RMK 43

$x_i^2 - x_j^2$ and $\sqrt{x_i^2 + x_j^2}$

3-P

0	142	0
-142	0	-142
0	142	0

101	18	101
18	-180	18
101	18	101

5-P

0	150	182	150	0
-150	0	32	0	-150
-182	-32	0	-32	-182
-150	0	32	0	-150
0	150	182	150	0

136	48	25	48	136
48	-80	-126	-80	48
25	-126	-206	-126	25
48	-80	-126	-80	48
136	48	25	48	136

CONGRESSION - A FAST REGRESSION TECHNIQUE WITH A GREAT NUMBER OF FUNCTIONS OF ALL PREDICTORS

by Olov Lönnqvist

SMHI SWEDISH METEOROLOGICAL AND HYDROLOGICAL INSTITUTE

RMK 43

CONGRESSION - A FAST REGRESSION TECHNIQUE WITH A GREAT NUMBER OF FUNCTIONS OF ALL PREDICTORS

by Olov Lönnqvist

Issuing Agency	Report number
SMHI S-60176 Norrköping Sweden	RMK 43
	Report date December 1984

Author (s)

Olov Lönnqvist

Title (and Subtitle)

Congression - A Fast Regression Technique with a Great Number of Functions of All Predictors

Abstract

The term Congression is used for an entirely new technique for Multiple Regression Analysis. The merits of the new technique are obvious. It is made possible to introduce a vast number of derived predictors in the analysis without prolonging the computing time.

In the examples given in the paper, 6 predictors are increased to 1068, and the analysis carried out in less than 25 per cent of the time now needed for 6 predictors.

The main features of Congression are the Grouping of data in Parties and the use of Grouping Diagrams as an interface towards a great number of pre-prepared Potential Functions of any two predictors; x_i^2, $x_i x_j$, $\sin x_i$ and $\sqrt{x_i^2 + x_j^2}$, are examples of such functions.

Key words

Multiple Regression
Nonlinear Regression
Derived Predictors
Grouping of data
Test Program for Regression
Empirical Functions

Supplementary notes	Number of pages	Language
	30	English

ISSN and title

0347-2116 SMHI Reports Meteorology and Climatology

Report available from:
SMHI
S-601 76 Norrköping

C O N T E N T S

1. INTRODUCTION

1.1 Original and derived predictors

In many applied sciences it is of great interest to find the relationship between one variable, here called the predictand, and a large number of other variables, predictors, which are known or suspected to influence the value of the predictand. The relationship, whether really functional or not, is given the form of a function which could be either linear or nonlinear in the predictors.

These adjectives, linear and nonlinear, frequently appear in the following text. It is therefore essential to explain the meaning of these words in this paper. We shall deal here only with such mathematical relationships which can be written as a sum of terms and which are in that sense linear. Our use of linear and nonlinear refers instead to the terms in the polynomial. With this definition, the relationship is said to be linear if it is a linear function of the original predictors $x_1, x_2 \dots x_n$.

It is called nonlinear if it also contains additional predictors, which are derived from the original ones. We shall mainly deal with derived additional predictors which are functions of one or two predictors, such as x_i^2, $x_i x_j$, $\sin x_i$, and $\sqrt{x_i^2 + x_j^2}$.

For many applications it might be essential to examine the possible relationship also with such derived predictors. Historical data are only available for certain parameters, but it is often suspected or even known that nonlinear functions of the same parameters are just as relevant in describing the unknown relationship.

Up til now it has been considered totally impossible to include a great number of additional predictors derived from all original predictors at the same time in a regression analysis, even after the advent of very fast computers. The necessary but highly unsatisfactory solution has been to include just a few of them. A somtimes difficult subjective choice has been needed.

This is no longer true, thanks to a quite new technique - Congression.

It will be shown in the present paper that it is indeed possible to include a large, almost unlimited, number of derived predictors in the analysis without loosing time in comparison with present methods.

A simplified version of the Congression technique suffices for solving the linear problem. Computing time is radically reduced. Calculations are easy to perform. Therefore, the Linear Congression technique is especially well fitted for computation by hand.

1.2 The term Congression

As was pointed out by DRAPER and SMITH (1981), the reasons for the choice of the term "regression" about hundred years ago were indeed vague and the word soon lost its proper meaning. The term "congression" seems more well-founded. It not only is a short word for "con-regression"; it indicates that data are in fact congregated and added together into groups which will be called Parties.

It is the Party Representatives - so to speak - which take Congress decisions; it is not all the individuals in the population. Or to be back in mathematical terminology - while in regression all data are used for computing the correlations and by that obtaining the regression coefficients, only Party mean-values are used for obtaining Congression coefficients.

The method recommended in this paper, though in general just called Congression, should in fact, to be more specific, be termed Five-Party Congression. Other methods will be demonstrated as well; Two-Party and Three-Party Congression. A detailed presentation of these simpler methods will be given and is thought to make an excellent introduction to the new and very special technique of Congression in general.

1.3 The Problem

We are going to study the relationship between one predictand Y and n predictors x_1, x_2 ... x_n. These predictors have been normalized. Thus

$$x_i = (X_i - \bar{X}_i)/\sigma_i \quad ,$$

where $\bar{X}_i$ and σ_i were based on N values; X_{i1}, X_{i2} ... X_{iN} .

The linear regression problem is now to estimate, by the method of least squares, the regression coefficients b_i in the equation

$$Y = \sum_{x=1,n} b_i x_i + \varepsilon \quad .$$

The percentage variation explained is often denoted R^2 and expressed as a percentage by multiplication by 100. It will here be referred to as the Variance Reduction.

The nonlinear problem in our study is to find all significant coefficients in the equation

$$Y = \sum_{\substack{i=1,n-1 \\ j=2,n \\ (j>i)}} \sum_{k=1,m} b_{ijk} f_k(x_i, x_j) + \varepsilon .$$

That means a study of all possible pairs of the predictors x_i and x_j.

There are $n(n-1)/2$ such pairs.

For each one of these pairs, as many as m functions are formed. By that, the number of predictors is drastically increased from n, initially, to

$$m\,n\,(n-1)/2 .$$

This means that if we want to study 70 different functions of each pair, and if the number of original predictors is 6, there will now be as many as 1050 predictors to analyse.

It will be shown that in spite of this very high figure, such an analysis can be carried out in reasonable time; in fact even in shorter time than is required for solving the linear problem with traditional methods.

The technique is called Congression. Before entering in a description of that technique and the tests which prove its merits, let us start looking at some simpler Congression methods which are easier to describe by means of numerical examples.

2. DETAILED PRESENTATION OF SOME SIMPLE METHODS

2.1 Two-Party Linear Congression

The values in N cases of each predictor are grouped into two Parties, designated P_i^+ and P_i^-. The grouping depends simply on whether x_i is positive or negative. Subsequently the N values of the predictand Y are referred to either Party with respect to the same predictor x_i.

$$\begin{cases} Y_i^+ = \frac{1}{N_i^+} \sum Y & \text{for } x_i \geq 0 \\ Y_i^- = \frac{1}{N_i^-} \sum Y & \text{for } x_i < 0 \end{cases}$$

where

$$N_i^+ + N_i^- = N \ .$$

For the predictors on the other hand, a common designation is used for all of them, namely

$$\begin{cases} x^+ = \frac{1}{\nu} \sum x \Delta x & \text{for } x_i > 0 \\ x^- = \frac{1}{\nu} \sum x \Delta x & \text{for } x_i < 0 \end{cases}$$

where x is normally distributed, and $\nu \rightarrow \infty$.

By this assumption

$$\begin{cases} x^+ = 0.8 \\ x^- = -0.8 \end{cases} .$$

(A more accurate value can be found from the equation

$$\frac{1}{\sqrt{2\pi}} \left[e^{-\frac{1}{2} x^2} + \int_0^x e^{-\frac{1}{2} x^2} \right] = \frac{1}{2}$$

Graphical interpolation gives $x = \pm 0.8031$. The table of the probability integral used for this purpose is the one reproduced as Annex VI by CONRAD and POLLAK (1950).)

Now the number of cases in the problem is drastically reduced from N to 2. For each one of those 2 cases, we have now got n predictands but only one predictor!

