Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 27 (2003) 701–714
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

The SROV program for data analysis and regression model identification

Mordechai Shacham [a,*], Neima Brauner [b]

[a] Department of Chemical Engineering, Ben Gurion University of the Negev, Beer-Sheva 84105, Israel
[b] School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel

Received 31 October 2001; received in revised form 21 October 2002; accepted 11 November 2002

Abstract

A new stepwise regression program (SROV) for the construction of optimal (stable and of highest possible accuracy) regression models comprised of linear combination of independent variables and their non-linear functions is described. The program uses for regression QR decomposition based on Gram–Schmidth orthogonalization, which is highly resilient to numerical error propagation. Variables are selected to enter the regression model according to their level of correlation with the dependent variable and they are removed from further consideration when their residual information gets below the noise level. The use of this program is demonstrated in two examples. In both examples the program identifies an optimal and stable regression model and several sub-optimal models. The existence of sub-optimal models provides additional insight regarding the relationships that exist between the explanatory variables, between the explanatory variables and the dependent variable and information on model related uncertainties caused by sample size and experimental error.

Keywords: Stepwise regression; Colinearity; Non-influential variable; Noise; Precision

1. Introduction

Precise analysis, modeling and regression of experimental data are key requirements for realistic and accurate modeling and simulation of physical phenomena. As model based simulation, design, control and optimization of chemical processes become increasingly more widespread, the requirements for more precise regression models for representing input data (e.g. physical and transport properties, phase equilibrium) become increasingly more severe.

Regression models can be partially theory based or completely empirical. In both cases, it is not known a-priori how many explanatory variables (independent variables, and/or their functions) and parameters should be included in the model for obtaining an optimal regression model. An insufficient number of explanatory variables result in an inaccurate model, which is characterized by a large variance. Some independent variables, which may have critical effects on the dependent variable under certain circumstances, may be left out of the correlation. On the other hand, including non-influential variables and/or variables which are collinear, renders an unstable model. The instability is characterized by typical ill effects, whereby adding or removing an experimental point from the data set may drastically change the parameter values. Also, the derivatives of the dependent variable are not represented correctly and extrapolation outside the region, where the measurements were taken, yields absurd results even for a small range of extrapolation. Shacham and Brauner (1997), Brauner and Shacham (1998a,b) provide several examples where regression models published in the chemical engineering literature are grossly inaccurate and/or unstable.

Over the years, many procedures have been introduced for selection of the optimal model in linear regression (for detailed reviews, see for example Daniel & Wood, 1980; Neter, Wasserman & Kutner, 1990). Wagner (1973) and Setzmann and Wagner (1989) have

* Corresponding author. Tel.: +972-8-646-1481; fax: +972-8-647-2916.
E-mail address: shacham@bgumail.bgu.ac.il (M. Shacham).

0098-1354/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0098-1354(02)00260-0

applied such procedures extensively to models of physicochemical and thermodynamic properties. In those procedures, various statistical tests, such as the F-test, t-test, R_p^2-criterion, C_p-criterion, PRESS$_p$-criterion, residual plot and other diagnostic plots are used to compare between different models and to decide which of the explanatory variables should be included or removed from the model in order to arrive at the 'best' subset of explanatory variables. (For details of the various statistical tests used, see for example, pp. 433–470 in Neter et al., 1990). Those tests may be applied while considering all possible combinations of the explanatory variables or in the framework of a stepwise regression procedure (such as forward stepwise regression, backward elimination, ridge regression, see for example Marquardt & Snee, 1975).

Stepwise regression programs that use statistical tests do not take advantage of the information concerning the signal to noise ratio in the data. This information is usually available when working with experimental data, but it is less common in observational data that statisticians usually work with. Shacham and Brauner (1999a,b) have described the principles and the algorithm of a new stepwise regression procedure with orthogonal variables (SROV), which uses indicators based on 'signal-to-noise' ratio. They have also demonstrated some of the potential applications and the advantages of the use of this procedure (Shacham & Brauner, 1999a,b; Brauner & Shacham, 1999). Experience in using the SROV algorithm for solving a wide range of problems has identified some necessary modifications of the algorithm.

In this paper the modified and extended SROV algorithm is presented. The algorithm has also been implemented as a MATLAB[1] program and it has become available on the Internet for downloading. The use of this new program is demonstrated in the paper.

In the next section two motivating examples that demonstrate the need for using stepwise regression, are presented. In Section 3, the revised and extended SROV algorithm is described while Section 4 includes details of the SROV program implementation and use. In Section 5, the motivating examples are solved using SROV. Finally some conclusions regarding the benefits in using SROV are presented.

The computations reported in the paper were carried out with MATLAB 5.3 and POLYMATH[2] 5.1

[1] MATLAB is a trademark of The Math Works, Inc. (http://www.mathworks.com).

[2] POLYMATH is copyrighted by M. Shacham, M.B. Cutlip and M. Elly (http://www.polymath-software.com).

2. Motivating examples

2.1. Example 1. Calibration of a near infrared reflectance instrument (Fearn, 1983)

This example concerns data analysis from a series of experiments performed to calibrate a near infrared reflectance (NIR) instrument for the measurement of protein content in ground wheat samples. Fearn (1983) reports the results of the experiments in two separate sets of data: a 'Calibration set' (shown in Table 1) and 'Prediction set' (shown in Table 2). The six independent variables $L1$–L6 are measurements of the reflectance of the NIR radiation by the wheat samples at six different wavelengths. These measurements are taken on a log(1/R) scale, where R is the reflectance, and are commonly referred to as 'log values'. The protein content (dependent variable) was measured by the standard Kjeldahl method. The objective of the calibration is to find a linear combination of the log values, which predicts the protein content. For this purpose, the linear equation:

$$y = \beta_0 + \beta_1 L1 + \beta_2 L2 + \beta_3 L3 + \beta_4 L4 + \beta_5 L5 + \beta_6 L6 \quad (1)$$

is fitted to the data, where y is the protein content and $\beta_0, \beta_1, \ldots, \beta_6$ are the model parameters. The calculated parameter values, the 95% confidence intervals on the parameter values, the variance and the linear correlation coefficients (R^2) for the two data sets are shown in Table 3. Note that in this table, as in the other tables reporting regression results, the number of significant digits seems excessive in light of the precision of the data. However, reporting the results with many significant digits is necessary for keeping the internal consistency of the results (parameter, variance and confidence interval values) and for comparing the accuracy achieved by different regression programs.

