Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

IO 74/37 19. september 1974

STATISTISK SENTRALBYRÅS REGRESJONSPROGRAM

45 variable, dobbel presisjon. Maskin: H6060

(Revisjon av notat IO 68/1)

av

Grete Dahl og Kjetil Sørlie

INNHOLD

1. GENERELT OM PROGRAMMET. ANVENDELSE OG OUTPUT

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.

Denne dokumentasjon av programmet er utarbeidet for kjøring på Honeywell Bull. (For kjøring på IBM-maskin, se arbeidsnotat IO 68/1.)

b. Programmet estimerer koeffisientene i en lineær regresjonslikning av formen:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots \beta_s X_{si} + U_i$$

for i=1,2 ... n, n er antall observasjonssett.

Y_i = observasjon nr. i av avhengig variabel
$X_{1i}, X_{2i} \ldots X_{si}$ = observasjonssett nr. i av de uavhengige variable
$\beta_0, \beta_1 \ldots \beta_s$ = konstanter
U_i = restledd

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, og en eller flere uavhengige variable kan utelates. Utskillingen av den avhengige og de (den) uavhengige variable som en ønsker å nytte i regresjonen, må spesifiseres i parameterkortene (regresjonskortene[1]).

d. Programmet kan gi følgende output:

1. Sum av kryssprodukt (Hvis ønsket på problemkortet[1], kol. 21)
2. Sum-matrise (" " " " , kol. 20)
3. Gjennomsnittene (" " " " , kol. 16)
4. Standard avvik (" " " " , kol. 17)
5. Sum av kryssprodukt fra middeltallene (" " " " , kol. 18)
6. Korrelasjonsmatrisen(" " " " , kol. 19)
7. Invertert korrelasjonsmatrise (" " " " , kol. 22)

---

1) En nærmere omtale av problem- og regresjonskortene er gitt i avsnitt 3.

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

} (Dersom ikke 00 i kol. 14-15 på problemkortet[1])

16. Restledd (Hvis ønsket på regresjonskortet[1], kol. 1-2)
    Durbin-Watson's observator (" " " " , kol. 1-2)
    X^2-test på normalitet (XJIKVA) (" " " " , kol. 1-2)

I tillegg til dette er det mulig å få tegnet histogram over restleddene, hvis disse ønskes i regresjonen, se vedlegg B.

e. Begrensninger

Maksimalt antall variable er 45, (dvs. sum av originale og/eller transformerte variable[2] i subrutine DATA må ikke overstige 45). Maksimalt antall observasjonssett er 99999. Minimalt antall observasjonssett er 2 større enn antall variable som inngår i regresjonslikningen.

2. PROGRAMMETS RUTINER. RETTLEDNING FOR PROGRAMMERING AV SUBRUTINE DATA

a. Programmets rutiner

Regresjonsprogrammet består av en hovedrutine ved navn REGRE, en spesiell subrutine DATA, samt fem subrutiner, CORRE, ORDRER, MINV, MULTR, GRUPPE, som alle nå ligger i subrutine biblioteket, 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. Det er videre mulig å kalle inn fra programbiblioteket et program som tegner histogram over restleddene, hvis restleddene ønskes i regresjonen. En del nye kort i tillegg til de som er omtalt under avsnitt 5, må da fylles ut (se vedlegg B).

---

1) Se note 1 side 2. 2) Om originale og transformerte variable, se avsnitt 2.

b. Rettledning for programmering av subrutine DATA

Subrutine DATA inneholder 6 faste kort, se avsnitt 5, pkt. 9a-f. (Kortet som er spesifisert under pkt. 9d, er bare nødvendig hvis restleddsbehandling.)

Foruten disse faste kort vil subrutine DATA bestå av ett FORMAT-kort, ett READ-kort og eventuelle kort for transformasjoner av de variable. Innholdet av disse kort varierer med det oppdrag en utfører, og er bl.a. bestemt av filebeskrivelsen, logisk enhetsnummer for INPUT-DATA (=5), og hvilke transformasjoner som ønskes foretatt.

Subrutine DATA skrives i FORTRAN.

