A program for optimal regression analysis

S. Grønås

Research Department

March 1982

This paper has not been published and should be regarded as an Internal Report from ECMWF. Permission to quote from it should be obtained from the ECMWF.
CONTENTS

1. Introduction
2. Data Structure
3. Subroutine REGR
4. Program REGR
5. An example of output from REGR
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 \( n \times p \) 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:

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

\[
\text{var}(y) = R^2 \text{var}(y) + \text{var}(e).
\]

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 program 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 \( 1 \times 2 \times 3 \times 4 \) 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 \( p=10 \) 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 Nordb, 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**  
inger 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,REGOBJJ,ID=MOS.
FIN(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)=1
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=1OH 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 10 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.
FIN(I=TAPE10)
EOR
```

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: \( N, NP, NFP, NS, NN, NTR, NFE \)

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

3. card(s): if \( NN \equiv 0 \) this card will be skipped
   - if \( NN \equiv 1 \) 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 \( NTR \equiv 0 \) 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:

<table>
<thead>
<tr>
<th>( N )</th>
<th>Formula</th>
<th>Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>( X(k) = X(I) + X(J) )</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>( X(k) = X(I) - X(J) )</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>( X(k) = X(I) \times X(J) )</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>( X(k) = \frac{X(I)}{X(J)} )</td>
<td>( X(J) \neq 0.0 )</td>
</tr>
<tr>
<td>15</td>
<td>( X(k) = X(I)^{**2} )</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>( X(k) = X(I)^{**3} )</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>( X(k) = X(I)^{**4} )</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>( X(k) = \text{SQRT}(X(I)) )</td>
<td>( X(I) \geq 0.0 )</td>
</tr>
<tr>
<td>19</td>
<td>( X(k) = \text{LOG}(X(I)) ) (e)</td>
<td>( X(I) &gt; 0.0 )</td>
</tr>
<tr>
<td>20</td>
<td>( X(k) = \text{EXP}(X(I)) )</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>( X(k) = \text{EXP}(-X(I)) )</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>( X(k) = \text{SIN}(X(I)) )</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>( X(k) = \text{ARCSIN}(X(I)) )</td>
<td>(-1.0 \leq X(I) \leq 1.0 )</td>
</tr>
<tr>
<td>24</td>
<td>( X(k) = \text{COS}(X(I)) )</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>( X(k) = \text{ARCCOS}(X(I)) )</td>
<td>(-1.0 \leq X(I) \leq 1.0 )</td>
</tr>
<tr>
<td>26</td>
<td>( X(k) = \text{TAN}(X(I)) )</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>( X(k) = \text{ARCTAN}(X(I)) )</td>
<td></td>
</tr>
</tbody>
</table>

   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: \( X(25) = X(23) \times 2.718 \).

   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,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.
NPR    : Index of the dependent variable, predictand
NPR    : Number of independent variable, number of predictors.
I1,I2,...,INPR : 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.

MOS,CM150000,T10.
ACCOUNT,ECOMMO.
ATTACH,TAP150,REGDATA8101,ID=MOS.
ATTACH,RP,REMAIN,ID=MOS.
FTN(I=RR)
LGO.
TESTFIELD
30,25,25,1,1,0,1
ZZ01ZZ02ZZ03ZZ04ZZ05ZZ06ZZ07ZZ08ZZ09ZZ10ZZ11ZZ12ZZ13ZZ14ZZ15ZZ16
VAR15,VAR16,VAR17,NONE,NONE,vNENONE,NENONE,NENONE,NENONE
0,17,16,12,3,4,5,6,7,8,9,10,11,12,13,14,15,16
-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,

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

**Selection:**

1. **Variables in the Regression Analysis:**
   - Number of variables evaluated: 17

**Means:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>X17</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X1</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X2</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X3</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X4</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X5</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X6</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X7</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X8</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X9</td>
<td>52.64E+04</td>
</tr>
</tbody>
</table>