Let us examine first the results for the 'Calibration set' of data. The residual plot for this set is shown in Fig. 1. The first impression is that the model represents the data adequately. There is a random distribution of the residuals with a maximal error below 5% and $R^2 = 0.982$. But comparing the parameter values with their 95% confidence intervals shows that some of the confidence intervals are much larger than the respective parameter values. Such a situation usually arises if non-influential independent variables are included in the model, or/and there is colinearity between some of the variables. Colinearity between the variables, or the inclusion of non-influential variables, renders the model unstable. The ill-effects of the instability of this particular model can be clearly seen by comparing the parameter values obtained by regressing the 'Calibration set' with those obtained by regressing the 'Prediction set'. The sign of three of the coefficients ($\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_5$)

Table 1
Results of calibration experiment with 24 samples for Example 1 (Fearn, 1983)

Sample number	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$	Protein (%)
1	468	123	246	374	386	−11	9.23
2	458	112	236	368	383	−15	8.01
3	457	118	240	359	353	−16	10.95
4	450	115	236	352	340	−15	11.67
5	464	119	243	366	371	−16	10.41
6	499	147	273	404	433	5	9.51
7	463	119	242	370	377	−12	8.67
8	462	115	238	370	353	−13	7.75
9	488	134	258	393	377	−5	8.05
10	483	141	264	384	398	−2	11.39
11	463	120	243	367	378	−13	9.95
12	456	111	233	365	365	−15	8.25
13	512	161	288	415	443	12	10.57
14	518	167	293	421	450	19	10.23
15	552	197	324	448	467	32	11.87
16	497	146	271	407	451	11	8.09
17	592	229	360	484	524	51	12.55
18	501	150	274	406	407	11	8.38
19	483	137	260	385	374	−3	9.64
20	491	147	269	389	391	1	11.35
21	463	121	242	366	353	−13	9.7
22	507	159	285	410	445	13	10.75
23	474	132	255	376	383	−7	10.75
24	496	152	276	396	404	6	11.47

Table 2
Prediction set with 26 samples for Example 1 (Fearn, 1983)

Sample number	$L1$	$L2$	$L3$	$L4$	$L5$	$L6$	Protein (%)
1	486	144	266	393	373	26	8.66
2	485	136	260	393	395	6	7.9
3	482	136	260	388	423	−2	9.27
4	443	112	232	346	355	−18	11.77
5	478	134	257	382	390	−5	9.7
6	449	113	233	351	343	−18	10.46
7	461	121	243	366	378	−14	10.17
8	503	155	280	403	414	6	11.1
9	493	146	271	390	378	−3	12.03
10	368	40	158	275	250	−63	9.43
11	462	114	237	367	331	−19	8.66
12	438	109	229	333	326	−28	14.44
13	478	127	252	384	378	−11	8.5
14	405	73	193	311	305	−44	10.41
15	498	146	273	403	415	0	9.72
16	442	106	226	341	303	−28	11.69
17	457	118	240	354	327	−23	12.19
18	439	103	224	339	325	−29	11.59
19	500	146	272	404	398	5	8.76
20	427	85	207	334	319	−36	8.6
21	479	128	253	384	382	−10	8.54
22	444	102	224	350	333	−27	9.34
23	458	118	239	362	355	−16	10.09
24	518	162	290	426	464	16	8.72
25	465	124	247	369	386	−13	10.87
26	457	120	242	363	411	−15	10.89

Table 3
Regression results for Example 1 for a linear model containing all six variables

Parameter	Calibration set		Prediction set	
	Value	95% confidence interval	Value	95% confidence interval
β_0	23.07423	20.88694	29.37223	13.18371
β_1	0.028124	0.17327	−0.16928	0.129687
β_2	0.001667	0.183913	−0.15365	0.105771
β_3	0.234909	0.163315	0.533368	0.153156
β_4	−0.24044	0.13339	−0.13627	0.081847
β_5	0.011839	0.012927	−0.00825	0.009345
β_6	−0.03558	0.096068	−0.06154	0.027228
Variance	0.048549		0.027176	
R^2	0.982149		0.991224	

has changed, thus there is uncertainty even in the direction of change in the dependent variable (the protein content) as a result of changes in the values of $L1$, $L2$ or $L5$. To obtain a stable model some of the variables must be removed. In Section 5, the SROV program will be used for the selection of variables to be included in a stable, accurate model.

2.2. Example 2. The cracking of n-heptanes to acetylene

Kunugi, Tamura and Naito (1961) investigated the thermal cracking of hydrocarbons (*n*-heptanes and methane) to acetylene. Himmelblau (1970) presented 16 data points from the results obtained by Kunugi et al. (1961). The data reported by Himmelblau are shown in Table 4. The independent variables are the reactor temperature (x_1), the mole ratio of hydrogen to *n*-heptanes (x_2) and the contact time (x_3). The dependent variable, y, is the conversion of *n*-heptane to acetylene. All the variables were normalized (divided by the largest

Table 4
Data for Example 2

Sample number	x_1^*	x_2	x_3	y
1	1300	7.5	0.012	49
2	1300	9	0.012	50.2
3	1300	11	0.0115	50.5
4	1300	13.5	0.013	48.5
5	1300	17	0.0135	47.5
6	1300	23	0.012	44.5
7	1200	5.3	0.04	28
8	1200	7.5	0.038	31.5
9	1200	11	0.032	34.5
10	1200	13.5	0.026	35
11	1200	17	0.034	38
12	1200	23	0.041	38.5
13	1100	5.3	0.084	15
14	1100	7.5	0.098	17
15	1100	11	0.092	20.5
16	1100	17	0.086	29.5

*, Definition of the variables: x_1, reactor temperature (°C); x_2, mole ratio of hydrogen to *n*-heptane; x_3, contact time (s); y, conversion of *n*-heptane to acetylene (%).