Formatet av input skal spesifiseres i FORMAT-kortet. Utgangspunktet for koding (punching) av dette kortet er filebeskrivelsen. Dette gjelder uansett om inputdata er på plate, bånd eller hullkort. Filebeskrivelsen gir en oversikt over hvor på båndet eller hullkortene hver opplysning (variabel) 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 fra felt til felt. I filebeskrivelsen er feltene og posisjonene nummererte. Posisjonene er nummerert fortløpende, f.eks. 1 ... 8 (for felt nr. 1), 9 ... 12 (for felt nr. 2), osv. ... 250 ... 255 (for felt nr. 30, som tenkes å være siste felt på denne recorden).

I en regresjonsanalyse er en ofte interessert i kvantitative variable som f.eks. sysselsetting, investering, bruttoproduksjon, realkapital osv. Disse variable finnes på båndet (kortene) og opptar hver et bestemt felt. I noen tilfelle kan en også være interessert i å nytte kvalitative variable som f.eks. næring, kommune, geografiske kriterier osv. i regresjonen. Behandlingen av slike kvalitative variable er helt analog. (En må her skaffe seg den kodeliste som er nyttet ved grupperingen av de kvalitative variable.)

De variable (input-data) som tas ut av filebeskrivelsen, nummereres D(1), D(2) ... D(N), der N er antall variable. Disse kalles originale variable, og danner utgangspunkt for koding (punching) av FORMAT-kortet. FORMAT-kortet angir hvilke variable, og feltlengden (antall posisjoner på båndet/kortene) for hver enkelt variabel, som tas ut av filebeskrivelsen.

Eksempel på utfylling av FORMAT-kortet:

Vi har en filebeksrivelse der hver record består av 255 posisjoner. I disse posisjoner har en opplysninger fra i alt 30 forskjellige variable, f.eks. sysselsetting, næring, bruttoproduksjon, formue, realkapital,

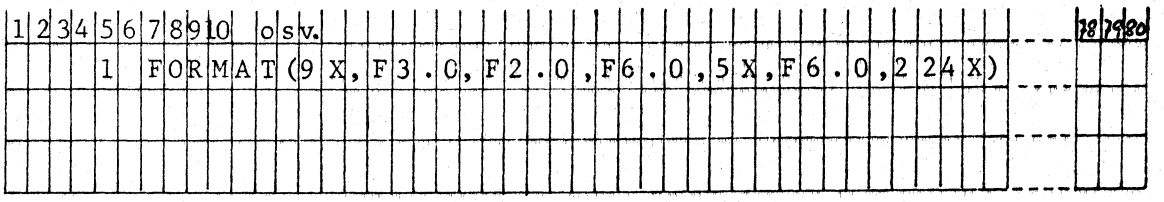
kommune osv. Opplysninger fra hver enkelt variabel finnes i bestemte posisjoner og hver variabel er tildelt ett feltnummer, f.eks.

"sysselsetting"	- felt	4, posisjon	10-12
"næring"	- "	5, "	13-14
"bruttoproduksjon"	- "	6, "	15-20
"formue"	- "	7, "	21-25
"realkapital"	- "	8, "	26-31
"kommune"	- "	9, "	32-35

osv.

Vi ønsker å ta ut av filebeskrivelsen variablene "sysselsetting" (D(1)), "næring" (D(2)), "bruttoproduksjon" (D(3)) og "realkapital" (D(4)). FORMAT-kortet vil da se slik ut:

Kolonnenummer på hullkortet:



9X betyr at de 9 første posisjoner på båndet (kortene) ikke leses, så leses 3 posisjoner for "sysselsetting" (F3.0), 2 posisjoner for "næring" (F2.0) osv. Dersom FORMAT-statementet ikke får plass i ett hullkort, fortsettes kodingen (punchingen) fra og med posisjon 7 i et nytt kort. Tallet 2 må da påføres i kolonne 6 i det nye kortet.

READ-kortet sørger for at data blir "lest inn" i maskinen. I READ-kortet kodes (punches) det logiske enhetsnummer for INPUT-DATA og nummeret på FORMAT-statementet, (kolonne 5 i FORMAT-kortet).

Eksempel på utfylling av READ-kort tilhørende FORMAT-kortet i eksemplet ovenfor: (ved ikke-transformasjon)

Fra og med posisjon 7 i hullkortet:

READ (5, 1) (D(J),J = 1,4)

I READ-kortet angir (D(J),J = 1,4) at det er 4 variable som blir "lest inn".

Dersom en i regresjonen ønsker å operere med de originale variable, er nå subrutine DATA ferdig programmert, og består av FORMAT-kort, READ-kort og de kort som er faste for subrutinen.

