE
```

2) Inndeling av diskret variabel i grupper

Eks.: Anta at vi tilstreber en næringsgruppering. La næringene være: Jordbruk, skogbruk, fiske, bergverk og industri, varehandel, bygg og anlegg, med de øvrige næringer som referansegruppe: La disse næringene f.eks. ha kodene: 01, 02, 04, 11 + 20 + 23 + 25 + 28 + 31, 60, 41, 51 + 67 + 69 + 70 + 73 + 79 + 81 + 89 + 99 henholdsvis. Vi ser at vi har en del aggregeringer av næringer her. Vi antar at den næringsvariable er D(1) og at vi hittil har definert 14 variable som ovenfor.

```
   DO 10 i = 15,20
10 D(i) = 0.0
   IF (D(1) - 01)    16,15,16
15 D(15) = 1.0
   GO TO 37
16 IF (D(1) - 02)    18,17,18
17 D(16) = 1.0
   GO TO 37
18 IF (D(1) - 04)    20,19,20
19 D(17) = 1.0
   GO TO 37
20 IF(D(1) - 11)   22,21,22
21 D(18) = 1.0
   GO TO 37

22 IF(D(1) - 20)   24,21,24
24 IF(D(1) - 23)   26,21,26
26 IF(D(1) - 25)   28,21,28
28 IF(D(1) - 28)   30,21,30
30 IF(D(1) - 31)   32,21,32
32 IF(D(1) - 60)   34,33,34
33 D(19) = 1.0
   GO TO 37
34 IF(D(1) - 41)   36,35,37
35 D(20) = 1.0
   GO TO 37
37 CONTINUE
```

3) Kommentarer til output

La regresjonsmodellen være $Y = X\beta + U$ eller om en vil:

$$Y_i = \beta_0 + \beta_1 X_n + \beta_2 X_{2i} + \ldots + \beta_s X_{si} + U_i \quad \text{for } i = 1, \ldots, n$$

der n er antall observasjoner.

1) Sum av kryssprodukt: $\sum_{i=1} D(J)_i \, D(K)_i$ der J og K er hvilke som helst tall fra 1 ..... M, der M er totalt antall definerte variable i subrutine DATA. Foran tallet står her angitt hvilke variable det tas kryssprodukt med hensyn på.

2) Sum-matrisen angitt rett og slett summen av hver variabel: $\sum_{i=1}^{n} D(J)_i$.

3) Gjennomsnittene: $\frac{1}{n} \sum_{i=1}^{n} D(J)_i$ for hver J.

4) Standardavvik: Kvadratroten av $\frac{1}{n-1} \sum_{i=1}^{n} (D(J)_i - D(\bar{J}))^2$ for hver J, altså empirisk standardavvik.

5) Sum av kryssprodukts-avvikene fra middeltallene: $\sum_{i=1}^{n} (D(J)_i - D(\bar{J}))\,(D(K)_i - D(\bar{K}))$. Foran disse tallene er også her angitt hvilke variable J og K det dreier seg om.

6) Korrelasjonsmatrisen består av ledd av typen r_{JK} der

$$r_{JK} = \frac{\sum_{i=1}^{n} (D(J)_i - D(\bar{J}))\,(D(K)_i - D(\bar{K}))}{\sqrt{\sum_{i=1}^{n} (D(J)_i - D(\bar{J}))^2 \sum_{i=1}^{n} (D(K)_i - D(\bar{K}))^2}}$$

altså empirisk korrelasjonskoeffisienter.

7) Den inverterte korrelasjonsmatrise.
Her skal bemerkes at det er ingen rutiner som sjekker nøyaktighetsgraden av inverteringen. Denne kan bli dårlig. Hvis en får mistanke om dette, bør en selv manuelt kontrollere noen tall i identiteten $R \cdot R^{-1} = I$.
Som vanlig står oppført (i - j)-te element foran tallet; i tillegg står det bak tallet hvilken eksponent det er opphøyet i. (10 som grunntall).