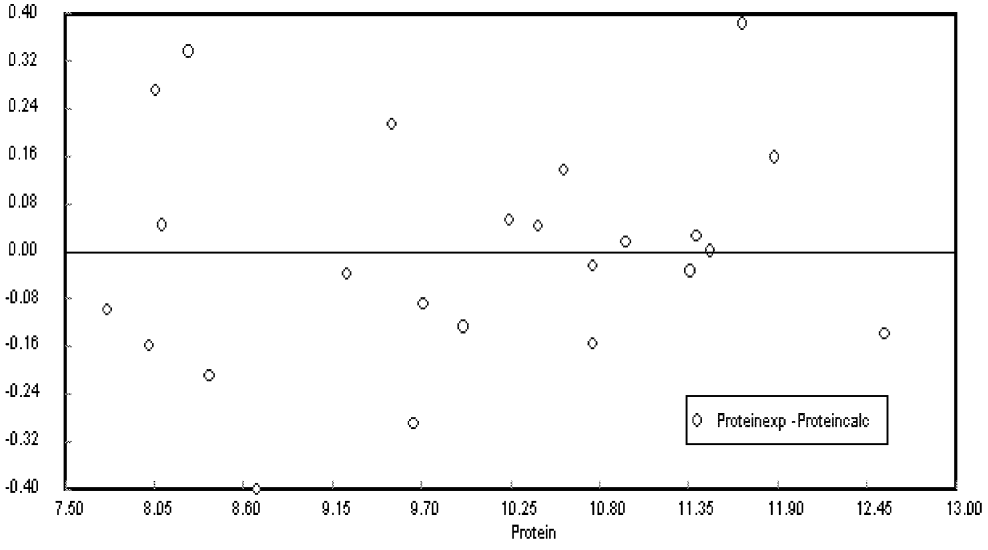

Fig. 1. Residual plot for Example 1 (calibration data set, linear model containing all six variables).

absolute value of the variable) before carrying out the regression.

Himmelblau (1970) suggested fitting a linear model to the data in Table 4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (2)$$

Regression yields the following parameter values (including 95% confidence intervals): $\hat{\beta}_0 = -2.4013783 \pm 2.3919682$, $\hat{\beta}_1 = 3.2655444 \pm 2.3661164$, $\hat{\beta}_2 = 0.1585668 \pm 0.1756745$ and $\hat{\beta}_3 = -0.0369135 \pm 0.4563804$. The variance is $s^2 = 0.0055645$ and $R^2 = 0.919815$. The residual plot for this model is shown in Fig. 2. Comparing the parameter values to their confidence intervals shows that the model is unstable. For two of the parameters, the confidence intervals are larger than the parameter values and in the case of $\hat{\beta}_3$, there is more than an order of magnitude difference. The residuals in Fig. 2 show a clear trend, implying an inadequate model.

Marquardt and Snee (1975) proposed using a full quadratic model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2 \quad (3)$$

for representation of these data. The calculated parameter values, the 95% confidence intervals on the parameter values, the variance and the linear correlation coefficients (R^2) for the quadratic model are shown in Table 5 and the residual plot in Fig. 3.

It can be seen that the quadratic model is much more appropriate for representing the data than the linear model. The use of the quadratic model yields a random residual distribution, higher value for R^2 ($= 0.9977$) and reduces the residual variance by more than an order of magnitude. But the full quadratic model is unstable: seven (out of the ten) parameters are smaller in absolute value than the respective confidence intervals (see Table 5). Thus, also in this case, there is a need to select the terms (the variables and their functions) to be included in a stable and appropriate model.

Table 5
Regression results for Example 2 for the full quadratic model with ten parameters

Parameter	Value	95% confidence interval
β_0	−71.6282	151.9582
β_1	137.0618	307.3667
β_2	8.764576	4.795331
β_3	26.71481	49.61517
β_4	−8.37465	4.65295
β_5	−26.6841	50.87425
β_6	−0.93886	1.009246
β_7	−64.4778	155.2602
β_8	−0.31784	0.299519
β_9	−2.20258	3.582665
Variance	0.0003186	
R^2	0.9977042	

3. The SROV (stepwise regression using ortogonalized variables) algorithm

The SROV algorithm has been described in Shacham and Brauner (1999a,b). Incorporation of the algorithm into a general-purpose program for solving a wide range of problems, required some modifications and additions to the algorithm. In the following, the updated algorithm is described.

A standard linear regression model can be written:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 \ldots + \beta_n \mathbf{x}_n + \boldsymbol{\varepsilon} \quad (4)$$

where $\mathbf{y}$ is an N-vector of the dependent variable, $\mathbf{x}_j$ ($j = 1, 2, \ldots n$) are N vectors of explanatory variables, β_0,

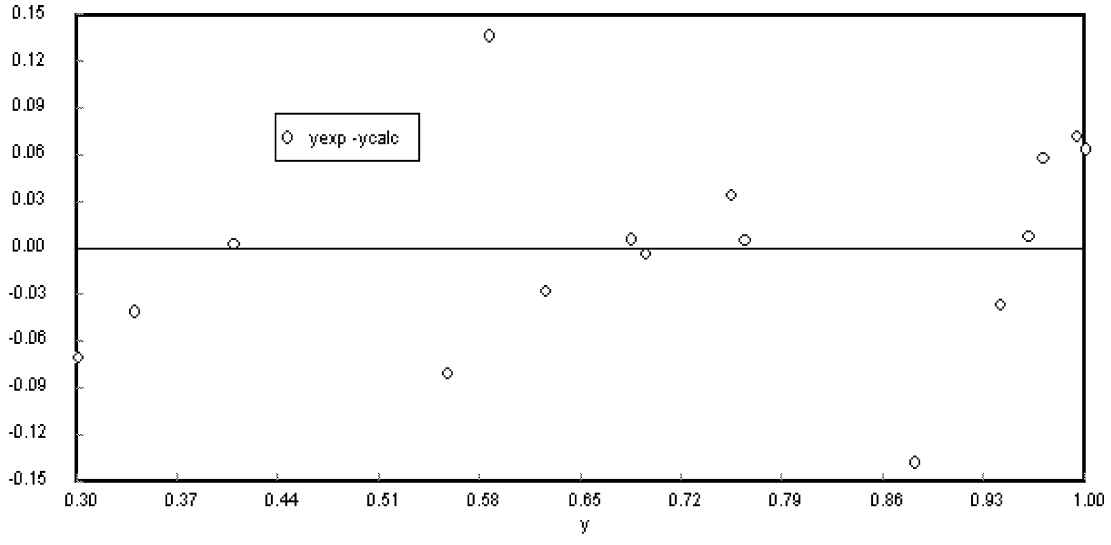

Fig. 2. Residual plot for Example 2 (four parameters linear model).