Party Name	Case No	Predictands	Predictor	Weight
P^+	1	$Y_1^+ , Y_2^+ \dots Y_n^+$	0.8	$w_i^+ = N_i^+/N \sim 0.5$
P^-	2	$Y_1^- , Y_2^- \dots Y_n^-$	-0.8	$w_i^- = N_i^-/N \sim 0.5$

The regression-coefficient estimates are obtained from the usual equation, which now takes the following form:

$$b_i = \frac{\sum xY_i}{\sum x^2} = \frac{0.80\ (w_i^+Y_i^+ - w_i^-Y_i^-)}{0.5\ (0.80^2 + 0.80^2)} \quad .$$

This can be written

$$b_i = \frac{1.25}{N} \left(\sum_{x_i \geq 0} Y - \sum_{x_i < 0} Y \right) \quad .$$

The following numerical example demonstrates the simplicity of the method.

Assumptions There are 20 cases. The data sets, x_1, x_2 and x_3, are all normally distributed random values, normalized and multiplied by 100. The predictand chosen for the test is an exact linear function of two of the predictors. Thus, let us assume that

$$Y = 0.8\ x_1 - 0.2\ x_3$$

for all the 20 cases.

Data

Case No.	x_1	x_2	x_3	Y
1	-66	136	-204	-12
2	82	107	-78	81
3	-10	47	-19	-4
4	-76	-190	179	-97
5	-130	-84	24	-109
6	3	-36	69	-12
7	-1	-31	-142	28
8	-195	-96	-48	-146
9	7	179	90	-13
10	-14	94	20	-15
11	29	-111	47	13
12	151	10	-69	135
13	-61	-67	158	-80
14	250	56	18	196
15	-53	84	-121	-18
16	-22	31	-24	-13
17	61	-158	104	28
18	51	-102	-25	46
19	118	68	-100	114
20	-121	63	120	-121

First Round The predictand Y_1 is obtained from Y by either keeping the Y-value (if $x_1 \geq 0$) or changing its sign (if $x_1 < 0$). The two other predictands, Y_2 and Y_3, are obtained in the same way. (It could be said that Y_1, Y_2 and Y_3 are the predictand Y as seen from the predictors x_1, x_2 and x_3, or rather, maybe, as filtered by them.)

We obtain

No.	Y_1	Y_2	Y_3
1	12	-12	12
2	81	81	-81
3	4	-4	4
:	:	:	:
20	121	-121	121
Mean	58.8	-21.0	33.0
b_i	0.734		

The mean with the highest absolute value, 58.8 , is obtained for Y_1 . Thus 0.73 x_1 is the first term in the regression. Hence, -0.73 x_1 is used to reduce the predictand. RP1 (the first residual predictand) is the result of the first round.

Final result The procedure is repeated, and we obtain the following table which shows how computations proceed through four rounds.

No.	OP	$-0.73x_1$	RP1	$+0.22x_3$	RP2	$-0.07x_1$	RP3	$-0.02x_3$	RP4
1	-12	48	36	-45	-9	5	-4	4	0
2	81	-60	21	-17	4	-6	-2	2	0
3	-4	7	3	-4	-1	1	0	0	0
4	-97	55	-42	39	-3	5	2	-4	-2
5	-109	95	-14	5	-9	9	0	0	0
6	-12	-2	-14	15	1	0	1	-1	0
7	28	1	29	-31	-2	0	-2	3	1
8	-146	142	-4	-11	-15	14	-1	1	0
9	-13	-5	-18	20	2	0	2	-2	0
10	-15	10	-5	4	-1	1	0	0	0
11	13	-21	-8	10	2	-2	0	-1	-1
12	135	-110	25	-15	10	-11	-1	1	0
13	-80	44	-36	35	-1	4	3	-3	0
14	196	-183	13	4	17	-18	-1	0	-1
15	-18	39	21	-27	-6	4	-2	3	1
16	-13	16	3	-5	-2	2	0	0	0
17	28	-45	-17	23	6	-4	2	-2	0
18	46	-37	9	-6	3	-4	-1	1	0
19	114	-86	28	-22	6	-8	-2	2	0
20	-121	88	-33	26	-7	8	1	-3	2

In excellent agreement with the assumption, we find

$$Y = 0.80\ x_1 - 0.20\ x_3\ .$$

As demonstrated in the example, the approximate method for finding the best predictor in each round, and the corresponding regression coefficient, has the advantage of using additions only and no multiplications. Although therefore this estimate is less accurate than by the traditional method, the residual is correct as such, since it is obtained from correct values of the predictand and the predictor involved, using all cases.

2.2 Three-Party Linear Congression

Next obvious step is to proceed from two to three Parties. Surprisingly enough, this does not lead to a more complicated method but a simpler one. Less additions are needed.

This is again best demonstrated by an example. Let us choose the same sets of data as before. Columns to the left in the following table demonstrate the grouping of predictors in three parties: + , o , and - . Also given are the three predictands corresponding to the predictors x_1, x_2 and x_3 , respectively.

	Parties according to			The corresponding predictands in the First Round		
No.	x_1	x_2	x_3	Y_1	Y_2	Y_3
1	-	+	-	12	-12	12
2	+	+	-	81	81	-81
3	o	o	o	0	0	0
4	-	-	+	97	97	-97
5	-	-	o	109	109	0
6	o	o	+	0	0	-12
7	o	o	-	0	0	-28
8	-	-	o	146	146	0
9	o	+	+	0	-13	-13
10	o	+	o	0	-15	0
11	o	-	o	0	-13	0
12	+	o	-	135	0	-135
13	o	-	+	0	80	-80
14	+	o	o	196	0	0
15	o	+	-	0	-18	18
16	o	o	o	0	0	0
17	o	-	+	0	-28	28
18	o	-	o	0	-46	0
19	+	+	-	114	114	-114
20	-	+	+	121	-121	-121
Mean				<u>50.6</u>	18.1	-31.2

Before compiling the table a decision had to be taken as to the Party limits. A straight-forward division into equal parts (33;33;33 %) would not be the optimum one. The optimum is (27;46;27). This will be shown in a later section. Since the predictors are normalized, the limits can be fixed to ±0.613. This value is derived from a tabulation of the probability integral. It holds true for normal distributions and shall be used whatever the actual distribution might be. It was used for forming the Parties in the table.

Because of the many zeros in the Y_i-sets, the number of additions is substantially reduced in comparison with the Two-Party method.

The tabulated probability integral also makes it possible to determine the typical (normal-distribution) average x-value within the two extreme Parties, namely +1.225 and - 1.225, respectively.

The Three-Party technique implies that the N cases this time are concentrated into three cases, as follows:

Party Name	Case No	Predictands	Predictor	Weight
P^+	1	$Y_1^+, Y_2^+ \dots Y_n^+$	1.225	$w_i^+ = N_i^+ / N \sim 0.27$
P^o	2	$Y_1^o, Y_2^o \dots Y_n^o$	0	$w_i^o = N_i^o / N \sim 0.46$
P^-	3	$Y_1^-, Y_2^- \dots Y_n^-$	-1.225	$w_i^- = N_i^- / N \sim 0.27$

Coefficients are obtained from

$$b_i = \frac{\sum xY_i}{\sum x^2} = \frac{1.225\ (w_i^+Y_i^+ - w_i^-Y_i^-) + 0}{0.27\ (1.225^2 + 1.225^2)} \quad ;$$

which can simply be written

$$b_i = \frac{1.51}{N} \left(\sum_{x_i \geq 0} Y - \sum_{x_i < 0} Y \right) .$$

The best estimate reported in the table, 50.6, equals 100 times the parenthesis divided by N. Multiplication by 1.51 gives $b_1 = 0.764$, a value that differs little from the value 0.73 in the Two-Party case. The computation goes on in much the same way, and again the final result is in good agreement with the assumption.

2.3 Three-Party Nonlinear Congression

Data grouped in 3 groups is a minimum requirement for analysing nonlinear relationships. However, a new ingenious approach must be applied in order to make it feasible. The idea is that the original predictors are dealt with two and two. Hence, if there are n original predictors, there are n(n-1)/2 pairs available for study.

Two steps are now needed in each round. The First Step is to find which one of the various Pairs of predictors that gives the "best" description of the predictand by explaining more of the variance than any other Pair. In other words, we have to find the "best" Empirical Function, $E_{3x3}(x_i,x_j)$, for describing the predictand, or any residual predictand, as a non-formulated function of two of the original predictors. The subscript, 3x3, indicates that the function is given by 9 discrete values and not in the form of an equation.

The second step is to find which one in a predefined set of potential mathematical functions that explains the variance of the Empirical function "better" than any other function.

The ingenious point here is that these functions are presented in the same format as the Empirical function, thus as $F_{3x3}(x_i,x_j)$.

Returning now to the first step, the following table will show how to define the nine groups formed by the two predictors, x_i and x_j, both of which are divided into three Parties. (x-values are multiplied by 100.)