**Standard Deviations:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>X17</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X1</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X2</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X3</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X4</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X5</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X6</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X7</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X8</td>
<td>52.64E+04</td>
</tr>
<tr>
<td>X9</td>
<td>52.64E+04</td>
</tr>
</tbody>
</table>

**Correlation Matrix:**

<table>
<thead>
<tr>
<th>Variable</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
<th>X9</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.0000</td>
<td>-0.2000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
<td>0.3532</td>
<td>0.6533</td>
<td>0.9984</td>
<td>0.5618</td>
</tr>
<tr>
<td>X2</td>
<td>-0.2000</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
<td>0.3532</td>
<td>0.6533</td>
<td>0.9984</td>
<td>0.5618</td>
</tr>
<tr>
<td>X3</td>
<td>-0.0937</td>
<td>-0.0937</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
<td>0.3532</td>
<td>0.6533</td>
<td>0.9984</td>
</tr>
<tr>
<td>X4</td>
<td>0.1016</td>
<td>0.1016</td>
<td>-0.0937</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
<td>0.3532</td>
<td>0.6533</td>
</tr>
<tr>
<td>X5</td>
<td>0.1547</td>
<td>0.1547</td>
<td>0.1016</td>
<td>-0.0937</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
<td>0.3532</td>
</tr>
<tr>
<td>X6</td>
<td>0.3532</td>
<td>0.3532</td>
<td>0.1547</td>
<td>0.3532</td>
<td>-0.0937</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
<td>0.1547</td>
</tr>
<tr>
<td>X7</td>
<td>0.6533</td>
<td>0.6533</td>
<td>0.3532</td>
<td>0.6533</td>
<td>0.1547</td>
<td>-0.0937</td>
<td>1.0000</td>
<td>-0.0937</td>
<td>0.1016</td>
</tr>
<tr>
<td>X8</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9984</td>
<td>0.9984</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>X9</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.5618</td>
<td>0.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

The table above represents the correlation matrix for the variables listed. Each entry shows the correlation between two variables, with values ranging from -1.0000 to 1.0000, indicating the strength and direction of the relationship between the variables.
OPTIMAL REGRESSION ANALYSIS
ONE PREDICTOR CASE

5 BEST RELATIONS

\[
\begin{align*}
x_{17} &= 0.7928402E+00 \times x_{15} + 0.1188656E+04 \quad R = 0.501400 \quad SE = 0.97E+02 \\
x_{17} &= 0.7205097E+00 \times x_{8} + 0.1565370E+04 \quad R = 0.543313 \quad SE = 0.10E+03 \\
x_{17} &= 0.7175971E+00 \times x_{16} + 0.1580142E+04 \quad R = 0.541765 \quad SE = 0.10E+03 \\
x_{17} &= 0.6738318E+00 \times x_{14} + 0.1793204E+04 \quad R = 0.534630 \quad SE = 0.10E+03 \\
x_{17} &= 0.6207518E+00 \times x_{7} + 0.2075385E+04 \quad R = 0.493222 \quad SE = 0.11E+03 
\end{align*}
\]

OPTIMAL REGRESSION ANALYSIS
TWO PREDICTOR CASE

5 BEST RELATIONS

\[
\begin{align*}
x_{17} &= 0.4912251E+00 \times x_{6} + 0.7531031E+00 \times x_{16} + 0.1143599E+04 \quad R = 0.662191 \quad SE = 0.99E+02 \\
x_{17} &= 0.5276384E+00 \times x_{7} + 0.6465165E+00 \times x_{16} + 0.7753417E+03 \quad R = 0.662515 \quad SE = 0.99E+02 \\
x_{17} &= 0.4875842E+00 \times x_{14} + 0.5275406E+00 \times x_{16} + 0.3707445E+02 \quad R = 0.650109 \quad SE = 0.91E+02 \\
x_{17} &= 0.5085986E+00 \times x_{8} + 0.5044578E+00 \times x_{16} + 0.3747478E+02 \quad R = 0.643974 \quad SE = 0.91E+02 \\
x_{17} &= 0.3291427E+00 \times x_{6} + 0.7325096E+00 \times x_{15} + 0.2024743E+03 \quad R = 0.634449 \quad SE = 0.92E+02 
\end{align*}
\]
### OPTIMAL REGRESSION ANALYSIS

#### THREE PREDICTOR CASE

<table>
<thead>
<tr>
<th>5 BEST RELATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X17 = -2691414E+00*X 1 +5093606E+00*X 6 =7678262E+00*X16 +1028883E+03</td>
</tr>
<tr>
<td>X17 = -7595930E+00*X 2 +7490059E+00*X 3 +9603047E+00*X15 +3897033E+03</td>
</tr>
<tr>
<td>X17 = -2219661E+00*X 1 +5024377E+00*X 7 +6609730E+00*X16 +4472069E+03</td>
</tr>
<tr>
<td>X17 = -2528000E+00*X 1 +4913646E+00*X14 +5387043E+00*X16 +1290547E+04</td>
</tr>
<tr>
<td>X17 = -7433322E+00*X 3 -7950801E+00*X 9 +9588974E+00*X15 +6259145E+03</td>
</tr>
</tbody>
</table>

### OPTIMAL REGRESSION ANALYSIS

#### FOUR PREDICTOR CASE

<table>
<thead>
<tr>
<th>5 BEST RELATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>X17 = -1086893E+01*X 2 +1214364E+01*X 3 -1359473E+01*X13 +2112621E+01*X14 +7443609E+03</td>
</tr>
<tr>
<td>X17 = -1059214E+01*X 2 +1176057E+01*X10 -1361353E+01*X13 +2133114E+01*X14 +6914015E+03</td>
</tr>
<tr>
<td>X17 = -7967512E+00*X 2 +8022326E+00*X 3 +611157E+00*X 8 +6132975E+00*X16 +1050521E+04</td>
</tr>
<tr>
<td>X17 = -8144446E+00*X 3 +6060822E+00*X 8 -8547541E+00*X 9 +6278162E+00*X16 +8440496E+03</td>
</tr>
<tr>
<td>X17 = -1123013E+01*X 9 +1183167E+01*X10 -1346778E+01*X13 +2158635E+01*X14 +1062776E+04</td>
</tr>
</tbody>
</table>

R IS THE MULTIPLE CORRELATION COEFFICIENT
SE THE STANDARD DEVIATION OF THE RESIDUAL