I en del tilfelle vil det imidlertid være av interesse å transformere en eller flere av de originale variable. Slike transformasjoner kan foretas i subrutine DATA, og en del nye kort for programmering av transformasjonene må fylles ut.

Ved transformasjoner vil READ-kortet se slik ut:

```
READ (5,1) (S(J),J = 1,N)
```

N = antall originale variable (innleste variable). Formatkortet punches på samme måte som beskrevet ovenfor. Under transformasjonen plasseres variablene i D, (se eksemplene nedenfor).

Eksempler på transformasjoner:

Programmering av en del 1-1-tydige transformasjoner

Addisjon:	$D(K) = S(J_1) + S(J_2)$
Subtraksjon:	$D(K) = S(J_1) - S(J_2)$
Multiplikasjon:	$D(K) = S(J_1) * S(J_2)$
Divisjon:	$D(K) = S(J_1) / S(J_2)$
Eksponentiell:	D(K) = DEXP (S(J))
Logaritmisk:	D(K) = DLOG (S(J)),S(J) > 0 grunntall:
Potensering:	D(K) = S(J) ** N,der N er naturlig tall
Kvadratrot:	D(K) = DSQRT(S(J)), $S(J) \geq 0$

Inndeling av en kontinuerlig variabel i grupper karakterisert ved binærvariable (dummy-variable)

Anta at f.eks. inntekt skal inndeles i 5 grupper: 0-10, 10-50, 50-100, 100-1000, 1000 →, alt i 1000 kroner og at 1000 → er referansegruppen. En ønsker å innføre binærvariable svarende til denne gruppering. Anta så at inntekt er originalvariabel nr. 5 (S(J)). Anta videre at det hittil er plassert 10 variable (originalvariable + transformerte variable) i subrutine DATA.

Kolonnenummer på hullkortet:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	osv.
							I	F	(	S	(	5	)	-	1	0	.	0	)	3	,	3	,	4					
					3		D	(	1	1	)	=	1	.	0														
							G	0		T	0		1	4															
					4		I	F	(	S	(	5	)	-	5	0	.	0	)	5	,	5	,	6					
					5		D	(	1	2	)	=	1	.	0														
							G	0		T	0		1	4															
					6		I	F	(	S	(	5	)	-	1	0	0	.	0	)	7	,	7	,	8				
					7		D	(	1	3	)	=	1	.	0														
							G	0		T	0		1	4															
					8		I	F	(	S	(	5	)	-	1	0	0	0	.	0	)	9	,	9	,	1	4			
					9		D	(	1	4	)	=	1	.	0														
							G	0		T	0		1	4															
				1	4		C	O	N	T	I	N	U	E															

Inndeling av en kvalitativ variabel i grupper karakterisert ved binærvariable (dummy-variable)

Anta f.eks. at en vil lage en næringsgruppering på et høyere aggregeringsnivå enn den som er nyttet i input-data. I input-data har næringene kodene 01, 02, 04, 11, 20, 23, 25, 28, 31, 60, 41, 51, 67, 69, 70, 73, 79, 81, 89 og 99.

Ved transformasjonen vil en slå sammen næringene i hovedgrupper på følgende måte:

Næringsgruppe 1 - kode 01
" 2 - " 02
" 3 - " 04
" 4 - kodene 11 + 20 + 23 + 25 + 28 + 31
" 5 - kode 60
" 6 - " 41
" 7 - kodene 51 + 67 + 69 + 70 + 73 + 79 + 81 + 89 + 99

Næringsgruppe 7 er referansegruppen. En ønsker å innføre binærvariable svarende til denne gruppering. En antar at den næringsvariable er S(1) og at det hittil er plassert 14 variable i subrutine DATA.