Group	Party combination	Definition	
$G1_{ij}$	− −	$x_i < -61$	$x_j < -61$
$G2_{ij}$	− o		$-61 \le x_j \le +61$
$G3_{ij}$	− +		$+61 < x_j$
$G4_{ij}$	o −	$-61 \le x_i \le +61$	$x_j < -61$
$G5_{ij}$	o o		$-61 \le x_j \le +61$
$G6_{ij}$	o +		$+61 < x_j$
$G7_{ij}$	+ −	$+61 < x_i$	$x_j < -61$
$G8_{ij}$	+ o		$-61 \le x_j \le +61$
$G9_{ij}$	+ +		$+61 < x_j$

Again, the new technique is certainly best explained by a numerical example.

<u>Assumptions</u> Let us use 20 cases and the same predictors, x_1, x_2 and x_3, as before. In this example the predictand will of course be a polynomial with nonlinear terms, say

$$Y = 0.67\, x_2 x_3 + 0.33\, x_3^2 .$$

<u>First Step</u> In the first round we obtain the following table. <u>OP</u> is as before, the original predictand.

No.	<u>OP</u>	(x_1,x_2)			(x_1,x_3)			(x_2,x_3)		
					P a i r s					
1	-46	-	+	G3	-	-	G1	+	-	G7
2	-36	+	+	G9	+	-	G7	+	-	G7
3	-5	o	o	G5	o	o	G5	o	o	G5
4	-120	-	-	G1	-	+	G3	-	+	G3
5	-11	-	-	G1	-	o	G2	-	o	G2
6	-1	o	o	G5	o	+	G6	o	+	G6
7	96	o	o	G5	o	-	G4	o	-	G4
8	37	-	-	G1	-	o	G2	-	o	G2
9	134	o	+	G6	o	+	G6	+	+	G9
10	14	o	+	G6	o	o	G5	+	o	G8
11	-28	o	-	G4	o	o	G5	-	o	G2
12	11	+	o	G8	+	-	G7	o	-	G4
13	12	o	-	G4	o	+	G6	-	+	G3
14	8	+	o	G8	+	o	G8	o	o	G5
15	-19	o	+	G6	o	-	G4	+	-	G7
16	-3	o	o	G5	o	o	G5	o	o	G5
17	-74	o	-	G4	o	+	G6	-	+	G3
18	19	o	-	G4	o	o	G5	-	o	G2
19	-12	+	+	G9	+	-	G7	+	-	G7
20	98	-	+	G3	-	+	G3	+	+	G9

The following Group Diagrams for the Empirical Functions are now compiled. The arithmetic mean of predictand values falling in each box is given as well as a weighting factor based on the number of cases. Figures below the diagrams give the Relative Variance. That means that the variance presented in the diagrams and calculated from the mean values and the corresponding weights has been divided by the variance of the predictand itself.

Evidently, in this case the Pair (x_2, x_3) is by far the best one to describe the predictand.

$E_{3x3}\ (x_1, x_2)$

G1 -31 .15	G2 0 .00	G3 26 .10
G4 -18 .20	G5 22 .20	G6 43 .15
G7 0 .00	G8 10 .10	G9 -24 .10

12.5 %

$E_{3x3}\ (x_1, x_3)$

G1 -46 .05	G2 13 .10	G3 -11 .10
G4 39 .10	G5 -1 .25	G6 18 .20
G7 -12 .15	G8 8 .05	G9 0 .00

6.4 %

$E_{3x3}\ (x_2, x_3)$

G1 0 .00	G2 4 .20	G3 -61 .15
G4 54 .10	G5 0 .15	G6 -1 .05
G7 -28 .20	G8 14 .05	G9 116 .10

41.6 %

Second Step In each round the "best" diagram found in the first step of the round is utilized in the second step to find out which one of the functions in a set of Potential Functions that should be selected as the best one to describe the Empirical Function.

Such a set has been prepared in advance. In the present example we shall assume that there are five functions available in the set; x_i ; x_j ; x_i^2 ; x_j^2 ; $x_i x_j$.

Three-by-three Group Diagrams for these functions can easily be constructed.

They take the following form:

x_i (→ x_j, ↓ x_i)

-122	-122	-122
0	0	0
122	122	122

x_j

-122	0	122
-122	0	122
-122	0	122

$x_i x_j$

150	0	-150
0	0	0
-150	0	150

x_i^2

150	150	150
0	0	0
150	150	150

x_j^2

150	0	150
150	0	150
150	0	150

For comparison purposes these diagrams are stored in normalized form together with the conversion constants which are needed for the further calculations. In the table below the normalized values are seen in the right columns (multiplied by 100) and also the conversion factors (in brackets). OP-values are deviations from the mean.

The second step of the first round gives the following results.

Group	$\underline{OP}_{2,3} - 4$	$w_{2,3}$	Functions				
			x_2 (1.111)	x_3 (1.111)	x_2^2 (1.338)	x_3^2 (1.338)	x_2x_3 (1.233)
G1	-	0.00	-136	-136	92	92	185
G2	0	0.20	-136	0	92	-108	0
G3	-65	0.15	-136	136	92	92	-185
G4	50	0.10	0	-136	-108	92	0
G5	-4	0.15	0	0	-108	-108	0
G6	-5	0.05	0	136	-108	92	0
G7	-32	0.20	136	-136	92	92	-185
G8	10	0.05	136	0	92	-108	0
G9	112	0.10	136	136	92	92	185
Correlation coefficient:			0.16	0.03	-0.09	0.00	<u>0.33</u>

Obviously, the product, x_2x_3, is the "best" function of the tested ones. The Congression Coefficient for this term is obtained as follows:

$$b_{x_2x_3} = 0.33 \cdot 1.233 = 0.41$$

The demonstrated procedure is repeated for each round; a first step by which the appropriate Pair is found, and a second step for finding the best function and the proper coefficient.

Final Result The computation of the first 4 rounds are summarized in the following table. The next 4 ones give further reduction terms, which amount to

$$-0.06\ x_2x_3; -0.03\ x_3^2;\ -0.03\ x_2x_3;\ -0.01\ x_3^2.$$

In all those rounds, $b_o = 0$.

Table showing the result of the first 4 rounds in the example.

No.	OP	$-0.41\ x_2x_3$	RP1	$-0.20\ x_3^2$	RP2	$-0.15\ x_2x_3$	RP3	$-0.08\ x_3^2$	RP4
1	-46	114	68	-83	-15	42	27	-33	-6
2	-36	34	-2	-12	-14	12	-2	-5	-7
3	-5	4	-1	-1	-2	1	-1	0	-1
4	-120	139	19	-64	-45	51	6	-26	-20
5	-11	8	-3	-1	-4	3	-1	0	-1
6	-1	10	9	-10	-1	4	3	-4	-1
7	96	-18	78	-40	38	-7	31	-16	15
8	37	-19	18	-5	13	-7	6	-2	4
9	134	-66	68	-16	52	-24	28	-6	22
10	14	-8	6	-1	5	-3	2	0	2
11	-28	21	-7	-4	-11	8	-3	-2	-5
12	11	3	14	-10	4	1	5	-4	1
13	12	43	55	-50	5	16	21	-20	1
14	8	-4	4	-1	3	-2	1	0	1
15	-19	42	23	-29	-6	15	9	-12	-3
16	-3	3	0	-1	-1	1	0	0	0
17	-74	67	-7	-22	-29	25	-4	-9	-13
18	19	-11	8	-1	7	-4	3	0	3
19	-12	28	16	-20	-4	10	6	-8	-2
20	98	-31	67	-29	38	-11	27	-12	15
b_o	4		22		2		8		0

The total result, after 8 rounds, equals

$$Y = 0.65\ x_2x_3 + 0.32\ x_3^2$$

in good agreemant with the assumption.

Comment

In view of the remarkable success of these examples, it must be observed that 20 cases are in fact much too few to warrant a successful analysis, whether by regression or congression. Although the predictor values were taken at random, it was checked that they did not deviate too much from normal distribution. This explains why data behaved faily well, which would certainly not always be the case with such a small number of data.

On the other hand, it should also be noted, that intercorrelation does not seem to make any harm. It can even be accepted, as in the last example, that there are boxes with no cases at all.

