REGNECENTRALEN
Uofficielt program
Frank Nymand nr. 3
2o. august 1962

Beskrivelse af program for

BEREGNING OG SAMMENFATNING AF LINEÆRE REGRESSIONSLINIER

1. Programmets funktion

Programmet arbejder på et datamateriale, der kan inddeles i q grupper via et gruppenummer (i = 1,2,....q), som hvert datasæt indeholder. Derudover består et datasæt af 1 uafhængig variabelværdi (x) samt m stokastiske variabelværdier, (y_j, hvor j = 1,2, ...,m).

For hver af de m stokastiske variable, uafhængigt af de øvrige, beregner programmet efter sædvanlige metoder regressionskoefficienterne i den lineære regression af y_j på x inden for hver af de q grupper. Derefter sammenfattes regressionskoefficienterne i de q grupper, under forskellige hypoteser om datamaterialets struktur, ligesom der beregnes teststørrelser svarende til visse hypoteser. Endelig beregnes også skøn over middelværdier og spredninger samt andre parameterskøn og teststørrelser til vurdering af regressionskoefficienterne.

2. Beregningsgrundlag

Forudsætninger og metode for programmets beregninger er nogenlunde som beskrevet af A. Hald: Statistical Theory with Engineering Applications, chap. 18.9.

Inden for den enkelte inddelingsgruppe er som middeltal og spredning på x og y angivet de sædvanlige centrale skøn ($\bar{x}$., $\bar{y}$., s(x), s(y) og s(y.)).

Hældningerne beregnes som $b = SAP/SAK_x$, og residualspredningerne fås af

$$s(y|x)^2 = SAKy|x/(n-2) = (SAK_y - SAP^2/SAK_x)/(n-2)$$

medens variansen på b beregnes som $s(b)^2 = s(y|x)^2/SAK_x$.

Herefter beskrives beregningen af de sammenfattede størrelser. Hvor der i det følgende er anvendt summationstegn uden indeks på de variable, betyder det blot, at der er summeret over alle q grupper, (i = 1,2,3,....q).

Ud fra den forudsætning, at alle q regressionslinier er parallelle, beregnes den fælles hældning ($b_.$), spredningen på denne ($s(b_.)$) samt den tilsvarende residualspredning ($s_.$).

$$b_. = \sum SAP / \sum SAK_x$$

$$s_.^2 = (SAK1 + SAK2)/(\sum n_i - q - 1),$$

og $$s(b_.)^2 = s_.^2 / \sum SAK_x,$$

hvor $SAK1 = \sum SAKy|x$

og $SAK2 = \sum (b_i - b.)^2 \, SAKx$

Her betegner n_i antallet af observationer i den enkelte inddelingsgruppe.

SAKxT, SAKyT og SAPT er de respektive kvadrat- og produktafvigelsessummer beregnet på hele materialet under ét. Endelig betegnes med Sx_i og Sy_i summerne i gruppe i af henholdsvis den uafhængige og den stokastiske variabel.

Hypotese 1, at $b. = 0$, testes ved et t-test, nemlig

$$t_2 = b./s(b.), \text{ hvor } f_2 = \sum n - q - 1.$$

Hypotese 2, at de q regressionslinier er parallelle, testes ved hjælp af

$$v_1^2 = \left[SAK2/(q-1)\right] \Big/ \left[SAK1/(\sum (n-2))\right]$$

med $f_T = q-1$ og $f_N = \sum (n-2)$

Der indføres nu en parameter, c, hvor $c = 0$, når gruppemiddeltallene på x er identiske; i modsat fald sættes $c = 1$. Derpå beregnes nogle hjælpestørrelser, $\hat{b}$, SAK3 og s.

$$\hat{b} = (SAPT - \sum SAP) \,/\, (SAKxT - \sum SAKx), \text{ hvis } c = 1,$$

$$\hat{b} = 0, \text{ hvis } c = 0,$$

$$SAK3 = SAKyT - \sum SAKy - \hat{b}\,(SAPT - \sum SAPxy),$$

og $$s^2 = (SAK1 + SAK2 + SAK3) \,/\, (\sum n - 2 - c).$$

Hypotese 3, linearitet af regressionsliniernes tyngdepunkter, kan under forudsætning af hypotese 2 og $c = 1$ testes med følgende

$$v_2^2 = SAK3/s.^2(q-1-c), \text{ idet } f_T = q-1-c \text{ og } f_N = \sum n-q-1$$

For $c = 0$ giver v_2^2 et test for hypotese 4.