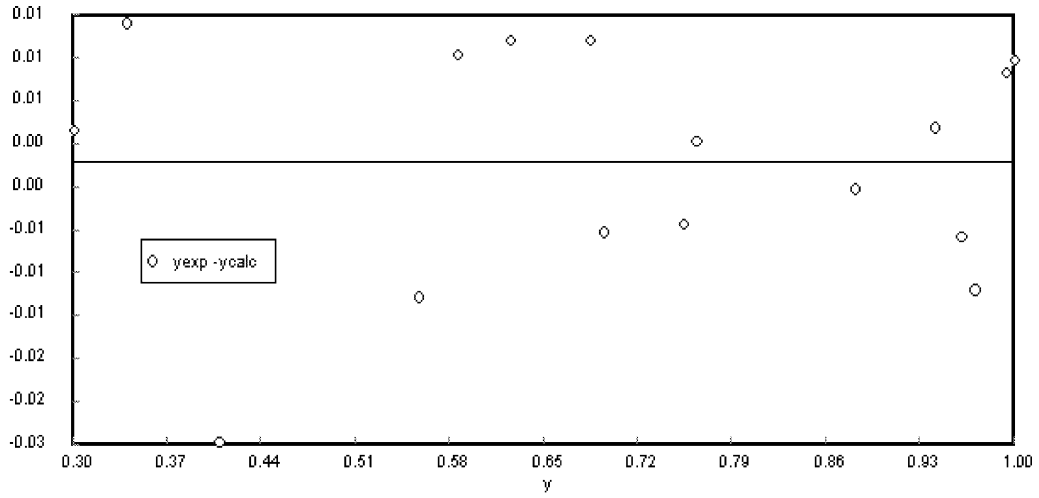

Fig. 3. Residual plot for Example 2 (quadratic model of ten parameters).

$\beta_1, \ldots, \beta_n$ are the model parameters to be estimated and $\boldsymbol{\varepsilon}$ is an N vector of stochastic terms (measurement errors). It should be noted that an explanatory variable could represent an independent variable or a function of one or more independent variables.

A certain error (disturbance, imprecision, noise) in the explanatory variables is also considered. Thus, a vector of an explanatory variable can be represented by:

$$\mathbf{x}_j = \hat{\mathbf{x}}_j + \delta\mathbf{x}_j \tag{5}$$

where $\hat{\mathbf{x}}_j$ is an N-vector of expected value of $\mathbf{x}_j$ and $\delta\mathbf{x}_j$ is an N-vector of stochastic terms due to noise.

The vector of estimated parameters $\hat{\boldsymbol{\beta}}^T = \hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_n$ is often calculated via the least-squares error approach by solving the normal equation:

$$\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \tag{6}$$

where $\mathbf{X} = [1, \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n]$ is an $N(n+1)$ data matrix and $\mathbf{X}^T\mathbf{X} = \mathbf{A}$ is the normal matrix. This method is subjected to accelerate numerical error propagation in cases of colinearity (see for example, Brauner & Shacham, 1998b). An alternative method for calculating the vector of the estimated parameters is the QR decomposition. It requires more arithmetic operations than the solution of the normal equations, but is less sensitive to numerical error propagation and as such, is more adequate for a general-purpose stepwise regression program. The QR decomposition solves the equation:

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} \tag{7}$$

by decomposing $\mathbf{X}$ into the product of a matrix $\mathbf{Q}$ with orthogonal columns and an upper triangular matrix $\mathbf{R}$. The SROV algorithm orthogonalizes the $\mathbf{Q}$ matrix by the Gram–Schmidt method (see for example Dahlquist, Bjork & Anderson, 1974), but the order in which the columns are orthogonalized and the numbers of columns that can be included in the regression model are determined by the unique algorithm that is described below.

The generation of the $\mathbf{Q}$ and $\mathbf{R}$ matrices is carried out simultaneously with the selection of the variables that should be included in the regression model. This is done in sequential steps, where at each step one of the explanatory variables, say x_p, is selected to enter the regression model. The explanatory variables, which have already been included in the regression model (at previous stages) are referred to as *basic variables*, and the remaining explanatory variables are the *non-basic variables*. At each step, the non-basic variables and the dependent variable are first updated, by subtracting the information which is collinear with the basic variables. This updating generates non-basic variables, which are orthogonal to the basic variables set. The description of the operations carried out in a single step of the algorithm follows.

3.1. Selection of x_p and update of the $\mathbf{Q}$ and $\mathbf{R}$ matrices and the $\check{\boldsymbol{\beta}}$ vector

The strength of the linear correlation between an explanatory variable $\boldsymbol{x}_j$ and a dependent variable $\mathbf{y}$ is measured by

$$YX_j = \mathbf{y}^T\mathbf{x}_j \tag{8}$$

where $\mathbf{y}$ and $\boldsymbol{x}_j$ are centered and normalized to a unit length. The value of $|YX_j|$ is in the range [0,1]. In a case of a perfect correlation between $\mathbf{y}$ and $\boldsymbol{x}_j$ ($\mathbf{y}$ is aligned in the $\boldsymbol{x}_j$ direction), $|YX_j| = 1$. In case $\mathbf{y}$ is unaffected by $\boldsymbol{x}_j$ (the two vectors are orthogonal), $YX_j = 0$. The inclusion of a variable $\boldsymbol{x}_p$, which has the highest level of correlation with $\mathbf{y}$ in the basic set ($|YX_j|$ value is the closest to one) will affect the maximal reduction of the variance of the regression model. Therefore, the criterion $\boldsymbol{x}_p = \boldsymbol{x}_j$ {max ($|YX_j|$} is used to determine which of

the non-basic variables should preferably be included in the regression model in the next step.

After the selection of x_p at step $k(\mathbf{x}_p^k)$, the **Q** and **R** matrices are updated using the following equations:

$$r_j^k = \frac{(\mathbf{x}_j^k)^T \mathbf{x}_p^k}{(\mathbf{x}_p^k)^T \mathbf{x}_p^k} \tag{9}$$

and

$$\mathbf{q}_j^{k+1} \equiv \mathbf{x}_j^{k+1} = \mathbf{x}_j^k - \mathbf{x}_p^k r_j^k \tag{10}$$

This update is carried out for all the columns associated with non-basic variables. At the same time the parameter value associated with $\mathbf{x}_p^k$ is calculated and the **y** vector is updated:

$$\check{\beta}_k = \frac{(\mathbf{y}^k)^T \mathbf{x}_p^k}{(\mathbf{x}_p^k)^T \mathbf{x}_p^k} \tag{11}$$

and

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \check{\beta}_k \mathbf{x}_p^k \tag{12}$$

Note that the updated non-basic explanatory variables include the residual subspace of explanatory variables that cannot be represented by the basic variables. Similarly, the updated dependent variable includes the residuals that cannot be explained by the basic variables. Thus, the current model variance is:

$$s^2 = \frac{(\mathbf{y}^{k+1})^T \mathbf{y}^{k+1}}{v} \tag{13}$$

where v is the number of degrees of freedom ($v = N - (k+1)$). The confidence interval, $\Delta\check{\beta}_p$ on a parameter estimate can be defined:

$$\Delta\check{\beta}_p = t(v, \alpha)\sqrt{s^2}(\mathbf{x}_p^k)^T \mathbf{x}_p^k \tag{14}$$

where $t(v, \alpha)$ is the statistical t distribution corresponding to v degrees of freedom and a desired confidence level, α.