Potential Functions for Three-Party Nonlinear Congression

Functions of one predictor — Primary functions

Functions of both predictors — Products of primary functions; Some examples of convenient functions (though unnecessary)

-136	-136	-136
0	0	0
136	136	136

x_i

185	0	-185
0	0	0
-185	0	185

$x_i x_j$

-192	-96	0
-96	0	96
0	96	192

$x_i + x_j$

0	-96	-192
96	0	-96
192	96	0

$x_i - x_j$

-136	0	136
-136	0	136
-136	0	136

x_j

-185	0	-185
0	0	0
185	0	185

$x_i x_j^2$

-61	-122	-182
61	0	-61
182	122	61

$2x_i + x_j$

-61	61	182
-122	0	122
-182	-61	61

$x_i + 2x_j$

92	92	92
-108	-108	-108
92	92	92

x_i^2

-185	0	185
0	0	0
-185	0	185

$x_i^2 x_j$

197	61	-73
-73	-73	-73
-73	61	197

$x_i(x_i + x_j)$

197	-73	-73
61	-73	61
-73	-73	197

$(x_i + x_j)x_j$

92	-108	92
92	-108	92
92	-108	92

x_j^2

181	-75	181
-75	-75	-75
181	-75	181

$x_i^2 x_j^2$

207	39	-130
-45	-45	-45
-130	39	207

$x_i(2x_i + x_j)$

207	-45	-130
39	-45	39
-130	-45	207

$(2x_i + x_j)\, x_j$

Special function of both predictors

101	18	101
18	-180	18
101	18	101

$\sqrt{x_i^2 + x_j^2}$

131	-11	131
-11	-154	-11
131	-11	131

$x_i^2 + x_j^2$

0	142	0
-142	0	-142
0	142	0

$x_i^2 - x_j^2$

In this table negative values are specially marked in order to better illustrate how distinctly these functions differ one from the other.

Note! In Three-Party Congression it is impossible to distinguish between x, $\sin x$, and $\mathrm{arctg}\ x$, on one hand, and between x^2, $|x|$, and $\cos x$, on the other.

3. THE RECOMMENDED METHOD - FIVE-PARTY CONGRESSION

3.1 Definitions

The above simple variants of the congression technique were described in detail by numerical examples in order to make the reader familiar with those approaches which are essential for the new technique.

For that reason a detailed presentation of the five-party method would be superfluous. We shall concentrate on the specific characteristics of the method.

One immediate question will be: How many parties are needed for a quite satisfactory regression analysis? By tests reported on in the next section, it will be shown that a five-party grouping gives quite acceptable results. There has been no time, so far, to investigate if much would be gained by increasing the number of parties to, say, six, seven or nine. Probably though, it would not be worth-while to do so.

After we have decided on the number of parties, the next problem is the selection of Party Intervals. It will be shown later that the Optimum grouping would be (10; 25; 30; 25; 10). However, a grouping frequently used in meteorology, (12.5; 25; 25; 25; 12,5), is almost as good. The latter has been chosen so far. Since it functions well, it might not be worth-while to make a change in the future.

The five Parties are defined as follows. As before, a probability-integral table has been used for obtaining the actual figures.

Party	Percentage	Interval	Standard mean value
P1	12.5	$x_i < -1.15034$	-1.665
P2	25.0	$-1.15034 < x_i < -0.31863$	-0.694
P3	25.0	$-0.31863 < x_i < +0.31863$	0
P4	25.0	$+0.31863 < x_i < +1.15034$	+0.694
P5	12.5	$+1.15034 < x_i$	+1.665

3.2 Potential Functions

On page 14, the Potential Functions for the three-party case were classified as Primary Functions, Products, Special Functions and Convenient Functions. We shall use the same classification here. Let us then start with the Primary Functions.

With the much better resolution offered by the five intervals, a larger number of functions can now be utilized. It is no longer impossible to differentiate between x, sin x and arctg x; or between x^2 , $|x|$ and cos x. In the following table we find specified the 8 primary functions which were chosen for the test runs presented in this paper. There are many other functions which could have been included as well, e.g. $\cos(1.813\ x)$, $1/(1-x^2)$, $\sin(2 \cdot 1.813\ x)$, $\cos(2 \cdot 1.813\ x)$.

(Note here and in the table, that 1.813 x radians is used in the functions so that the standard deviation of the angle equals 1.0, provided x has a rectangular distribution!)

Primary functions of x	P1	P2	P3	P4	P5
x	-1.655	-0.964	0.000	0.694	1.655
x^2	2.898	0.536	0.000	0.536	2.898
x^3	-5.35	-0.45	0.00	0.45	5.35
$\|x\|$	1.655	0.694	0.000	0.694	1.655
arctg 2x	-1.264	-0.910	0.000	0.910	1.264
arctg(2x-1.36)	-1.354	-1.213	-0.914	0.019	1.047
arctg(2x+1.36)	-1.047	-0.019	0.914	1.213	1.354
sin(1.813 x)	-0.373	-0.902	0.000	0.902	0.373

Since the primary functions can be functions either of x_i or x_j , this gives us 16 potential functions.

Going now to functions of both predictors, the functions above give us 64 product functions.

As to special functions of both predictors, only one was included in the test set, namely $\sqrt{x_i^2 + x_j^2}$. Others could of course have been included as well.

Whether or not also to include so-called "Convenient Functions" is really a question that can only be answered by experience. Such functions are unnecessary but might have the effect of shortening the computations by diminishing the number of rounds. This effect should be balanced against the increased handling time, which seems to be almost negligible.

In the test runs, only three convenient functions were included, namely $x_i + x_j$, $x_i - x_j$, and $(x_i + x_j)(x_i - x_j)$.

This means in total that as many as 84 Potential Functions were prepared, stored and used in the first Program Package utilized for the test runs.

4. TESTS

4.1 Test-Program Design

Scientific reports on the use of Congression for solving various problems in meteolology will be published in due course. For the purpose of the present paper specially designed test runs would be more interesting, provided they answer the following questions:

1) Is Congression comparable in quality with traditional regression for analysing linear relationships in general?

2) Is Congression, due to its partly approximate nature, inferior to traditional regression for analysing "messy" relationships?

3) Is Congression effective in finding the right predicors and their functional relationships in case they are nonlinear?

In the special tests which have been run at the SMHI computer, Congression has been compared with the standard method available at SMHI for Linear Regression, which is afast program, quite convenient to the user. It is based on a program for Stepwise Regression by Forward Selection available in the IMSL Program Library. Comparisons were made at the same time with both Two-Party and Three-Party Linear Congression. Of these two methods, the Three-Party variant (3-PL) is the fastest one and also gives slightly better results, although the difference in this respect is small.

As stated above, the tests were specially designed. The number of cases in each run was 1000. Seven predictor sets were constructed as follows:

One of the predictors, x_1, was taken entirely at random among numbers between 0 and 999, having a rectangular distribution.

In order to let the remaining predictors be affected of both autocorrelation and intercorrelation and to have typical random deviations from the normal distribution, those predictors were taken from various meteorological data, which happened to available when the program was written.

Some of the more or less random characteristics of these predictors are shown in the following two tables; the first showing the intercorrelations; the second one the five-party distributions as compared with Normal and Rectangular Distribution.

Note the very special predictor, labeled x_x. It is the "unknown" predictor. It will be involved in the second group of tests, but it will not be available for the analysis. This was meant to make those examples rather typical for "problems with messy data"; an expressive term used by DRAPER and SMITH (1981). That textbook also mentions that what is here called the "unknown" predictor, is called "latent" or "lurking" by other authors.

Correlations between test predictors

	x_1	x_2	x_3	x_4	x_5	x_6	x_x
x_1	100	-5	1	-1	-1	-1	-5
x_2	-5	100	-5	7	-11	-7	-5
x_3	1	-5	100	18	-23	19	7
x_4	-1	7	18	100	-20	17	5
x_5	-1	-11	-23	-20	100	-36	-7
x_6	-1	-7	19	17	-36	100	-11
x_x	-5	-5	7	5	-7	-11	100

Distribution

	P1	P2	P3	P4	P5
x_1	17	24	18	25	16
x_2	7	36	28	17	12
x_3	10	29	25	25	11
x_4	11	26	29	23	11
x_5	14	29	22	23	14
x_6	14	29	22	23	14
x_x	7	36	28	17	12
Normal	13	25	25	25	13
Rectangular	16	24	19	24	16

In the first 12 test runs, the relationship between the predictand and some of the predictors, as varied from one test to the other, is always given by an exact function. In the following 12 runs the same functional relationships are used again, but this time the unknown "lurking" predictor has been added to the function. Note that the "noise" added in that way is therefore of a sophisticated nature, since it is to some extent correlated with the given predictors.

4.2 Computing Time

The test runs gave a first estimate of the Computing Times typical for the compared methods. Since the 24 test runs were always performed in sequence, the times reported in the following table are all means of 24 runs. It is felt that the comparison is a fair one, although a somewhat different handling of data on peripheral units and other program differences might slightly influence the times registered by the computer.