Kolonnenummer på hullkortet:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	osv.
						I	F	(	S	(	1	)	-	0	1	)	1	6	,	1	5	,	1	6			
			1	5		D	(	1	5	)	=	1	.	0													
						G	0		T	0		3	7														
			1	6		I	F	(	S	(	1	)	-	0	2	)	1	8	,	1	7	,	1	8			
			1	7		D	(	1	6	)	=	1	.	0													
						G	0		T	0		3	7														
			1	8		I	F	(	S	(	1	)	-	0	4	)	2	0	,	1	9	,	2	0			
			1	9		D	(	1	7	)	=	1	.	0													
						G	0		T	0		3	7														
			2	0		I	F	(	S	(	1	)	-	1	1	)	2	2	,	2	1	,	2	2			
			2	1		D	(	1	8	)	=	1	.	0													
						G	0		T	0		3	7														
			2	2		I	F	(	S	(	1	)	-	2	0	)	2	4	,	2	1	,	2	4			
			2	4		I	F	(	S	(	1	)	-	2	3	)	2	6	,	2	1	,	2	6			
			2	6		I	F	(	S	(	1	)	-	2	5	)	2	8	,	2	1	,	2	8			
			2	8		I	F	(	S	(	1	)	-	2	8	)	3	0	,	2	1	,	3	0			
			3	0		I	F	(	S	(	1	)	-	3	1	)	3	2	,	2	1	,	3	2			
			3	2		I	F	(	S	(	1	)	-	6	0	)	3	4	,	3	3	,	3	4			
			3	3		D	(	1	9	)	=	1	.	0													
						G	0		T	0		3	7														
			3	4		I	F	(	S	(	1	)	-	4	1	)	3	7	,	3	5	,	3	7			
			3	5		D	(	2	0	)	=	1	.	0													
						G	0		T	0		3	7														
			3	7		C	O	N	T	I	N	U	E														

3. INPUT. PROBLEMKORT OG REGRESJONSKORT

Input i programmet er parameterkortene (pkt. a), og datakort/ magnetbånd (pkt. b). I avsnitt 5, nedenfor, er hele kortoppsettet for regresjonsprogrammet spesifisert. I dette avsnitt blir det gitt en nærmere orientering om utfyllingen av problem- og regresjonskortene (parameterkortene).

a. Parameterkort (gule kort)

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 ≠ 99. For bruk av nr. 00 se *)
Kol. 7-11	Antall observasjonssett
Kol. 12-13	Antall variable plassert i subrutine DATA
Kol. 14-15	Antall regresjonskort. (Dersom en setter 0 her vil ikke programmet foreta noen regresjon, men en kan få den output som en kan velge fra kol. 16-22 i problemkortet. I dette tilfellet kan en bare ha et sampel i hver kjøring.)
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) Regresjonskort

På regresjonskortet skal angis en avhengig variabel samt alle uavhengige variable i en likning. Enhver variabel kan velges som den avhengige. Ved å lage flere regresjonskort for hvert problemkort kan brukeren ved samme kjøring estimere koeffisientene i flere regresjonslikninger. Hvis en ønsker histogram over restleddene må det imidlertid kun brukes én regresjonslikning for hver kjøring.

---

* 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).

Kortet utfylles på følgende måte:

Kol. 1- 2	01 Hvis restledd ønskes som output. 00 " " ikke ønskes som output. Når en setter 00, blir ikke rutinen som beregner restledd påkallet, og enhver form for restleddbehandling er umulig.
Kol. 3- 4	Nummer på den avhengige variable[1)]
Kol. 5- 6	Antall uavhengige variable
Kol. 7- 8	Nummer på første uavhengige variable[1)]
Kol. 9-10	Nummer på annen uavhengige variable[1)]
	osv. OBS! Dersom man har mer enn 33 uavhengige variable, så setter man variabel 34 og oppover i et nytt kort fra kol. 1.

Eksempler på utfylling av problem- og regresjonskort

Anta at input-data finnes på kort og at antall observasjonssett er 30. Antall variable er 6. En ønsker å estimere koeffisientene i en likning av formen

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

og en annen likning

(2) $V(6) = A \cdot V(2) + B \cdot V(3) + C \cdot V(5)$,

hvor V(i) betegner variabel nr. i. Restleddene ønskes som output.

Utfylling av problemkortet:

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

Utfylling av regresjonskortet (likning 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

Utfylling av regresjonskortet (likning 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

1) Variablene gis de numre de har fått ved innlesningen eller eventuelt ved transformasjonen i subrutine DATA.

b. Datakort (hvis data er på kort)
   Magnetbånd (hvis data er på tape)

4. LOGISKE ENHETSNUMRE

a. Logisk enhetsnr. 1 brukes for parameterkort
b. Logisk enhetsnr. 8 brukes for output
c. Logisk enhetsnr. 13 brukes for MELLOMLAGER
d. Logisk enhetsnr. 5 brukes for INPUT-DATA
e. Logisk enhetsnr. 7 brukes for RESTLEDD
f. Logisk enhetsnr. 6 brukes for restledd til HISTOGRAM-programmet.