3.2. Criteria for removing non-basic variables from consideration

An explanatory variable is removed from consideration for inclusion in the regression model when its residual information is at the noise level. For this purpose, two indicators are consulted. The first is the CNR_j, which measures the signal-to-noise ratio of YX_j and is defined by:

$$\mathrm{CNR}_j^k = \left\{ \frac{|(\mathbf{y}^k)^T \mathbf{x}_j^k|}{\sum_{i=1}^{N} (|x_{ij}^k \varepsilon_i^k| + |y_i^k \delta x_{ij}^k|)} \right\} \tag{15}$$

A value of $\mathrm{CNR}_j^k \gg 1$ signals that the correlation between $\mathbf{x}_j^k$ and $\mathbf{y}^k$ is significantly larger than the noise level. Thus, an accurate value of YX_j^k can be calculated. But when $\mathrm{CNR}_j^k \leq 1$, the noise in YX_j^k, as affected by $\delta\mathbf{x}_j^k$ and $\boldsymbol{\varepsilon}^k$ is as large as, or even larger than $|YX_j^k|$. If this is the case, no reliable value for YX_j^k can be obtained and the respective variable should not be included in the regression model.

The second indicator is the TNR_j^k, which measures the signal-to-noise ratio in an explanatory variable $\mathbf{x}_j^k$. It is defined in terms of the corresponding Euclidean norms (Shacham & Brauner, 1999a)

$$\mathrm{TNR}_j^k = \frac{||\mathbf{x}_j^k||}{||\delta\mathbf{x}_j^k||} = \left\{ \frac{(\mathbf{x}_j^k)^T \mathbf{x}_j^k}{(\delta\mathbf{x}_j^k)^T \delta\mathbf{x}_j^k} \right\}^{1/2} \tag{16}$$

A value of $\mathrm{TNR}_j^k \gg 1$ indicates that the (non-basic) explanatory variable x_j^k, contains valuable information. On the other hand, a value of $\mathrm{TNR}_j^k \leq 1$ implies that the information included in x_j^k is mostly noise, and therefore, it should not be added to the basic variables.

It should be noted that the denominators of Eqs. (15) and (16) represent the error in YX_j and $\mathbf{x}_j^k$, respectively, as propagated through the orthogonalization process. The propagated error is estimated by carrying out the orthogonalization process simultaneously with two data sets. The first is the original data set, while to the second is a perturbed data set obtained after introducing a normally distributed error with mean of $\delta\mathbf{x}_j$ (or $\boldsymbol{\varepsilon}$ in the case of the dependent variable). Subtracting the corresponding $\mathbf{x}_j^k$ (or $\mathbf{y}^k$) vectors obtained using the two data sets provides the required error estimates.

The selection of new variables (from among the non-basic variables) to be added to the basic variables in the SROV algorithm stops when for all the non-basic variables either $\mathrm{CNR}_j \leq 1$ or $\mathrm{TNR}_j \leq 1$.

3.3. Model parameters and confidence intervals in terms of the original variables

The results of variable selection for inclusion in the regression model at each stage are stored in the matrices **Q** and **R** and in the vector $\check{\boldsymbol{\beta}}$. Matrix **Q** is orthogonal; its columns are associated with the subset of the explanatory variables that are included in the regression model. Matrix **R** is an upper triangular matrix with 1 (one) on the diagonal associated with the basic variables. The vector $\check{\boldsymbol{\beta}}$ contains the regression coefficients associated with the orthogonalized variables. For practical use, parameter values associated with the original (non-orthogonalized) variables (the vector $\check{\boldsymbol{\beta}}$) are preferred. These can be simply calculated from the following equation:

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\check{\boldsymbol{\beta}} \tag{17}$$

To calculate the confidence intervals for the original parameters, the diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}$ are needed. These can be obtained using the $(\mathbf{Q}^T\mathbf{Q})^{-1}$ and

$(\mathbf{R})^{-1}$ matrices:

$$(\mathbf{X}^T\mathbf{X}) = \mathbf{R}^{-1}(\mathbf{Q}^T\mathbf{Q})^{-1}(\mathbf{R}^T)^{-1} \quad (18)$$

Since $\mathbf{Q}^T\mathbf{Q}$ is a diagonal matrix and $\mathbf{R}$ is an upper triangular matrix, their inverses can be very easily and accurately calculated. Therefore, this algorithm yields results of high accuracy even for problems that are considered highly co-linear and ill conditioned.

3.4. The two phases of the SROV algorithm

The SROV algorithm consists of two phases. In the first phase, an initial (nearly optimal) solution is found. In the second phase, the basic variables are rotated in an attempt to improve the model.

In the first phase, the steps of basic variable selection, followed by orthogonalization, are repeated until for all the non-basic variables either $\mathrm{CNR}_j = 1$ or $\mathrm{TNR}_j = 1$. The execution of the first phase may already yield the optimal solution (stable, with minimal variance) if the correlation between the original explanatory variables is weak (they are nearly orthogonal). However, if there is a considerable colinearity among the explanatory variables, the order in which they enter the basis may change their effect on the reduction of the variance. In such cases, changing the order of variable selection (by rotation) can lead to a solution with a smaller variance. Therefore, during the rotation phase, the variables in the basis are rotated so that each of them is tested versus the non-basic variables and reselected as the last one to enter the basis.

Before starting a new (rotation) phase, all the variables are set back to their original values. Only the order at which they entered the basis in the previous phase is retained. If due to this rotation, a new variable enters the basis, a new rotation cycle (a new phase) is initiated. Otherwise, when all basic variables are reselected as the last to enter the basis, an optimal regression model has been obtained.

4. Implementation of the SROV algorithm and use of the SROV program

The SROV algorithm was implemented as a collection of MATLAB M-files. The SROV program can be downloaded from the ftp site: ftp://ftp.bgu.ac.il/shacham/SROV. To use the program, the user should provide data file/s and an M-file that specify the problem to be solved. Explanation concerning the content of these files follows.