8) Korrelasjonskoeffisienten mellom den avhengige og hver av de uavhengige variable er åpenbart

$$R_J = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) (D(J)_i - D(\bar{J}))}{\sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sum_{i=1}^{n} (D(J)_i - D(\bar{J}))^2}}$$

der J nå står for en av de uavhengige variable i denne spesifikke regresjonen.

9) Regresjonskoeffisientene. Disse kan lett nedskrives på matriseform:

$\hat{\beta} = (X'X)^{-1} X'Y$, eller om en vil at $\hat{\beta}$ er løsningen av likningssystemet:

$(X'X)\ \hat{\beta} = X'Y$.

10) La $S^2 = \frac{1}{n-s-1} (Y - X\hat{\beta})' (Y - X\hat{\beta})$ eller om en vil $S^2 = \frac{1}{n-s-1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \beta_1 X_{1i} \ldots\ldots \hat{\beta}_s X_{si})^2$. Da er den empiriske varians på regresjonskoeffisienten $\hat{\beta}_i$ lik S^2 multiplisert med det j-te diagonalledd i matrisen $(X'X)^{-1}$, og det empiriske standardavvik $STD\hat{\beta}_j$ (for $\hat{\beta}$) er lik kvadratroten av dette uttrykk.

11) Angir t-verdien til testing av hypotesen H: $\beta_j = 0$ mot $\beta_j \neq 0$. Man forkaster altså hypotesen hvis

$$t = \frac{|\hat{\beta}_j|}{STD\hat{\beta}_j} > t_{1-\frac{\varepsilon}{2},\ n-s-1}$$

12) Konstantleddet er minste kvadraters estimat av β_o : $\hat{\beta}o = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \ldots\ldots - \hat{\beta}s\bar{X}o$.

13) Multippel korrelasjonskoeffisient angir følgende størrelse $R_{o\ .\ 1\ \ldots\ldots\ s}$ der

$$R^2_{o\ .\ 1\ \ldots\ldots\ S} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{is})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

eller om en vil:

$$R^2_{o\,.\,1\,.....\,s} = 1 - \frac{(Y - X\beta)'(Y - X\beta)}{(Y - \bar{Y})'(Y - \bar{Y})}$$

14) Standardavvik på estimatet betyr kvadratroten av et forventningsrett estimat på variansen på residualleddet, σ^2. Det er rett og slett lik S som er angitt i 10).

15) F-verdien angir forholdet mellom de to Mean Squares på listen, nemlig

$$F = \frac{\frac{1}{s}\left|\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{1i} ..... \hat{\beta}_s X_{si})^2\right|}{\frac{1}{n-s-1}\Sigma\,(Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{1i} ..... \hat{\beta}_s X_{si})^2}$$

eller om en vil på matriseformen:

$$F = \frac{1/s\,\left|(Y - \bar{Y})'(Y - \bar{Y}) - (Y - X\hat{\beta})'(Y - X\hat{\beta})\right|}{1/n-s-1\,(Y - X\hat{\beta})'(Y - X\hat{\beta})}$$

16) Restleddene kan beregnes for hver observasjon ved $\hat{U}_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{1i} ..... \hat{\beta}_s X_{si}$, for i = 1,....,n.

17) Durbin-Watson's test.