5. KONTROLLKORT OG KORTOPPSETTET ELLERS FOR REGRESJONSPROGRAMMET

Alle kontrollkort (blå kort) har samme faste format, nemlig

Kol. 1	Å
Kol. 2- 7	blank
Kol. 8-15	tekstfelt A
Kol. 16-80	" B

Tekstfeltene blir heretter bare omtalt som A og B. (Ubenyttede kolonner i A og B blir stående blanke.)

Alle kort - kontrollkort, parameterkort, eventuelle datakort og kortene for subrutinen DATA - skal ligge i den rekkefølgen de her blir omtalt.

1. Kontrollkort for jobidentifikasjon

A IDENT
B SSBnnnnXXhhp,SB41kkNNN5500290
ikke-faste parametre her er:
nnnn = statistikknr.
hh = nøkkel for statistikknr. - fås på Driftskontoret
p = prioritet, gyldige koder A, B, C, D
kk = kontornr.
nnn = brukers initialer

2. Kontrollkort som gir tillatelse til bruk av Byråets programbibliotek

A USERID

B SSBÅXXXXXXXX

XXXXXXXX er en kode som fås opplyst på Systemkontoret

3. Kontrollkort for limit på job

A LIMITS

B tt,32K,,LLK

tt = tidskode:

05 = 3 min.
10 = 6 min.
15 = 9 min.
osv.

LL= antall linjer (1K = 1024 linjer)

Hvis hele jobben tar mindre enn 3 minutter og antall linjer som skrives antas å være under 10240 (10K), kan dette kortet utelates.

4. Kontrollkort for loading av programmet

A OPTION

B FORTRAN

5. Kontrollkort for kompilering av programmet

A FORTY

B blank

Hvis programlistingen ønskes utelatt settes NLSTIN i B.

6. Kontrollkort for kalling på programmet fra programbibliotek

A PRMFL

B S*,R/W,S,SSB/SOURSE/S0001/S5500290

7. Kontrollkort for kompilering av subrutine DATA

A FORTY

B blank

8. Kontrollkort for oversettelse fra IBM-koder til HB-koder

A INCODE

B IBMEL

9. Kortene for subrutine DATA

Faste kort til denne er:

a) SUBROUTINE DATA (M,D,IPR)

b) DIMENSION D(1) (hvis ikke transformasjon)
DIMENSION S(8) (hvis transformasjon fra f.eks. 8 innleste variable)

c) DOUBLE PRECISION D

Så følger READ-kort, FORMAT-kort og kort for eventuelle transformasjoner av data. Til slutt kommer kortene

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

d) er bare nødvendig hvis restleddbehandling (se avsnitt 1.d, pkt. 16).

e) RETURN

f) END

10. Kontrollkort for tilgang til subrutinebibliotek

A LIBRARY

B A4

11. Kontrollkort for eksekvering av programmet

A EXECUTE

B blank

12. Kontrollkort for kalling av subrutiner i biblioteket SUBLIB

A PRMFL

B A4,R,R,SSB/SUBLIB

13. Kontrollkort for limit på eksekvering

A LIMITS

B tt,20K,,LLK

20K står for programstørrelsen

tt og LL har samme betydning som i pkt. 3, men med den forskjell at det her gjelder eksekveringsaktiviteten istedetfor hele jobben

14. Kontrollkort for allokering av mellomlager (disk)

A FILE

B 13,X1R,10L

15. Kontrollkort for montering av tape

A TAPE

B 05,X2D,,nnnnn,,,3

nnnnn = båndnummer (volumnr.)

