TECHNICAL MEMORANDUM

55

A program for optimal regression analysis

S. Grønås

Research Department

March 1982

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

C O N T E N T S

1. INTRODUCTION

The program REGR is built on the theory for linear multiple regression analysis, which is thoroughly described in textbooks on statistics. It deals with data containing two or more variables measured on a set of events or set of objects. For example, we may have a set of variables consisting of observed and forecasted temperatures at a given station and for a set of each day in a season. In general we have p variables and n events so that the data matrix contains n x p pieces of information.

Multiple regression analysis uses the method of least squares to establish linear relation between one of the variables y, called predictand, and the remaining variables x_i, predictors. Coefficients will be computed giving the relation:

$$y = \sum_{i=1}^{p-1} a_i x_i + e,$$ where e is the residual, a random variable representing what is not explained by the predictors. The multiple correlation coefficient will determine the fraction size of explained variance by the predictors from the relation:

$$var(y) = R^2 var(y) + var(e).$$

Very often the number of predictors is high and we are interested in a small subset which explains an optimal amount of the variance of the predictand. In computer libraries on statistics, the so-called screening method is often offered for this purpose. Here first the best single predictor will be picked out, then a second predictor, which together with the first predictor explains the highest amount of variance, will be added, then a third predictor and so on.

Our program will investigate all possible linear combinations of one, two, three and four variables and point out the five best relations for each step, i.e. the five best single predictors, the five best sets of two predictors etc. In this way we keep the number of predictors on a low level and get the simplest possible relations between predictand and the predictors. For p predictors the programm will consider $\frac{p(p-1)(p-2)(p-3)}{1 \times 2 \times 3 \times 4}$ different residuals before having a list of the best relations of four predictors. The computer time will therefore increase considerably with increasing number of predictors. We have just used CYBER 175 to run the program, for $p \leq 10$ the CPU time is a couple of seconds and for p=99, which is the maximum value, the CPU time is about 100 sec. The program was first written by Jack Nordø, The Norwegian Meteorological Institute.

2. DATA STRUCTURE

The program (executed in Cyber) reads the nxp data matrix from tape 50. All variables (predictand and predictors) for one event (one value of n) is read in one piece using the statement:

```
READ(50) (DAT(I),I=1,P)
```

where P is the number of variables. Maximum value of P is 99.

3. SUBROUTINE REGR

Usually a special program has to be written in order to provide the data matrix to tape 50. A subroutine version of REGR is then available and may be used in connection with this data sorting program.

The source program is found in file REGSUB, id=MOS*. A binary compiled version is stored in file REGOBJ, id=MOS.

First all the predictors are written to tape 50 using the statement:

```
WRITE(50) PRED,
```

where PRED is an array containing up to 99 predictors. Then a call statement may follow:

```
CALL REGR(NAME,N,NP,IRES,IPT,NPR,JPR)
```

NAME	problem name up to ten characters
N	number of cases in the sample
NP	number of variables (predictand + predictors)
IRES	=0 no output of estimates and residuals =1 output of estimates and residuals
JPR	integer array dimensioned as PRED and containing the number in the row of variables in PRED to be selected for the analysis.
IPT	the number of the element in JPR to be pointing at the variable in PRED to be picked out as predictand.
NPR	number of predictors selected.
Comment	The NPR+1 first elements in JPR are used. Let NPR+1=7, IPT=3 and the seven first elements of JPR be 7,8,23,24,84,85,86. Then JPR(3)=23, and variable 23 in PRED are chosen as predictand and variables 7,8,24,84,85,86 are predictors.

* This id will be changed to an id belonging to Product Development Group

A sketch of a program using the subroutine follows:-

```
      ATTACH,RRLDR,REGOBJ,ID=MOS.
      FTN(N=REG)
      LOAD(REG,RRLDR)
      EXECUTE.
      ....
      *EOR
      PROGRAM TEST(,TAPE50,...)
      DIMENSION PRED(99),JPR(99)
      ....
      DO 1 I=1,99
      PRED(I)=1
      JPR(I)=I
      1 CONTINUE

C     If there are many variables, it might be convenient to let JPR(I) be fixed.
      .....
C     Read data of interest
      .....
      Do 2 I=1,NN
C     NN is the number of cases
C     Within this loop the variables are defined and stored in PRED
      .....
      Write (50) PRED

      2  CONTINUE
      NAME=10H PROBL TEMP
      CALL REGR (NAME,NN,99,0,1,10,JPR)

C     NN observations
C     99 possible variables
C     No output of estimates and residuals
C     The first parameter of JPR gives the predictand
C     There are ten predictors in this selection
C     If the array JPR has not been changed variable 1
C     is the predictand and variables 2,3,4,5,6,7,8,9,10 and 11
C     are the 10 predictors

      END
*     EOR
```