Typical computing times

(1000 cases, one predictand)

Method	Number of Predictors	Total time (seconds)
Conventional Linear Regression (LIN)	6	90
Three-Party Linear Congression (3-PL)	6	12
Five-Party Nonlinear Congression (5-P)	6 + 1062	21

Evidently, the time differences are striking. It is almost incredible that the much more comprehensive calculations of Nonlinear Congression takes less than a quarter of the time needed for Linear Regression. You can also ask yourself why the very speedy linear congression technique has not been introduced long ago. The time saving is astonishing, as is indeed the simplicity of the technique.

4.3 Tests with exact solutions

No.			Variance Reduction
1		$0.4\ x_2 + 0.6\ x_3$	
	LIN	$0.40x_2 + 0.60x_3$	100
	3-PL	$0.40x_2 + 0.60x_3$	100
	5-P	$0.39x_2 + 0.61x_3$	100
2		$0.4\ x_5 + 0.6\ x_6$	
	LIN	$0.40x_5 + 0.60x_6$	100
	3-PL	$0.40x_5 + 0.59x_6$	100
	5-P	$0.41x_5 + 0.60x_6$	100
3		$1.0\ x_5 + 1.0\ x_6$	
	LIN	$1.00x_5 + 1.00x_6$	100
	3-PL	$0.99x_5 + 1.00x_6$	100
	5-P	$1.00x_5 + 1.00x_6$	100
4		$1.0\ x_5 - 1.0\ x_6$	
	LIN	$1.00x_5 - 1.00x_6$	100
	3-PL	$0.99x_5 - 1.00x_6$	100
	5-P	$0.98x_5 - 0.98x_6$	100
5		$1.0\ x_2x_5 - 1.0\ x_2x_6$	
	LIN	$-0.22x_2 - 0.14x_5\ -0.04$	2
	3-PL	$-0.19x_2 - 0.12x_5\ -0.04$	3
	5-P	$0.98x_2x_5 - 1.04x_2x_6$	100
6		$1.0\sqrt{x_5^2 + x_6^2}$	
	LIN	$0.15x_5 + 0.15x_6 + 0.04x_1\ +1.27$	7
	3-PL	$0.04x_5 + 0.04x_6 + 0.05x_1 - 0.04x_3 + 1.27$	5
	5-P	$1.00\sqrt{x_5^2 + x_6^2}$	100
7		$1.0(x_3+x_4)(x_3-x_4)$	
	LIN	-	0
	3-PL	-	0
	5-P	$0.98(x_3+x_4)(x_3-x_4)$	100
8		$1.0\ x_2\ \sin(1.813\ x_1)$	
	LIN	-	0
	3-PL	$0.04x_2 + 0.05x_3 + 0.04x_5 - 0.04$	1
	5-P	$1.00\,x_2 \sin(1.813\ x_1)$	100
9		$1.0\ \sin(1.813\ x_1)$	
	LIN	$0.58\ x_1\ +0.01$	64
	3-PL	$0.61\ x_1\ +0.01$	64
	5-P	$1.02\sin(1.813x_1)$	100

No.			Variance Reduction
10		$1.0\, x_5 \lvert x_6 \rvert$	
	LIN	$0.76x_5-0.23x_6-0.02$	75
	3-PL	$0.78x_5-0.20x_6-0.02$	75
	5-P	$1.00x_5 \lvert x_6 \rvert$	100
11		$1.0\sqrt{x_3^2+x_4^2}+1.0\, x_5 \lvert x_6 \rvert$	
	LIN	$0.75x_5-0.22x_6+0.10x_3+1.20$	47
	3-PL	$0.73x_5-0.20x_6+0.05x_3+0.03x_4+1.20$	48
	5-P	$0.96\sqrt{x_3^2+x_4^2}+0.99x_5 \lvert x_6 \rvert$	100
12		$1.0\, x_6 \text{arctg}\, 2x_5$	
	LIN	$-0.19x_6-0.12x_5-0.07x_4-0.33$	5
	3-PL	$-0.10x_6-0.05x_5-0.06x_4-0.33$	4
	5-P	$1.00x_6 \text{arctg}\, 2x_5$	100

4.4 Tests with a lurking predictor involved

13		$0.2\, x_2+0.3\, x_3+0.5\, x_x$	
	LIN	$0.17x_2+0.34x_3-0.06x_5-0.09x_6$	37
	3-PL	$0.16x_2+0.32x_3-0.06x_5-0.05x_6$	38
	5-P	$0.15x_2+0.27x_3-0.05x_5-0.07x_6$	37
14		$0.2\, x_5+0.3\, x_6+0.5\, x_x$	
	LIN	$0.14x_5+0.21x_6-0.03x_2+0.04x_3$	17
	3-PL	$0.14x_5-0.20x_6-0.04x_2+0.04x_3-0.04x_1+0.04x_4$	17
	5-P	$0.17x_5+0.21x_6-0.04x_2$	16
15		$1.0\, x_5+1.0\, x_6+1.0\, x_x$	
	LIN	$0.88x_5+0.84x_6$	49
	3-PL	$0.84x_5+0.84x_6$	49
	5-P	$0.81x_5+0.81x_6$	49
16		$1.0\, x_5-1.0\, x_6+1.0\, x_x$	
	LIN	$0.88x_5-1.16x_6$	74
	3-PL	$0.86x_5-1.19x_6$	74
	5-P	$0.97x_5-1.12x_6$	74
17		$1.0\, x_2x_5-1.0\, x_2x_6+1.0\, x_x$	
	LIN	$-0.29x_2-0.29x_5-0.20x_6-0.04$	4
	3-PL	$-0.30x_2-0.26x_5-0.18x_6-0.04$	4
	5-P	$0.84\, x_2x_5-1.00\, x_2x_6-0.09x_2-0.09x_6$	70
18		$1.0\sqrt{x_5^2+x_6^2}+1.0\, x_x$	
	LIN	$-0.08\, x_2+1.27$	0
	3-PL	–	0
	5-P	$0.78\sqrt{x_5^2+x_6^2}-0.08x_6-0.08x_2+0.17 \lvert x_5 \rvert$	26

No.			Variance Reduction
19		$1.0\,(x_3 + x_4)(x_3 - x_4) + 1.0\,x_x$	
	LIN	$0.23x_3$	1
	3-PL	-	0
	5-P	$0.97(x_3 + x_4)(x_3 - x_4)$	85
20		$1.0\,x_2 \sin(1.813\,x_1) + 1.0\,x_x$	
	LIN	$0.13x_3 - 0.14x_6 \quad -0.04$	2
	3-PL	$0.11x_3 - 0.12x_6 - 0.07x_1 - 0.05x_5 - 0.04$	2
	5-P	$0.97x_2 \sin(1.813x_1) - 0.80x - 0.08x$	32
21		$1.0 \sin(1.813x_1) + 1.0\,x_x$	
	LIN	$0.53x_1 - 0.09x_2 - 0.10x_5 - 0.13x_6 \quad -0.01$	21
	3-PL	$0.55x_1 - 0.09x_2 - 0.08x_5 - 0.13x_6 + 0.04x_4 + 0.03x_3 - 0.01$	22
	5-P	$0.94 \sin(1.813x_1) \quad -0.08x_2 - 0.08x_6$	31
22		$1.0\,x_5 \lvert x_6 \rvert + 1.0\,x_x$	
	LIN	$0.65x_5 - 0.40x_6 + 0.09x_3 \quad -0.02$	40
	3-PL	$0.65x_5 - 0.37x_6 + 0.07x_3 - 0.09x_1 - 0.06x_2 - 0.02$	39
	5-P	$0.89x_5 \lvert x_6 \rvert - 0.18x_6 - 0.06x_2$	50
23		$1.0\,x_5 \lvert x_6 \rvert + 1.0 \sqrt{x_3^2 + x_4^2} + 1.0\,x_x$	
	LIN	$0.65x_5 - 0.40x_6 + 0.16x_3 + 0.09x_4 \quad +1.20$	29
	3-PL	$0.64x_5 - 0.36x_6 + 0.15x_3 + 0.07x_4 - 0.03x_1 + 1.20$	29
	5-P	$0.87x_5 \lvert x_6 \rvert + 0.77 \sqrt{x_3^2 + x_4^2} \quad -0.19x_6 - 0.06x_2$	58
24		$1.0\,x_6 \operatorname{arctg} 2x_5 + 1.0\,x_x$	
	LIN	$-0.25x_5 - 0.37x_6 - 0.09x_2 \quad -0.33$	7
	3-PL	$-0.19x_5 - 0.29x_6 - 0.18x_2 - 0.10x_1 + 0.02x_3 - 0.33$	7
	5-P	$0.88x_6 \operatorname{arctg} 2x_5 - 0.09x_6 - 0.09x_2$	45

4.5 Conclusions

In cases with exact linear solutions (nos. 1-4) all three methods succeed to 100 per cent. However, when the exact solutions are nonlinear, only the Five-Party Congression is successful. The linear methods either fail completely (nos. 5-8 and 12) or do find the right predictors but of course not the proper function.