18) Kji-kvadrat-testing av normalitet i restleddene

Til denne test trenges ingen programmering. Likevel skal det anføres en del kommentarer av mer statistisk natur. Tall-linjen er delt inn i 7 grupper symmetrisk på hver side av 0: (0 - 0,5), (0,5 - 1,0),..... (2,5 - 3,0) (3,0 ——>) og symmetrisk. For å teste normalitet har en først standardisert restleddene

$$X_i = \frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}}{S\,\hat{\varepsilon}} \quad \text{der } S\,\hat{\varepsilon} = \frac{1}{n-1}\sum_{1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$$

Vår hypotese er da: H : X-ene er N(0,1). Testkriteriet er: Forkast når:

$$\chi^2 = \sum_{1}^{14}\frac{(h_i - n.p_i)^2}{npi} > \chi^2_{13}$$

der h_i er hyppigheten av X-ene i gruppe i og p_i er den teoretiske sannsynlighet for å falle i denne gruppe beregnet under H ved hjelp av tabellene over normalfordelingen. En kan kanskje lure på om det ikke finnes en annen gruppeinndeling som gir bedre styrke på testen. En måte å gå fram på er da først å forutsette at gruppen er slik at den teoretiske sannsynlighet for at observasjonen skal falle i en gruppe er lik for alle grupper. Deretter velges det antall grupper som maksimerer styrkefunksjonen. Dette er vist i M.G. Kendall and Stuart, Bind II, side 437 - 440. Det finnes også andre optimalitetskriterier, f.eks. at man velger den gruppeinndeling som minimaliserer innengruppen-variansene, etc.

4. Sluttkommentar

På grunn av tidsnød har en måttet stoppe den videre bearbeiding av regresjonsprogrammet. Av ting som i framtiden kan være av interesse å få gjort, er å få lagt inn en del rutiner til kontroll av beregningenes nøyaktighetsgrad. Dette gjelder f.eks. inverteringen av korrelasjonsmatrisen. Videre kunne det være interessant å få programmert et konfidensintervall for prediksjonsverdien av den endogene variable, som omtalt ovenfor; kanskje kunne en også klare å få det mer generelt til å omfatte konfidensintervall for et hvilket som helst vektorrom av lineærformer i en hvilken som helst variansanalysesituasjon. (I virkeligheten er det jo ingen formell forskjell mellom variansanalyse og regresjonsanalyse).

Histogram for restledd, hvis disse ønskes

Hvis en ønsker tegnet et histogram over restleddene ved regresjonsanalysen, kan man mellom de to siste kort i regresjonsprogrammet legge følgende kort-bunke:

```
// JOB HIST
// OPTION LINK
   INCLUDE DASCR
// EXEC FORTRAN
      SUBROUTINE BOOL (ICODE, NS, R, T)                          BOLL 001
      DIMENSION R(1)                                             BOOL 002
      RETURN                                                     BOOL 003
      END                                                        BOOL 004
/ *
// EXEC FORTRAN
      SUBROUTINE MATIN (ICODE, A, ISIZE,IROW,ICOL,IS,IER,NS)     MATIN069
      DIMENSION A(1)                                             MATIN070
      DIMENSION CARD (7)                                         MATIN071
    1 FORMAT (7F10.0)                                            MATIN072
    2 FORMAT (I6,2I4,I2)I3                                       MATIN073
      IDC = 7                                                    MATIN075
      IER = 0                                                    MATIN076
      READ (5,2)ICODE,IROW,ICOL,IS,NS                            MATIN077
      CALL LOC (IRON,ICOL,ICNT,IROW,ICOL,IS)                     MATIN078
      IF(ISIZE - ICNT) 6,7,7                                     MATIN079
    6 IER = 1                                                    MATIN080
    7 IF (ICNT) 38,38,8                                          MATIN081
    8 ICOLT = ICOL                                               MATIN082
      IROCR = 1                                                  MATIN083
   11 IRCDS = (ICOLT - 1) / IDC + 1                              MATIN087
      IF (IS - 1) 15,15,12                                       MATIN088
   12 IRCDS = 1                                                  MATIN089
   15 DO 31 K = 1, IRCDS                                         MATIN093