4. PROGRAM REGR

When the data matrix is already stored on a permanent file (tape 50), the program REGR might be used. The source code are found on REGMAIN,id=MOS, and the program can be executed by the following control cards:

```
ATTACH,TAPE10,REGMAIN,ID=MOS,MR=1.
FTN(I=TAPE10)
LGO.
```

A set of cards must be available, all numbers are read in free format (numbers separated by commas).

1. card: NAME , a problem name up to 10 characters, format (A10)

2. card: N,NP,NPP,NS,NN,NTR,NFE

N : number of cases
NP : number of variables
NPP : number of variables after transformations
NS : number of selections ($\geq$1)
NN : if (NN.EQ.0) no variable names will be read
if (NN.EQ.1) variable names will be read
NTR : number of transgeneration cards (see card 4)
NFE : if (NFE.EQ.0) F format is used in output
if (NFE.EQ.1) E format is used in output

3. card(s): if (NN.EQ.0) this card will be skipped
if (NN.EQ.1) the variable name will be read using the format 16A4. This means that there are five letters for each variable and 70 characters on each card.

4. card(s): Transgeneration cards. If NTR=0 there will be no transgeneration cards. Otherwise there must be NTR transgeneration cards. Each card should have the following variables:

K1,K2,K3,K4,C,

K1 : the number of the variable to receive the new value (Index k below)
K2 : the number of the first operator (index I below)
K3 : the operation code (code N below)
K4 : the number of the second operator (index J below)
C : a constant. If no constant is needed, put C=0.0.

The codes have the following meaning:

N	Formula	Restrictions
11	X(K)=X(I)+X(J)	
12	X(K)=X(I)-X(J)	
13	X(K)=X(I)*X(J)	
14	X(K)=X(I)/X(J)	X(J)$\neq$0.0
15	X(K)=X(I)**2	
16	X(K)=X(I)**3	
17	X(K)=X(I)**4	
18	X(K)=SQRT(X(I)))	X(I)$\geq$0.0
19	X(K)=LOG(X(I)) (e)	X(I)>0.0
20	X(K)=EXP(X(I))	
21	X(K)=EXP(-X(I))	
22	X(K)=SIN(X(I))	
23	X(K)=ARCSIN(X(I))	-1,0$\leq$X(I)$\leq$1.0
24	X(K)=COS(X(I))	
25	X(K)=ARCCOS(X(I))	-1,0$\leq$X(I)$\leq$1.0
26	X(K)=TAN(X(I))	
27	X(K)=ARCTAN(X(I))	

The angles in the trigonometric functions are in radians.

For code 11,12,13 and 14 it is possible to change one of the operators to be the constant C. Instead of the variable number put 0.

Example : 25,23,13,0,2.718 mean: X(25)=X(23)*2.718.

Additional transformation could easily be built into the program.

5. card(s): Selection cards. There must be NS selection cards plus a termination card. Each card should have the following parameters:

IOUT,NPT,NPR,I1,I2,....INPR

IOUT : if IOUT.EQ.0 no estimates and residuals will be printed.
if IOUT.EQ.1 estimates and residuals will be printed.
NPT : Index of the dependent variable, predictand
NPR : Number of independent variable, number of predictors.
I1,I2...,INRR : Index of the independent variables, altogether NPR in number.

Example:

0,8,10,2,3,4,21,22,23,24,25,18,17
variable 8 is the predictand and ten variables with indexes 2,3,4,21,22,24,25,18,17 will be the predictors in this selection.

If there are more predictors than can be filled on one card, just continue on a second card.

The last card, the termination card, should just contain -1, for all parameters.