4.1. Specification of the experimental data and the error estimates

The experimental (or observed) data should be stored column-wise in a text (ASCII) file, where the dependent variable data is stored in the last column (note that only one dependent variable is allowed at this time).

There are several options to specify the error. These options can be selected by setting the value of two parameters of the SROV program, namely the *data_file_type* and the *error_type* parameters (for description of the various parameters and their default values, see Table 6). The estimated error in the data can be specified as absolute or relative (%) error level for the whole set of data associated with a variable. Otherwise, an absolute error can be specified for each individual data point. This error matrix is stored in a separate file and loaded to the program.

If the error level is specified, the program generates an error matrix assuming that the specified error level represent an average error. Following the suggestion by Stewart (1987), the elements of the error matrix are calculated from the equation:

$$\delta x_{ij} = \frac{5Rn}{3}\varepsilon_{xj} \quad (19)$$

where Rn is a random number with a zero mean and a unit variance and ε_{xj} is the specified error level in the jth independent variable. The calculation of the error vector for the dependent variable is carried out using the same equation, but replacing ε_{xj} by ε_y and δx_{ij} by ε_j. There are two options for generating the random numbers. If the program parameter *rand_type* is set at zero, the random number generator is reset to its 0th state in every run of the program. If *rand_type* = 1, the random number generator is reset to a different state for each run.

Error level estimates are often unavailable for observational or experimental data. In such cases, the assumption that the data are accurate to all reported figures and subject only to rounding errors, can provide error level estimates. Based on this assumption, Stewart (1987) recommended the use of the expression 0.3×10^{-t} for error level estimation, where t is the digit at which rounding occurs.

4.2. Definition of the regression model

There are several options available for model definition and they are governed by the parameters: *model*, *freeparm*, *transform* and *maxorder*. The model can be linear, including only the user-specified independent variables. If a quadratic model is requested, the program generates additional explanatory variables by co-multiplication of the independent variables. If there is only one independent variable, a polynomial model can be

Table 6
SROV program parameters, parameter interpretations and default values

Number	Parameter	Value and interpretation	Default value
1	data_file_type	0-No error matrix is specified	0
		1-Error matrix is specified	
2	error_type	0-Absolute error	0
		1-Relative error (%)	
3	freeparm	0-No free parameter in the model	
		1-There is a free parameter in the model	1
4	inter_level	0-Lowest level of interaction	0
		1-Highest level of interaction	
5	maxorder	1-30-Maximal degree of a polynomial term	20
6	model	0-Linear	
		1-Quadratic (several independent variables)	1
		1-Polynomial (one independent variable)	
7	plot_level	0-No graph plotting	
		1-Residual plot	1
		2-Residual, normal probability and calculated versus experimental values	
8	prob_title		(['default problem title'])
9	rand_type	0-Generator is reset to 0th state	0
		1-Generator is reset to a different state, every time	
10	transform	0-No data transformation	
		1-Standardization	1
		2-Normalization	
		3-Transformation to the range $[-1,+1]$	

requested, in which case additional explanatory variables containing various degrees of the independent variable (up to the degree specified by the parameter *maxorder*) are generated. The options to generate a linear, quadratic or polynomial model with (*freeparm* = 1) or without (*freeparm* = 0) a free parameter are offered.

Before generating the additional explanatory variables, data transformation is carried out, if so requested. The transformations available are: standardization, normalization and transformation to the $[-1, +1]$ range. Standardization is a common practice in statistical studies and involves removing the mean and then normalizing to unit standard deviation each of the independent variables. Note that the dependent variable is not transformed.

Normalization is commonly used in engineering and science to generate dimensionless variables. Dividing elements of every data column with the maximal (absolute) value in the same column normalizes the independent and dependent variables.

Shacham and Brauner (1997) found the transformation to the range of $[-1, +1]$ especially useful in polynomial regression, as it minimizes the interdependency between the model parameter values. This transformation is defined by:

$$\mathbf{z} = \frac{2\mathbf{x} - x_{\max} - x_{\min}}{x_{\max} - x_{\min}} \tag{20}$$

As in the case of standardization, the dependent variable is not transformed.

4.3. Controlling user interaction and display of results

The program can carry out the complete SROV algorithm (as described in the previous section) automatically. However, the user may override the decisions regarding the variables that are selected to the basis and regarding the maximal number of variables to be included. To let the program run without intervention (batch mode), the parameter *inter_level* must be set at zero. In the batch mode, only results obtained after the completion of the initial phase and after each rotation phase are reported. These results include all the $\hat{\beta}$ and $\check{\beta}$ values, the respective confidence intervals, the variance and sum of squares of the errors. In the batch mode, no graphs are plotted.

In the interactive mode (*inter_level* = 1), the program stops after every step that involves a decision, waiting for the user response to either confirm or change the decision suggested by the program. To assist the user decision, the values of YX_j, CNR_j and TNR_j are displayed for the three non-basic variables with the highest absolute YX_j values. In the interactive mode, various plots can be requested. With *plot_level* = 1, a residual plot is displayed for the solution found at the end of each phase. With *plot_level* = 2, a normal probability plot is also displayed and a plot of experi-

mental data points and calculated curve is prepared (only in case of one independent variable).

4.4. Loading the data, input of error levels and change of default parameter values

The commands to load a particular set of data, input the values of ε_{xj} and ε_y and changing the default parameter values can be carried out interactively or by running an M-file which includes these commands. The details of the various commands will be explained and demonstrated in the next section in connection with the two motivating examples.

5. Finding optimal solutions for the motivating examples with the SROV program

5.1. Example 1

The file that includes the commands for loading the data, setting the estimated error levels and changing default program parameter values for Example 1 is shown in Appendix A. The data for this problem is stored in the text file *Fearn.dat*. This file has to be loaded and its content is transferred to a matrix *xyData0*, which is a variable recognized by the program. A problem title can be specified for documentation purposes, by entering it into the variable: *prob_title*. By default, the program carries out standardization of the data. In this example, the original data is used without any transformation (*transform* = 0). To obtain a linear model, the parameter *model* is set at zero.

There is no information regarding the precision of the data, thus it is assumed that the data is accurate up to all reported figures. Following the discussion in the previous section, the average error level in $L1, L2, \ldots L6$ is set at 0.3 and in the independent variable (% protein) at 0.003. These values are entered into the vector *errx0* and the variable *erry0*.

The data file and the command file (shown in Appendix A) are sufficient for a complete definition of this problem. The second part of the Appendix shows the results that are displayed during the initial basic-variables selection phase, when an interactive mode of operation is selected.