When the analysis is complicated by the unknown lurking predictor, the three methods are still quite comparable in linear cases (nos. 13-16), but the results are sometimes rather bad (No. 14). In the remaining cases the Five-Party method is again successful in finding the right function, although the coefficients are now less close to the correct answer.

Looking finally on fictitious effects of intercorrelation among predictors and their correlation with the lurking predictor, it is interesting to note that the three methods often agree on these terms of no significance.

Returning now to our three questions, the answers can be formulated in this way:

1) Yes, for analysing linear relationships the methods seem quite comparable.

2) No, not even in most difficult cases does Congression fail compared with traditional regression.

3) Congression is quite successful in nonlinear cases. The very short time needed on the computer is remarkable.

5. DISCUSSION

5.1 Different procedures for Multiple Regression Analysis

The Congression technique is a quite new way to tackle the problem of multiple regression. It has not been possible to find in the literature any method with even the slightest resemblance to the Congression approach, neither in the comprehensive guide to Applied Regression by DRAPER and SMITH (1981), nor in the more general textbook by KENDALL and STUART (1979).

The main features of the Congression techniqus are the Grouping of data in Parties and the use of Grouping Diagrams as an interface towards a library of potential functions of any two predictors. Instead of dealing with the original N cases, one predictand and say, 10 predictors, the analysis is carried out with 25 cases only, one "predictor" and as many "predictands" as there are pairs of the original predictor. The 10 predictors mentioned above, mean 45 pairs.

Congression in its present form utilizes a stepwise forward selection procedure. As the first step in each round, we search for the "best" pair of predictors, whether already used or not. As the second step, we search for the "best" function, still whether already used or not, to describe how the predictors in the best pair are connected with the predictand (or the residual predictand); the terms predictor and predictand again used in their original sense. Obviously, the final congression coefficient is sometimes the result of an iterative process.

Congression opens up an entirely new field of thoughts. For that reason many ideas that have come to my mind have not yet been explored. All those interested in the problem are invited to take part in the further exploration. Among other things it would be interesting to investigate to what extent the various selection procedures used so far in multiple regression could be carried out using the Congression technique. It should be noted that techniques which are supplementary to regression, such as "cross-validation", can of course be applied together with Congression as well.

5.2 The Effect of increasing the Number of Parties

With the various variants of the Congression technique, the original number of cases, say N = 1000, is substantially reduced to only 2 (for 2-PL), 3 (for 3-PL), 9 (for 3-P) or 25 (for 5-P). The great saving is evidently gradually lost with the introduction of more Parties. The number of boxes in the Grouping Diagram and by that in the Function Diagrams increase at the same rate. Hence a further increase in the number of parties leads to 49 cases and boxes for a Seven-Party method, and 81 for a Nine-Party one. Although this would at the same time lead to better resolution and a more detailed description of the Potential Functions, it also means that the number of cases falling in each box is reduced, which might lead to more "zeros" in the boxes and by that a less successful analysis.

It is my opinion that Five Parties happens to be the optimum solution and that it would not be worth-while, except possibly for very special cases with complicated functions, to further increase the number of Parties.

The remarkable improvement in the resolution of Function Diagrams when going from 3 to 5 Parties is demonstrated in the following figure.

$x_i^2 - x_j^2$ and $\sqrt{x_i^2 + x_j^2}$

3-P

$x_i^2 - x_j^2$:

0	142	0
-142	0	-142
0	142	0

$\sqrt{x_i^2 + x_j^2}$:

101	18	101
18	-180	18
101	18	101

5-P

$x_i^2 - x_j^2$:

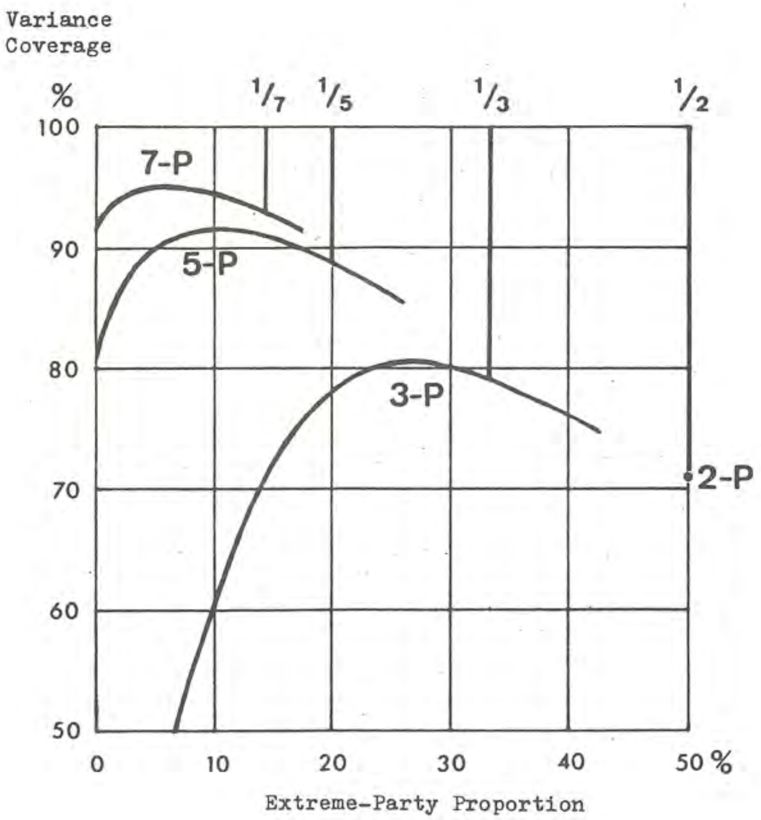
0	150	182	150	0
-150	0	32	0	-150
-182	-32	0	-32	-182
-150	0	32	0	-150
0	150	182	150	0

$\sqrt{x_i^2 + x_j^2}$:

136	48	25	48	136
48	-80	-126	-80	48
25	-126	-206	-126	25
48	-80	-126	-80	48
136	48	25	48	136

Comparison of the Function Diagrams for
$(x_i + x_j)(x_i - x_j)$ and $\sqrt{x_i^2 + x_j^2}$
in the case of 3-P and 5-P Congression.

Next figure shows how the optimum proportions of the Parties can be obtained, in particular the size of the most extreme Parties; curves are shown for 3, 5 and 7 Parties, as well as the singular point for the 2-Party case.



Let us summarize what can be read out of this diagram. For Two Parties there is no question; both parties are extreme and comprise 50 % each. The Variance Coverage is 71 %. For Three-Party Congression, the extreme parties should comprise each 27 %. This optimum value gives 81 % coverage. The optimum for Five-Party Congression is reached when each extreme party comprises 10 % the data, if normally distributed. This means that 91 % of the total variance is represented by the Party mean values. The corresponding figures for an optimum Seven-Party method are 6 and 95, respectively.

These findings are all illustrated in the following figure. The proportions actually used in the present study are shown as well.

2-P	50	50						71 %
3-P	27	46	27					81 %
5-P	10	25	30	25	10			91 %
7-P	6	14	18	24	18	14	6	95 %
5-P now used	12.5	25	25	25	12.5			91 %

The percentages to the right indicate how much of the total variance is represented by the Party mean values, provided data are normally distributed.

5.3 Selection of A Set of Potential Functions

The test experiments have shown that computer time can be saved by using more Potential Functions than are really necessary. Such extra functions could be termed "Convenient Functions". It is my opinion that it would pay to be generous in choosing functions for the "set" or "library" of pre-prepared potential functions. The present number, 84, is by no means the ultimate solution.

A somewhat flexible Program Package for Congression will certainly be made available to users. If the user then regards the number of functions as excessive, he should be encouraged to judge from his own point of view, whether some functions should be removed from the set already from start.

There will be devices to do so. It will also be possible to watch which functions come out in a first preliminary run, and then to decide which of them should perhaps be deleted before the final run. It should be remembered that there seldom exists something that could be called "the true solution" to a regression problem. There will always be an element of judgement. Or, as said by DRAPER and SMITH (1981):

> "The use of multiple regression techniques is a powerful tool only if it is applied with intelligence and caution."

5.4 <u>The Effect of Histogram Deviations from Normal</u>

Let us go back to Test Run No. 11, which was successful as far as Five-Party Congression is concerned. The "best" Empirical Function of the First and Second Round are shown below, as well as the Potential Functions which were picked up as "best" describing these Empirical Functions.