EXAMPLE: An example of a set up of cards is given below.

```
4  MOS,CM150000,T10.
   ACCOUNT,ECOMMO.
6  ATTACH,TAPE50,REGDATA8101,ID=MOS.
   ATTACH,RR,REGMAIN,ID=MOS.
8  FTN(I=RR)
   LGO.
10 TESTFIELD
   30,25,25,1,1,0,1
12 ZZ01ZZ02ZZ03ZZ04ZZ05ZZ06ZZ07ZZ08ZZ09ZZ10ZZ11ZZ12ZZ13ZZ14ZZ15ZZ16
   VAR15VAR16VAR17NONENONENONENONENONENONENONENONE
14 0,17,16,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16
   -1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,
```

Here there are 30 cases and 25 potential predictors. There is just one selection and no transformation of the variables. Variable names are read. The selection card says that variable 17 is the predictand and there are 16 predictors. There will be no printing of estimates and residuals.

5. An example of output from REGR

```
INPUT INFORMATION:
------------------

PROBLEM IDENTIFICATION:                          TESTFIELD
NUMBER OF OBSERVATIONS:                             30
NUMBER OF VARIABLES BEFORE TRANSFORMATIONS:         25
NUMBER OF VARIABLES AFTER TRANSFORMATIONS:          25
NUMBER OF SELECTIONS:                                1

-DATA TRANSFORMATIONS:
NONE
```

MULTIPLE REGRESSION.....TESTFIELD

SELECTION..... 1

NUMBER OF VARIABLES IN THE REGRESSION ANALYSIS: 17

EVALUATED IN THE ORDER:

17 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

MEANS

X17=	X 1=	X 2=	X 3=	X 4=	X 5=	X 6=	X 7=	X 8=	X 9=
.528E+04	.526E+04	.525E+04	.524E+04	.522E+04	.519E+04	.517E+04	.516E+04	.515E+04	.527E+04

X10=	X11=	X12=	X13=	X14=	X15=	X16=	X
.526E+04	.525E+04	.522E+04	.519E+04	.517E+04	.516E+04	.515E+04	

STANDARD DEVIATIONS

X17=	X 1=	X 2=	X 3=	X 4=	X 5=	X 6=	X 7=	X 8=	X 9=
.119E+03	.125E+03	.130E+03	.126E+03	.120E+03	.114E+03	.951E+02	.870E+02	.898E+02	.122E+03

X10=	X11=	X12=	X13=	X14=	X15=	X16=	X
.127E+03	.122E+03	.116E+03	.111E+03	.945E+02	.874E+02	.899E+02	

CORRELATION MATRIX

X17	1.0000															
X 1	-.2410	1.0000														
X 2	-.2000	.8746	1.0000													
X 3	-.0937	.6365	.8652	1.0000												
X 4	.0161	.3785	.5556	.8457	1.0000											
X 5	.1417	.1540	.2482	.5380	.8624	1.0000										
X 6	.3532	.0487	-.0146	.2079	.5826	.8547	1.0000									
X 7	.4533	-.0727	-.2354	-.2210	.0602	.3725	.7327	1.0000								
X 8	.5433	-.0108	-.1446	-.2798	-.1808	.0392	.3711	.8250	1.0000							
X 9	-.1921	.8569	.9976	.8691	.5623	.2523	-.0150	-.2419	-.1461	1.0000						
X10	-.1020	.6353	.8603	.9880	.8369	.5325	.2132	-.2005	-.2652	.8695	1.0000					
X11	-.0026	.4451	.6105	.8506	.9728	.8553	.6043	.1177	-.1380	.6155	.8594	1.0000				
X12	.1396	.2411	.3308	.5305	.7726	.9148	.8347	.4141	.0912	.3300	.5256	.8305	1.0000			
X13	.3158	.0629	.0841	.2013	.4202	.6677	.8091	.5889	.2838	.0827	.1938	.4937	.8496	1.0000		
X14	.5346	.0229	-.0135	-.0594	.0721	.3397	.6081	.6951	.5815	-.0207	-.0721	.1530	.5585	.8513	1.0000	
X15	.5814	-.0087	-.0824	-.2413	-.2831	-.1161	.1684	.4409	.6104	-.0830	-.2441	-.2292	.0692	.3949	.7524	1.0000
X16	.5418	.0360	.0008	-.1344	-.2755	-.2570	-.0684	.1393	.4196	.0093	-.1310	-.2717	-.1768	.0422	.3708	.8966
	X17	X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X10	X11	X12	X13	X14	X15

```
X15   1.0000
X16    .6135  1.0000
       X15    X16     X
```