Hypotese 4 udsiger, at alle regressionslinierne er identiske. Hvis hypotese 2 og 3 ikke er forkastede, og hvis c=1, kan hypotese 4 testes med t_1, hvor

$$t_1^2 = (\hat{b} - b.)^2 (SAKxT - \sum SAKx)(\sum SAKx) \,/ s^2(SAKxT),$$

og hvor $f_1 = \sum n - 3$

Hypotese 5, som er arbejdshypotese for de 4 øvrige hypoteser, udsiger, at gruppe-residualspredningerne er identiske. Hypotesen testes ved hjælp af G. Rasch's X^2-test, som er beskrevet i F. Abildgaard Jørgensen: Kompendium i statistik for pædagoger og psykologer, 1957.

Når Li betegner den naturlige logaritme til residualvariansen i gruppe i, og $f_i = n_i-2$, fås følgende teststørrelse:

$$X^2 = \frac{1}{2}\left[\sum f_i L_i^2 - (\sum f_i L_i)^2 / \sum f_i\right],$$

med $f = q - 1$

Det må bemærkes, at dette test forudsætter $f_i \geq$ lo, og at residualvarianserne ikke er nul.

Til slut beregnes q konstanter (a_i), som f.eks. kan bruges til en eventuel korrigering af de afhængige variable.

$$a_i = (Sy_i - b.Sx_i)/n_i - \sum (Sy_i - b.Sx_i)/ \sum n_i$$

3. Programmets struktur

Programmet er kodet til elektronregnemaskinen DASK efter NL4-konventioner. Sekvenslageret og indlæseprogrammet i NL1 benyttes.

Følgende sekvenser findes i programmet:

1) HS - en hovedsekvens, som administrerer de øvrige sekvenser ved hjælp af følgende

2) Tromleadministration - der sørger for at bringe ferritlagerafsnit fra og til tromlen,

3) Akkumulatorsekvens - som under datainlæsningen beregner summer, kvadratsummer og produktsummer,

4) Storingssekvens - som omregner mellemresultater fra faste til flydende tal,

5) Regressionssekvens - som udfører de egentlige beregninger, og samtidig administrerer følgende

6) Udskriftssekvens - der udskriver beregningsresultaterne,

7) Den lille flydende - en hjælpesekvens, som administrerer flydende regning og funktionsberegning vedhjælp af sekvenslageret.

4. Lagerdisponering

Programmet disponerer over tromlen og ferritlageret på følgende måde:

Betegnelse	Ferritlagerceller	Tromlekanaler
HS	316 - 4o7	
Tromleadministration	4o8 - 456	444
Akkumulatorsekvens	416 - 471	446
Storingssekvens	0 - 75	43o - 432
Regressionssekvens	6oo - 1o37	414 - 428
Udskriftssekvens	47o - 593	4o6 - 4o8
Den lille flydende	64 - 123	412
SKL-arbejdslager	0 - 63	41o
Arbejdslager 1	472 - 471 + 6q (m+1)	
- 2		1oo - 98 + 6q
- 3	124 - 315	
Til retablering af dataadministration		434 - 442

NB! For q = 12 og m = 2o benyttes arbejdsceller helt op til hac 1984.

Under indlæsning af styringsinformation benyttes HS samt af NL1 halvcellerne: 1792 - 2o47.
Under dataindlæsningen benyttes HS, Akkumulatorsekvensen samt Arbejdslager 1. Derefter gemmes hac 0 - 315 på tromlen, og under resten af gennemløbet benyttes alle de resterende lagerafsnit, nævnt i ovenstående skema. Til slut retableres afsnittet: hac 0 - 315.