First Round

$E_{5x5}(x_5,x_6)$

-189	-99	-14	-83	-216
-86	-50	-2	-45	-126
-10	1	21	-18	18
101	36	8	67	105
258	126	-3	144	0

$0.835\ F_{5x5}(x_5 | x_6 |)$

-246	-104	-24	-104	-246
-104	-43	-9	-43	-104
0	0	0	0	0
104	43	9	43	104
246	104	24	104	246

Accumulated Variance Reduction: 63.3 %

Second Round

Residual $E_{5x5}(x_3,x_4)$

101	70	49	78	140
67	-20	-45	-23	64
64	-46	-99	-51	38
59	-24	-52	-18	84
98	55	48	72	155

$0.506\ F_{5x5}(\sqrt{x_3^2 + x_4^2})$

69	24	13	24	69
24	-40	-64	-40	24
13	-64	-104	-64	13
24	-40	-64	-40	24
69	24	13	24	69

Accumulated Variance Reduction: 98.0 %

It took another two rounds to reach the final variance reduction, 99.9 % .

Let us now look at the corresponding Two-Dimensional Histograms and their Deviation from Normal Distribution. (Since there were 1000 cases, the figures are permillages.)

	Histograms					Deviations from Normal Distribution				
First Round (x_5, x_6)	4	23	31	40	40	-12	-8	0	9	24
	23	68	63	78	54	-8	5	1	15	23
	31	63	50	49	23	0	1	-13	-13	-8
	40	78	49	40	18	9	15	-13	-23	-13
	40	54	23	18	0	24	23	-8	-13	-16
Second Round (x_3, x_4)	8	32	34	20	5	-8	1	3	-11	-11
	38	86	89	55	23	7	23	27	-8	-8
	27	74	66	65	21	-4	12	3	3	-10
	25	52	76	59	36	-6	-11	14	-4	5
	8	18	26	34	25	-8	-13	-5	3	9

These diagrams demonstrate that the technique works well even when data deviate considerably from normal distribution. Not even a "zero" in one of the boxes (marked by a cross) seems to do any harm.

However, it might happen that predictors are so strongly correlated that there will be many zero-boxes in the Histogram. Experience so far, though limited, has shown that this might complicate the analysis. It has happened that the program picks up a function for which the most heavy boxes in the grouping diagram coincide with those zero-boxes. The computer program must include some device to prohibit such a bad choice.

In lectures given on Automatic Interpretation of Forecast Charts I have discussed a similar problem, LÖNNQVIST (1978). I mentioned in that case, that I had used a technique where information was borrowed from close-by boxes in a systematic way and to an extent depending on the number of cases falling in the box itself. This would probably not be a good procedure in the case of regression. It would be safer to introduce a system where certain Potential Functions are automatically blocked, depending on where and to which extent zeros appear in individual Pair-Histograms. In the worst case, one of the predictors might have to be removed as being too closely correlated with another predictor.

5.5 Empirical Functions and Damped Congression

Some slight variations in the technique described so far have already been partly explored.

5.5.1 As already mentioned, the result of the first step in each round is in fact a presentation of an Empirical Function, which describes the predictand (or its residuals) in 25 discrete points as a function of two predictors. For some very special problems that picture of the predictand, properly analysed by isolines, might be more relevant than the corresponding functional relationship. Just to mention one case, such a problem might concern the geographical distribution of a climatological property by latitude and longitude. In such a case the empirical solution should be recorded as the first "term" in a regression-relation based entirely on Empirical Functions. The effect of this "term" should then be subtracted from all the original predictand values in order to get a residual predictand, and so on.

5.5.2 This subtraction in each individual case might cause complications since the function is given in discrete points. There might therefore be a need for some interpolation method, as mentioned by LÖNNQVIST (1978). Now the best way would be to express the Empirical Function, nevertheless, in functional terms. Thus, the search for a new best pair of predictors should await carrying out a number of rounds by which as much as possible of the five-by-five-picture is described by a combination of Potential Functions available in the program package.

5.5.3 It might be advisable, even in other cases, to modify the forward selection method so that for each "best" pair found, there are always <u>two</u> rounds used to try to describe the relationship with the aid of available functions. Some experiments with this technique look promising, at least for solving some problems. The matter must certainly be studied much more in depth.

5.5.4. The same holds for another alternative to the straight forward technique. It could be called the "cautious" method. A Damping Factor, say 0.5 , is introduced. Whenever then a Congression Coefficient has been obtained in the usual way, it is multiplied by the damping factor before being used for modifying the predictand (or its residuals).

The analysis consequently proceeds more slowly than usual, but the idea is to prevent, at least partly, possible detrimental effects of a predictor chosen on vague grounds. Experiments seem to indicate that this might be useful when dealing with very "messy" data. However, more experience is needed.

5.6 Further outlook

Many other questions will certainly come up in due time. For instance, there might be an advantage, at least for solving very specific problems, to proceed from the present study of Pairs to a study of Triples of available predictors. The computing time will then of course be drastically increased. On the other hand, that would be the only way to study systematically relationships such as

$$\sqrt{x_i^2 + x_j^2 + x_k^2}, \quad x_i \sqrt{x_j^2 + x_k^2}, \text{ and } \quad x_i \sin(x_j - x_k).$$

It might even happen that techniques similar to Congression could be used with success for solving other problems in statistics and maybe in other fields of mathematics.

6. STATEMENT

The idea behind the Congression technique come suddenly to my mind when dealing with a very tricky meteorological problem. Less than a month later the first experiment was carried out on 12 September 1984. It was quite successful. The potentialities of the new technique were proven. The surprisingly short computing time was noted. By and large, the test runs were the same as those reported in this paper.

The simple linear versions were not thought of until later.

A first lecture on the new technique was presented at SMHI on 11 October 1984 under the title: "Congression - a New Effective Tool in Statistics".

The manuscript was delivered on 29 November 1984.

R E F E R E N C E S

Conrad, V. and L.W. Pollak, 1950: Methods in Climatology. Harvard University Press.

Draper, N. and H. Smith, 1981: Applied Regression Analysis. Second Edition. John Wiley & Sons, Inc.

Kendall, Sir M. and S. Stuart, 1979: The Advanced Theory of Statistics, Vol. 2, Inference and Relationship. Forth Edition. Ch.Griffin & Company, Ltd.

Lönnqvist, O., 1978: Automatic Interpretation of Forecast Charts. ECMWF Seminars 1978.

SMHI
HO
kuy/L/SCR

SMHI Rapporter, HYDROLOGI OCH OCEANOGRAFI (RHO)

Nr 1 Weil, J G
Verification of heated water jet numerical model
Stockholm 1974

Nr 2 Svensson, J
Calculation of poison concentrations from a hypothetical accident off the Swedish coast
Stockholm 1974

Nr 3 Vasseur, B
Temperaturförhållanden i svenska kustvatten
Stockholm 1975

Nr 4 Svensson, J
Beräkning av effektiv vattentransport genom Sunninge sund
Stockholm 1975

Nr 5 Bergström, S och Jönsson, S
The application of the HBV runoff model to the Filefjell research basin
Norrköping 1976

Nr 6 Wilmot, W
A numerical model of the effects of reactor cooling water on fjord circulation
Norrköping 1976

Nr 7 Bergström, S
Development and application of a conceptual runoff model
Norrköping 1976

Nr 8 Svensson, J
Seminars at SMHI 1976-03-29--04-01 on numerical models of the spreading of cooling water
Norrköping 1976

Nr 9 Simons, J, Funkquist, L and Svensson, J
Application of a numerical model to Lake Vänern
Norrköping 1977

Nr 10 Svensson, S
A statistical study for automatic calibration of a conceptual runoff model
Norrköping 1977

Nr 11 Bork, I
Model studies of dispersion of pollutants in Lake Vänern
Norrköping 1977

Nr 12 Fremling, S
Sjöisars beroende av väder och vind, snö och vatten
Norrköping 1977

Nr 13 Fremling, S
Sjöisars bärighet vid trafik
Norrköping 1977

Nr 14 Bork, I
Preliminary model studies of sinking plumes
Norrköping 1978

Nr 15 Svensson, J and Wilmot, W
A numerical model of the circulation in Öresund
Evaluation of the effect of a tunnel between Helsingborg and Helsingör
Norrköping 1978

Nr 16 Funkquist, L
En inledande studie i Vätterns dynamik
Norrköping 1978

Nr 17 Vasseur, B
Modifying a jet model for cooling water outlets
Norrköping 1979

Nr 18 Udin, I och Mattsson, I
Havsis- och snöinformation ur datorbearbetade satellitdata - en metodstudie
Norrköping 1979