At each step of the initial phase, the variable number, the values of YX_j, TNR_j and CNR_j are displayed for the first three variables with the highest absolute YX_j values. The program selects the variable with the highest absolute YX_j value to enter the basis next (provided that both $\mathrm{TNR}_j > 1$ and $\mathrm{CNR}_j > 1$), but the user can override this selection. After a variable has entered the basis, the respective $\check{\beta}$ and confidence interval values, as well as the current model variance, are displayed. The statistical indicators (the variance and confidence intervals) are calculated at this stage in order to provide the user a basis for comparison with the indicators based on signal to noise ratio and reassessment of the error level estimates.

In this particular example, YX_2 (YX_j associated with the independent variable $L2$) has the highest absolute value ($YX_2 = 0.55154$). Consequently this variable is selected to enter the basis first. The order of the first three non-basic variables (according to their YX_j values) after completion of the first step is $L4$, $L1$ and $L6$, where all $\mathrm{TNR}_j > 1$ and $\mathrm{CNR}_j > 1$. Variable $L4$ is selected to enter the basis at step two. This addition leads to reduction of the variance by an order of magnitude, while its parameter value is significantly different from zero (the confidence interval is still significantly smaller (in absolute value) than the respective parameter value). The order of the first three non-basic variables (according to their YX_j values) after completion of the second step is $L5$, $L3$ and $L6$, where only $\mathrm{CNR}_5 > 1$. Thus, only $L5$ can be still added to the regression model. Adding this variable leads to a moderate decrease of the variance, however, the associated parameter value is still significantly different from zero. At this point, no more variables satisfying the criterion: $\mathrm{CNR}_j > 1$ are left. Hence, the phase of the initial base selection has been completed and the regression results, in terms of the original non-orthogonalized variables, are displayed. These results are shown in Table 7. The model obtained in the initial phase includes only three, out of the six, variables: $L2$, $L4$ and $L5$. This model is stable, since all the confidence intervals are smaller than the respective parameter values.

The 1st rotation phase identifies a solution that includes variables $L3$, $L4$ and $L5$, which is also stable and has significantly smaller variance than the basic solution. An additional rotation phase shows that this model cannot be improved further. The results of the 1st rotation are also shown in Table 7. Comparing the variance and the R^2 values of this stable solution ($s^2 = 0.05057$ and $R^2 = 0.978$) with those of the unstable solution that contains all six variables ($s^2 = 0.04855$ and $R^2 = 0.991$), see Table 3 shows very little differences. Thus, the inclusion of additional variables in the model renders it unstable without improving its accuracy.

In Table 8, the final results of the SROV program are shown for the 'prediction set' data and for a data set that combines the data from the 'calibration' and 'prediction' sets. It can be seen that in both cases the stable regression model includes three out of the six variables: $L3$, $L4$ and $L6$. The coefficients of $L3$ and $L4$ are very similar to the coefficients obtained using the data of the 'calibration' set, but in these last two models $L5$ was replaced by $L6$.

Table 7
Results of the SROV program for the 'Calibration set' of Example 1

Parameter	Initial phase		1st rotation	
	Value	95% Confidence interval	Value	95% Confidence interval
β_0	54.65315	6.181407	32.61907	2.792364
β_1	0	–	0	–
β_2	0.230258	0.025349	0	–
β_3	0	–	0.242654	0.018364
β_4	−0.2132	0.030033	−0.23087	0.021878
β_5	0.015601	0.010774	0.008339	0.007428
β_6	0	–	0	–
Variance	0.104039		0.0505747	
R^2	0.954995		0.9781223	

It can be concluded that the SROV has identified a minimum variance, stable model for both the 'calibration' and the 'prediction' sets. However, based on the limited data of these two samples alone, there is still an uncertainty whether L5 or L6 should be included in the regression model as to better represent the whole calibration curve.

5.2. Example 2

The file that includes the commands for loading the data, setting the estimated error levels and changing program default parameter values for Example 2 is shown in Appendix B. The data for this problem are stored in the text file: *marquardt.dat*. This file has to be loaded and its content is transferred to a matrix *xyData0* (recognized by the program). The problem title is entered by the variable: *prob_title*. By default, the program carries out standardization of the data, in this case normalized data (transform = 2) is used.

The average error levels for the various variables are determined using the same considerations as in Example 1. The only exception is for the temperature; following the information provided by Kunugi et al. (1961) concerning the experimental error, an average error level of 2.5 °C is assumed.

The results obtained by the SROV program for this example are summarized in Table 9. At the initial phase, a regression model containing five variables: x_1, x_2, x_1x_2, x_2x_3 and x_2^2, with a variance of $s^2 = 0.000712$ is identified. Three consecutive rotations yield three additional solutions with consecutively decreasing variances. The solution of the lowest variance is obtained at the completion of the 3rd rotation. The model contains five variables: x_1, x_2, x_1x_2, x_2x_3 and x_1^2. This model is stable (the 95% confidence intervals are smaller, in absolute value than the respective parameter values). The values of the variance and the linear correlation coefficient are very close to those obtained with the full quadratic model of nine explanatory variables ($s^2 = 0.0004186$ and $R^2 = 0.99497$ for the optimal model, compared with $s^2 = 0.0003186$ and $R^2 = 0.9977$ for the full quadratic model). The residual plot of the optimal model (not shown) is also very similar to the residual plot of the full quadratic model (see Fig. 3).

It can be concluded that the SROV program has identified a stable regression model that represents the data adequately. In addition to this 'optimal' model, several sub-optimal' models have also been identified. These 'sub-optimal' solutions may also be considered

Table 8
Results of the SROV program for the 'Prediction set' and 'Combined sets' of Example 1

Parameter	Prediction set		Combined sets	
	Value	95% confidence interval	Value	95% confidence interval
β_0	23.401965	3.1706242	23.42421	2.920564
β_1	0	–	0	–
β_2	0	–	0	–
β_3	0.2661082	0.0163408	0.259497	0.011405
β_4	−0.2146058	0.0141748	−0.21018	0.010862
β_5	0	–	0	–
β_6	−0.0483577	0.0147088	−0.04618	0.0139
Variance	0.044495		0.049382	
R^2	0.983362		0.978452	

Table 9
Results of the SROV program for Example 2

Parameter (variable)	Value				95% confidence interval
	Base solution	1st rotation	2nd rotation	3rd rotation	
β_0	−4.6104	−2.0825	0.39421	7.5908	6.7437
β_1 (x_1)	5.5134	–	−5.4159	−21.0655	14.598
β_2 (x_2)	3.387	4.1783	4.5374	7.2047	2.3536
β_3 (x_3)	–	–	–	–	–
β_4 (x_1x_2)	−3.1136	−3.9455	−4.3427	−7.2904	2.3302
β_5 (x_1x_3)	–	–	–	–	–
β_6 (x_2x_3)	0.20879	0.056781	–	−0.53664	0.47242
β_7 (x_1^2)	–	3.0065	5.9635	14.5226	7.8744
β_8 (x_2^2)	−0.31782	−0.2787	−0.248	–	–
β_9 (x_3^2)	–	–	–	–	–
Variance	0.00071207	0.00053764	0.00044592	0.00041857	
R^2				0.994974	

for representing the data, particularly in cases where additional considerations dictate their preference over the 'optimal' solution.