5. Styringsinformation

Programmet skal tilføres styringsinformation i form af følgende 6 parametre perforeret på en 5-kanalstrimmel i nævnte rækkefølge. Styringsstrimlen indlæses ved hjælp af indlæseprogrammet i NL1.

```
CSF     S = sagsnummer
dA00    d = indhopsadressen i dataadministrationen
aA00    a = begyndelsesadresse for placering af datasættet
CpF     p = kommaplacering for de variable
mA00    m = antal afhængige variable
qA00    q = antal inddelingsgrupper
 E
```

De 6 parametre placeres i hac 399 - 4o4 i indlæsningsrækkefølge, og de kan udelades "bagfra", hvis programmet i forvejen er tilført information om de parametre, man vil udelade.

Programmets kapacitet fremgår af følgende begrænsninger:

$S \leq 9999$

$1 \leq m \leq 31$

$1 \leq q \leq 51$

og $q\,(m+1) \leq 252$

6. Datamaterialet

Dette består af et antal datasæt, hvert indeholdende m+2 tal, nemlig først et gruppenummer (i = 1,2,3,...., eller q) og dernæst m+1 variabelværdier (1 x-værdi og m y-værdier).

Datamaterialet behøver ikke at være sorteret efter gruppenumre.

Af hensyn til regnemaskinens cellekapacitet må kvadratsummen af værdierne af hver variabel være mindre end 2^{2p+1} i hver inddelingsgruppe.

Teorien for regressionsanalysen forudsætter yderligere, at grupperne hver indeholder mindst 3 datasæt, og at SAKx > 0 i hver af disse grupper.

7. Dataadministration

Programmet er ufuldstændigt i den forstand, at det ikke indeholder nogen dataadministration. Rekvirenten må selv indlægge en sådan.

Dataadministrationen skal ved hvert indhop anbringe et datasæt i halvcellerne a til a+m+1:

I hac a anbringes $i \cdot 2^{-19}$, hvor i = 1,2,3,...,eller q,

i hac a+1 anbringes $x \cdot 2^{-p}$, og

i hac a+j+1 sættes $y_j \cdot 2^{-p}$, hvor j = 1,2,...,m

Her er x og y altså faste tal i halvceller - men bemærk, at rekvirenten selv vælger a og p.

Udhop sker normalt med 2D1o, men med 1D1o, hvis forrige datasæt var det sidste.

Dataadministrationen kan i ferritlageret disponere over halvcelle 0 - 315 og halvcellerne max(472+6q(m+1), 1o38) til 1791 samt over kromlekanalerne 1oo+6q til 4o4 og 448 til 51o.

Halvcellerne 472+6q(m+1) til 1o37 og kanalerne 98 til 98+6q er også ledige, indtil sidste datasæt er indlæst; men da benyttes disse afsnit og retableres ikke.

Noget lignende er tilfældet med halvcelle 1792 til 1983, hvor indlæseprogrammet i NL1 anbringes, hver gang styringsinformationen skal indlæses.

8. Outputdata

Fejludskrift kan forekomme som 8-kanals skrivemaskineoutput. Beregningsresultaterne fremkommer som 8-kanalsperforatoroutput, idet der udskrives et skema for hver afhængig variabel.

Efter overskriften følger hovedskemaet, hvis q første linier indeholder parameterskøn for de enkelte grupper, hvorefter der følger en tilsvarende totallinie, hvori alle grupperne indgår under ét. Disse q+1 linier har følgende indhold:

Kolonne	1:	gruppenummer, i
-	2:	antal datasæt, n
-	3:	middeltal, x.
-	4:	spredning, s(x)
-	5:	middeltal, y.
-	6:	spredning, s(y)
-	7:	spredning, s(y.)
-	8:	hældning, b
-	9:	position, y. - bx.
-	10:	residualspredning, $s(y\|x)$
-	11:	spredning, s(b)

De næste 4 linier indeholder:

b.	s.	s(b.)		
t_2	f_2	v_1^2	f_T	f_N
t_1	f_1	v_2^2	f_T	f_N
X^2	f			

Til slut udskrives a_1, a_2,...., a_q i én eller flere linier i nummerorden.