OPTIMAL REGRESSION ANALYSIS — R IS THE MULTIPLE CORRELATION COEFFISIENT
ONE PREDICTOR CASE — SE THE STANDARD DEVIATION OF THE RESIDUAL

5 BEST RELATIONS

X17= .7928402E+00*X15	.1188656E+04	R= .581400	SE= .97E+02
X17= .7205097E+00*X 8	.1565370E+04	R= .543313	SE= .10E+03
X17= .7175971E+00*X16	.1580142E+04	R= .541765	SE= .10E+03
X17= .6738318E+00*X14	.1793204E+04	R= .534630	SE= .10E+03
X17= .6207518E+00*X 7	.2075385E+04	R= .453252	SE= .11E+03

OPTIMAL REGRESSION ANALYSIS — R IS THE MULTIPLE CORRELATION COEFFISIENT
TWO PREDICTOR CASE — SE THE STANDARD DEVIATION OF THE RESIDUAL

5 BEST RELATIONS

X17= .4912251E+00*X 6	.7531031E+00*X16	-.1143559E+04	R= .668191	SE= .89E+02
X17= .5276384E+00*X 7	.6465165E+00*X16	-.7763417E+03	R= .662615	SE= .89E+02
X17= .4876842E+00*X14	.5275406E+00*X16	.3707445E+02	R= .650109	SE= .91E+02
X17= .5085986E+00*X 8	.5044578E+00*X16	.5747478E+02	R= .643974	SE= .91E+02
X17= .8291427E+00*X 6	.7325059E+00*X15	-.2024743E+03	R= .635449	SE= .92E+02

OPTIMAL REGRESSION ANALYSIS

THREE PREDICTOR CASE

R IS THE MULTIPLE CORRELATION COEFFISIENT

SE THE STANDARD DEVIATION OF THE RESIDUAL

5 BEST RELATIONS

X17= -.2691414E+00*X 1 .5093606E+00*X 6 .7678262E+00*X16 .1028889E+03 R= .724915 SE= .82E+02

X17= -.7595900E+00*X 2 .7490059E+00*X 3 .9603047E+00*X15 .3897903E+03 R= .707512 SE= .84E+02

X17= -.2219861E+00*X 1 .5024377E+00*X 7 .6609738E+00*X16 .4472069E+03 R= .701873 SE= .85E+02

X17= -.2528000E+00*X 1 .4913646E+00*X14 .5387043E+00*X16 .1290647E+04 R= .701809 SE= .85E+02

X17= .7433322E+00*X 3 -.7950801E+00*X 9 .9588974E+00*X15 .6259145E+03 R= .701438 SE= .85E+02

OPTIMAL REGRESSION ANALYSIS

FOUR PREDICTOR CASE

R IS THE MULTIPLE CORRELATION COEFFISIENT

SE THE STANDARD DEVIATION OF THE RESIDUAL

5 BEST RELATIONS

X17= -.1084893E+01*X 2 .1214364E+01*X 3 -.1359473E+01*X13 .2112621E+01*X14 .7445609E+03 R= .781618 SE= .74E+02

X17= -.1059214E+01*X 2 .1176057E+01*X10 -.1361353E+01*X13 .2133114E+01*X14 .6914015E+03 R= .778053 SE= .75E+02

X17= -.7967512E+00*X 2 .8022326E+00*X 3 .6111157E+00*X 8 .6132925E+00*X16 -.1050521E+04 R= .771483 SE= .76E+02

X17= .8144446E+00*X 3 .6060822E+00*X 8 -.8547541E+00*X 9 .6276162E+00*X16 -.8440496E+03 R= .770022 SE= .76E+02

X17= -.1123913E+01*X 9 .1183167E+01*X10 -.1346778E+01*X13 .2138635E+01*X14 .1062776E+04 R= .767872 SE= .76E+02