Dette kortet skal bare være med hvis data er på tape. Hvis data er på kort, skal det erstattes med kortene i pkt. 16

16. Kontrollkort for datakortene

A DATA

B 05

Så følger et INCODE-kort helt maken til kortet i pkt. 8. Deretter kommer datakortene (se 3.b)

17. Kontrollkort for parameterkortene

A DATA

B 01

Så følger igjen et INCODE-kort som i pkt. 8

Heretter kommer parameterkortene:

a) et sannsynlighetskort

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

b) problemkort og regresjonskort som beskrevet i 3.a).

c) et sluttkort:

Kol.	1- 4	blank
"	5- 6	99
"	7-80	blank

18. Kontrollkort for outputliste

A SYSOUT

B 08

19. Kontrollkort nødvendig hvis restleddbehandling (se avsn. 1.d, pkt. 16)

a) Her kreves allokering av nok en outputliste, vi trenger kontrollkortet beskrevet ved

   A SYSOUT

   B 07

b) Det må allokeres en file til bruk for HISTOGRAM-programmet, dette gjøres med følgende kontrollkort

   A FILE

   B 06,H1S,10L

20. Kontrollkort som markerer slutt på job

A ENDJOB

B blank

Se at det er overensstemmelse mellom de logiske enhetsnumre i avsnitt 4 og de samme numre (filekodene) i kontrollkortene TAPE, FILE, DATA og SYSOUT.

6. 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.

Vedlegg A

KOMMENTARER TIL OUTPUT

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

$Y_i = \beta_0 + \beta_1 X_n + \beta_2 X_{2i} + \ldots + \beta_s X_{si} + U_i$ for i=1, ..., 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 $S^2 = \frac{1}{n-s-1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{1i} \ldots\ldots - \hat{\beta}_s X_{si})^2$. 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 : $\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \ldots\ldots - \hat{\beta}_s \bar{X}_s$

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

$$R^2_{o\,.\,1\,\ldots\ldots\,s} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \sum_{j=1}^{s} \hat{\beta}_j X_{ij})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

eller om en vil:

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

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

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

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

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 $\hat{U}_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 X_{1i} \ldots\ldots \hat{\beta}_s X_{si}$, for i=1, ..., n.

17) Durbin-Watson's test.

18) Kji-kvadrat-testing av normalitet i restleddene
Til denne test trengs 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 $\rightarrow$) og symmetrisk. For å teste normalitet har en først standardisert restleddene

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

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 sammsynlighet 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 optimalitets-kriterier, f.eks. at man velger den gruppeinndeling som minimaliserer innen-gruppen-variansene, etc.

HISTOGRAM FOR RESTLEDD, HVIS DISSE ØNSKES

For å få printet ut et histogram for restleddene må antall observasjonssett (records) være ≤ 1 800.

Kortene må ligge umiddelbart foran Å ENDJOB (pkt. 20, avsnitt 5) og i den rekkefølgen de her blir beskrevet.

1. Kontrollkort, kortoppsettet

Å	IDENT	SSB .., osv. (som pkt. 1, avsnitt 5)
Å	OPTION	FORTRAN
Å	FORTY	NLSTIN
Å	PRMFL	S*,R/W,S,SSB/SOURSE/S0001/S5540320
Å	LIBRARY	A4
Å	EXECUTE	
Å	PRMFL	A4,R,R,SSB/SUBLIB
Å	LIMITS	,19K
Å	FILE	07,H1R,10L
Å	FILE	13,X2R
Å	DATA	05

3 parameterkort (se pkt. 2)

Til slutt kommer så Å ENDJOB-kortet

2. Parameterkort

Det skal være 3 parameterkort:

1. kort (problemkort)

Kol.	1- 2	blanke
"	3- 6	firesifret problemnr. (etter ønske)
"	7-10	antall observasjoner (max 1 800)
"	11-19	konstant: [000100001]
"	20-80	blanke

2. kort (betingelseskort)

Kol.	1- 4	konstant: [0001]
"	5-80	blanke

3. kort (grensekort)

Kol.	1-10	nedre grense (helt tall, f.eks. - 4)
"	11-18	blanke
"	19-20	antall intervall (max 20, min 3)
"	21-30	øvre grense (helt tall, f.eks. 4)
"	31-80	blanke

Histogramprogrammet er egentlig mer generelt enn nødvendig for vår anvendelse, derfor opereres det med konstante parametre.

Hvis en ønsker at nedre grense for histogrammet skal være variabelens minimumsverdi og øvre grense variabelens maksimumsverdi, settes kol. 1-10 og kol. 21-30 i grensekortet blanke. Det histogram en får tegnet, vil generelt bestå av to intervaller færre enn angitt i grensekortet (altså max 18) mellom nedre og øvre grense. Disse intervallene er alle like lange. I de to ytterintervallene registreres observasjoner < nedre grense og > øvre grense. (Disse to intervallene blir naturligvis tomme hvis nedre grense = minimumsverdi og øvre grense = maximumsverdi.)