9. Eksempel

Her gives til eksempel et meget simpelt datamateriale, og det tilsvarende output.

i:	1	1	1	1	2	2	2	2	3	3	3	4 4 4
x:	4	5	6	7	3	3	4	6	2	4	6	1 5 6
y:	1	2	2	2	2	3	3	3	3	4	4	4 5 5

m = 1 og q = 4

Output:

Sag 1

Beregning og sammenfatning af 4 lineære regressioner på 1 uafhængig variabel.

variable nr 1

1	4	5.5oo	1.291	1.75o	0.5000	0.2500	0.3000	0.1000	0.3873	0.1732
2	4	4.000	1.414	2.750	0.5000	0.2500	0.1667	2.083	0.5401	0.2205
3	3	4.000	2.000	3.667	0.5774	0.3333	0.2500	2.667	0.4082	0.1443
4	3	4.000	2.646	4.667	0.5774	0.3333	0.2143	3.810	0.1543	0.04124
	14	4.429	1.742	3.071	1.207	0.3225	-0.01087	3.120	1.256	0.2000

0.2273 0.3541 0.06165

3.687	9	0.1024	3	6
-4.118	11	25.18	2	9
2.128	3			

-1.565 -0.2240 0.6926 1.693

---

10. Betjeningsvejledning

Programmet, der bl.a. findes på en kp.4-strimmel, indlæses ved hjælp af indlæseprogrammet i NL1, og når man har verificeret, at dette sidste står intakt i DASK, er fremgangsmåden følgende:

1) Sæt 8-kanals-omskifterne på 0 og 7. (Perforatoroutput kræver papirformat A4 på tværs med margin = 0 - stopsymbolet skal fungere som formularskift).
2) Sæt 18-ordren på oppakning.
3) Med indhop i hac 1987 indlæses programstrimlen, hvorefter maskinen stopper med 1987 A 30.
4) Indlæs dataadministrationen og eventuelle modificeringer.
5) Med indhop til hac 316 indlæses styringsstrimlen, hvorefter maskinen stopper i hac 318 med 319 A 30.
6) Indlæs datamaterialet ved start - efter endt indlæsning fremkommer 8-kanals perforatoroutput, hvorefter der stoppes i hac 378 med 316 A 30.
7) Maskinen står nu klar til ved start at begynde med punkt 5). (Hvis man ikke ønsker at indlæse styringsstrimlen ved fornyet kørsel, kan man manuelt foretage indhop i hac 319 og derefter fortsætte med instruksens punkt 6).

Fejludskrift: "overløb" på skrivemaskinen betyder, at en kvadratsum har oversteget 2^{2p+1}. IRC indeholder nu adressen på den famøse variabelværdi - stående i ferritlageret - som har fået bægeret til at flyde over. DASK er stoppet i hac 398 med 316 A 30, klar til at starte helt forfra ved start.

Hvis det sker, at maskinen efter dataindlæsningen "løber løbsk", skal man manuelt foretage indhop i hac 377 for at kunne begynde korrekt forfra. Dataadministrationen bliver herved retableret, idet den på dette tidspunkt er gemt på tromlen.

11. Modificeringsmuligheder

Indlæsningen af styringsinformation foregår ved hjælp af de 3 første hac i programmet, så hvis man lader programmet starte i hac 319 i stedet for hac 316, sker denne indlæsning ikke.

Således kan man ved at ændre indholdet af hac 378 fra 316 A 30 til 319 A 30 opnå, at programmet ved gentagen brug regner videre med den styringsinformation, der indlæstes inden første programgennemløb.

Som nævnt under omtalen af styringsinformationen er denne placeret i hac 399 til hac 4o4. Man kan da tilføre programmet styringsinformation ved at ændre indholdet af disse celler.