      READ (7,100) CARD(1)                                       MATIN094
  100 FORMAT (F20.10)                                            MATIN095
      IF (IER) 16,16,31                                          MATIN098
   16 L = 0                                                      MATIN099
      JS = (K - 1 ) * IDC + ICOL - ICOLT + 1                     MATIN103
      JE = JS + IDC - 1                                          MATIN104
      IF (IS - 1) 19,19,17                                       MATIN105
```

```
      17 JE = JS                                                    MATIN106
      19 DO 30 J = JS, JE                                           MATIN110
         IF (J - ICOL) 20,20,31                                     MATIN111
      20 CALL LOC (IROCR, J,IJ, IROW, ICOL, IS)                     MATIN112
         L = L + 1                                                  MATIN113
      30 A(IJ) = CARD(L)                                            MATIN114
      31 CONTINUE                                                   MATIN115
         IROCR = IROCR + 1                                          MATIN116
         IF (IROW - IROCR) 38,35,35                                 MATIN117
      35 IF(IS - 1) 37,36,36                                        MATIN118
      36 ICOLT = ICOLT - 1                                          MATIN119
      37 GO TO 11                                                   MATIN120
      38 READ (7,1) CARD(1)                                         MATIN121
         IF (CARD(1) - 9.E9) 39,40,39                               MATIN122
      39 IER = 2                                                    MATIN123
      40 RETURN                                                     MATIN124
         END                                                        MATIN125
   / *
   // EXEC LINKEDT
```

Parameterkort (se nedenfor)

Også dette program ligger på disk i maskinen, slik at en selv må fylle ut kun noen få kort, de såkalte parameterkort:

Først kommer ASSGN-kortene:

a) Data-input: logisk nr. 7 SYS 004
   Kontrollkort: 5 SYS 002
   OUtput: 6 SYS 003
   Arbeidstap: 13 SYS 001

slik at ASSGN-kortene da blir:

```
// ASSGN SYS 002, X'014'
// ASSGN SYS 003, X'00E'
// ASSGN SYS 004, X'184'
// ASSGN SYS 004, X'183'
```

Deretter følger et EXEC-kort:

```
// EXEC
```

b) Parameterkort (fylles ut selv)

i) Problemkort

Kol.	1 - 2	Blanke
"	3 - 6	Firesifret problemnummer (etter ønske)
"	7 - 10	Antall observasjoner
"	11 - 14	Antall variable (i vårt tilfelle: 01)
"	15 - 16	Matrisens form (her 00)
"	17 - 19	Antall seleksjoner (her 001 vanligvis)

Eventuelt: Hvis data finnes på kort, så kommer disse etter problemkortet og etter disse kommer så et slutt-data-kort hvor man i kol. 1 setter et 9-tall som betyr at vi ikke har flere data-kort (dette trengs altså ikke hos oss).

ii) Betingelseskort

Kol. 1 - 2. Antall betingelser (hos oss 00)

" 3 - 4. Nr. på den variable som skal plottes (01)

(Her kommer så inntil 3 kort som angir betingelser hvis antall betingelser i kol. 1 - 2 ε (0,21). Disse trengs altså ikke i vårt tilfelle).

iii) Grensekort

Kol. 1 - 10. Nedre grense (helt tall f.eks. -4).

" 11 - 20. Antall intervall (max. 20). Deles inn i like lange intervall).

" 21 - 30. Øvre grense (helt tall - f.eks. 4).

Det skal jer bemerkes at hvis en i stedet er interessert i å la nedre grense være lik variabelens minimumsverdi og øvre grense lik variabelens maksimumsverdi, så setter en i grensekortet nedre grense = øvre grense, begge f.eks. lik 00. Da gjør maskinen resten.

Etter siste problem må det ligge et blankt kort.

Her er gjengitt beskrivelse for bruk spesielt på restleddene i regresjonsanalysen. Dette programmet er imidlertid mye mer generelt; det gir anledning til, ved noen få endringer, å "tegne" simultane fordelinger under opptil 21 betingelser, f.eks. kan det brukes til å lage flerdimensjonale tabeller. For slike mer generelle anvendelser av programmet henvises til Systemkontoret.