Nr 19 Ambjörn, C och Gidhagen, L
Vatten- och materialtransporter mellan Bottniska viken och Östersjön
Norrköping 1979

Nr 20 Gottschalk, L och Jutman, T
Statistical analysis of snow survey data
Norrköping 1979

Nr 21 Eriksson, B
Sveriges vattenbalans. Årsmedelvärde (1931-60) av nederbörd, avdunstning och avrinning
Norrköping 1980

Nr 22 Gottschalk, L and Krasovskaia, I
Synthesis, processing and display of comprehensive hydrologic information
Norrköping 1980

Nr 23 Svensson, J
Sinking cooling water plumes in a numerical model
Norrköping 1980

Nr 24 Vasseur, B, Funkquist, L and Paul, J F
Verification of a numerical model for thermal plumes
Norrköping 1980

Nr 25 Eggertsson, L-E
HYPOS - ett system för hydrologisk positionsangivelse
Norrköping 1980

Nr 26 Buch, Erik
Turbulent mixing and particle distribution investigations in the Himmerfjärd 1978
Norrköping 1980

Nr 27 Eriksson, B
Den "potentiella" evapotranspirationen i Sverige
Norrköping 1980

Nr 28 Broman, B
On the spatial representativity of our oceanographic measurements
Norrköping 1981

Nr 29 Ambjörn, C, Luide, T, Omstedt, A, Svensson, J
En operationell oljedriftsmodell för norra Östersjön
Norrköping 1981

Nr 30 Svensson, J
Vågdata från svenska kustvatten 1979 - 1980
Norrköping 1981

Nr 31 Jutman, T
Stationsnät för vattenföring
Norrköping 1981

Nr 32 Omstedt, A, Sahlberg, J
Vertical mixing and restratification in the Bay of Bothnia during cooling
Norrköping 1982

Nr 33 Brandt, M
Sedimenttransport i svenska vattendrag
Norrköping 1982

Nr 34 Bringfelt, B
A forest evapotranspiration model using synoptic data
Norrköping 1982

Nr 35 Bhatia, P K, Bergström, S, Persson, M
Application of the distributed HBV-6 model to the Upper Narmada Basin in India
Norrköping 1984

Nr 36 Omstedt, A
A forecasting model for water cooling in the Gulf of Bothnia and Lake Vänern
Norrköping 1984

Nr 37 Gidhagen, L
Coastal Upwelling in the Baltic - a presentation of satellite and in situ measurements of sea surface temperatures indicating coastal upwelling
Norrköping 1984

Nr 38 Engqvist, A, Svensson, J
Water turnover in Himmerfjärd 1977 - a simulation study
Norrköping 1984

Nr 39 Funkquist, L, Gidhagen, L.
A model for pollution studies in the Baltic Sea
Norrköping 1984

SMHI
HO
kuy/L/SCR

SMHI Rapporter, METEOROLOGI OCH KLIMATOLOGI (RMK)

Nr 1 Thompson, T, Udin, I, and Omstedt, A
Sea surface temperatures in waters surrounding Sweden
Stockholm 1974

Nr 2 Bodin, S
Development on an unsteady atmospheric boundary layer model.
Stockholm 1974

Nr 3 Moen, L
A multi-level quasi-geostrophic model for short range weather predictions
Norrköping 1975

Nr 4 Holmström, I
Optimization of atmospheric models
Norrköping 1976

Nr 5 Collins, W G
A parameterization model for calculation of vertical fluxes of momentum due to terrain induced gravity waves
Norrköping 1976

Nr 6 Nyberg, A
On transport of sulphur over the North Atlantic
Norrköping 1976

Nr 7 Lundqvist, J-E, and Udin, I
Ice accretion on ships with special emphasis on Baltic conditions
Norrköping 1977

Nr 8 Eriksson, B
Den dagliga och årliga variationen av temperatur, fuktighet och vindhastighet vid några orter i Sverige
Norrköping 1977

Nr 9 Holmström, I, and Stokes, J
Statistical forecasting of sea level changes in the Baltic
Norrköping 1978

Nr 10 Omstedt, A, and Sahlberg, J
Some results from a joint Swedish-Finnish sea ice experiment, March, 1977
Norrköping 1978

Nr 11 Haag, T
Byggnadsindustrins väderberoende, seminarieuppsats i företagsekonomi, B-nivå
Norrköping 1978

Nr 12 Eriksson, B
Vegetationsperioden i Sverige beräknad från temperaturobservationer
Norrköping 1978

Nr 13 Bodin, S
En numerisk prognosmodell för det atmosfäriska gränsskiktet grundad på den turbulenta energiekvationen
Norrköping 1979

Nr 14 Eriksson, B
Temperaturfluktuationer under senaste 100 åren
Norrköping 1979

Nr 15 Udin, I, och Mattisson, I
Havsis- och snöinformation ur datorbearbetade satellitdata - en modellstudie
Norrköping 1979

Nr 16 Eriksson, B
Statistisk analys av nederbördsdata. Del I. Arealnederbörd
Norrköping 1979

Nr 17 Eriksson, B
Statistisk analys av nederbördsdata. Del II. Frekvensanalys av månadsnederbörd
Norrköping 1980

Nr 18 Eriksson, B
Årsmedelvärden (1931-60) av nederbörd, avdunstning och avrinning
Norrköping 1980

Nr 19 Omstedt, A
A sensitivity analysis of steady, free floating ice
Norrköping 1980

Nr 20 Persson, C och Omstedt, G
En modell för beräkning av luftföroreningars spridning och deposition på mesoskala
Norrköping 1980

Nr 21 Jansson, D
Studier av temperaturinversioner och vertikal vindskjuvning vid Sundsvall-Härnösands flgplats
Norrköping 1980

Nr 22 Sahlberg, J and Törnevik, H
A study of large scale cooling in the Bay of Bothnia
Norrköping 1980

Nr 23 Ericson, K and Hårsmar, P-O
Boundary layer measurements at Klockrike. Oct. 1977
Norrköping 1980

Nr 24 Bringfelt, B
A comparison of forest evapotranspiration determined by some independent methods
Norrköping 1980

Nr 25 Bodin, S and Fredriksson, U
Uncertainty in wind forecasting for wind power networks
Norrköping 1980

Nr 26 Eriksson, B
Graddagsstatistik för Sverige
Norrköping 1980

Nr 27 Eriksson, B
Statistisk analys av nederbördsdata. Del III. 200-åriga nederbördsserier
Norrköping 1981

Nr 28 Eriksson, B
Den "potentiella" evapotranspirationen i Sverige
Norrköping 1981

Nr 29 Pershagen, H
Maximisnödjup i Sverige (perioden 1905-70)
Norrköping 1981

Nr 30 Lönnqvist, O
Nederbördsstatistik med praktiska tillämpningar (Precipitation statistics with practical applications)
Norrköping 1981

Nr 31 Melgarejo, J W
Similarity theory and resistance laws for the atmospheric boundary layer
Norrköping 1981

Nr 32 Liljas, E
Analys av moln och nederbörd genom automatisk klassning av AVHRR data
Norrköping 1981

Nr 33 Ericson, K
Atmospheric Boundary layer Field Experiment in Sweden 1980, GOTEX II, part I
Norrköping 1982

Nr 34 Schoeffler, P
Dissipation, dispersion and stability of numerical schemes for advection and diffusion
Norrköping 1982

Nr 35 Undén, P
The Swedish Limited Area Model (LAM). Part A. Formulation
Norrköping 1982

Nr 36 Bringfelt, B
A forest evapotranspiration model using synoptic data
Norrköping 1982

Nr 37 Omstedt, G
Spridning av luftförorening från skorsten i konvektiva gränsskikt
Norrköping 1982

Nr 38 Törnevik, H
An aerobiological model for operational forecasts of pllen concentration in th air
Norrköping 1982

Nr 39 Eriksson, B
Data rörande Sveriges temperaturklimat
Norrköping 1982

Nr 40 Omstedt, G
An operational air pollution model using routine meteorological data
Norrköping 1984

Nr 41 Persson, C and Funkqvist, L
Local scale plume model for nitrogen oxides - Model description
Norrköping 1984

Nr 42 Stefan Gollvik
Estimation of orographic precipitation by dynamical interpretation of synoptic model data.

ISSN 0347-2116

SMHI

SWEDISH METEOROLOGICAL AND HYDROLOGICAL INSTITUTE
S-60176 Norrköping, Sweden. Tel. +46 11 15 80 00. Telex 644 00 smhi s.