6. Program tests and further developments

In the course of the program development, many benchmark problems involving data analysis and regression model identification were solved using SROV. Some of the results have been already reported in detail. In Shacham and Brauner (1999a), for example, a polynomial model is fitted to heat capacity data of solid 1-propanol and a quadratic model was used to correlate the yield of precipitation of $CaHPO_4 \cdot 2H_2O$ as function of three variables. Brauner and Shacham (2002) present two examples. One involves fitting a power-law expression of dimensionless numbers to heat transfer data after linearization of the expression (by taking logarithm of the two sides of the expression). In the second example, vapor pressure data is correlated using a bank of 23 nonlinear expressions involving various nonlinear functions of T_R (reduced temperature). SROV is used to select the variables to be included in an optimal model to represent $\ln P_R$ (reduced vapor pressure).

Work is currently underway to prepare and put on the web a large collection of benchmark problems, which were used for testing the SROV program. This library will contain the data and the optimal model/s that were identified by SROV.

It may often happen that the dependent variable data cannot be adequately represented by a linear (or a quadratic model) because of non-Gaussian error distribution. Such a situation can be detected using the residual and normal probability plots provided by the SROV program. In such cases, the Box–Cox (maximum likelihood) transformation of the dependent variable can be applied to linearize the regression model and obtain normal error distribution. The SROV program can at the present be used in conjunction with the Box–Cox transformation: the SROV is put in an internal loop, while in the outer loop, a minimization of the variance is carried out by changing the Box–Cox parameter. Work is currently under way to include the Box–Cox parameter search as an integral part of the SROV program.

7. Conclusions

The use of the SROV program for solving regression and data analysis problems involving linear, polynomial and quadratic models has been demonstrated. It has been shown that if the components of the model (the explanatory variables) cannot be determined a priory, using a stepwise regression procedure (such as SROV) is a must. Including non-influential or collinear explanatory variables in the model may lead to an unstable model where, for example, there is uncertainty regarding the direction of the change of the dependent variable due to changes in some of the explanatory variables.

In the examples presented, the SROV has identified the optimal model (a stable model with minimal variance) and identified also several sub-optimal models. The sub-optimal models help to discriminate between non-influential and collinear variables. Non-influential variables do not show up in any of the sub-optimal models and they can be safely removed from any further consideration. Variables that show up in some sub-optimal models and do not show up in others, are associated with a certain level of colinearity. In such cases and with a limited sample size, the distinction between the optimal and sub-optimal models may be sample dependent. A regression model, which seems optimal for one particular sample may not be the

optimal for a different sample of the same accuracy level, of the same or a different size. Additional user considerations may dictate replacement of the optimal model by one of the sub-optimal models.

It can be concluded that in addition to providing an optimal and stable solution, the SROV program provides also insight regarding the relationships that exist between the explanatory variables, between the explanatory variables and the dependent variable and information on model related uncertainties caused by sample size and experimental errors.

Appendix A: Commands' file and partial results of the SROV program for Example 1

Commands file

```
load Fearn.dat
xyData0 = Fearn;
prob_title = (['Calibration of a Near Infrared Reflectance Instrument']);
transform = 0;
model = 0;
for i = 1:6
   errx0(i) = 0.3;
end
erry0 = 0.003;
```

Selection of the variables included in the model in the initial stage

Starting initial base selection. Press a key to continue

Var. No.	x^*y (norm.)	TNR	CNR
2	0.55154	57.814	39.117
3	0.53734	67.448	46.611
1	0.46667	72.266	47.651

The new base variable selected is var. No.2. Press enter to accept or type in a different number. Type in 0 (zero)to finish >

Stage No.	Beta	Variance	Conf. interval
1	0.027591	1.3987	0.018

Var. No.	x^*y (norm.)	TNR	CNR
4	−0.95174	9.489	12.44
1	−0.8832	4.4786	5.9117
6	−0.71211	4.7727	4.4372

The new base variable selected is var. No.4. Press enter to accept or type in a different number. Type in 0 (zero) to finish >

Stage no.	Beta	Variance	Conf. interval
2	−0.18543	0.13773	0.026439

Var. No.	x^*y (norm.)	TNR	CNR
5	0.55972	19.791	2.5538
3	−0.45512	1.6544	0.93572
6	−0.30526	3.9019	0.79514

The new base variable selected is var. No.5. Press enter to accept or type in a different number. Type in 0 (zero) to finish >

Stage number	Beta	Variance	Conf. interval
3	0.015601	0.099085	0.010482

Initial base selection finished. Press a key to display the results.

Appendix B: Commands' file for the SROV program for Example 2

```
load marquardt.dat
xyData0 = marquardt;
prob_title = ['Thermal cracking of hydrocarbons to acetylene'];
transform = 2;
errx0(1) = 2.5;
errx0(2) = 0.03;
errx0(3) = 0.0003;
erry0 = 0.03;
```

References

Brauner, N., & Shacham, M. (1998a). Role of range and precision of the independent variable in regression of data. *American Institute of Chemical Engineers Journal 44* (3), 603–610.

Brauner, N., & Shacham, M. (1998b). Identifying and removing sources of imprecision in polynomial regression. *The Journal of Mathematics and Computers in Simulation 48* (1), 77–93.

Brauner, N., & Shacham, M. (1999). Regression diagnostic using an orthogonalized variable based stepwise regression procedure. *Computers and Chemical Engineering 23 Supplement*, S327–S331.

Brauner, N., & Shacham, M. (2002). A software toolbox for data analysis and regression, considering data precision and numerical error propagation. In: *Proceedings of the ESCAPE 12 conference*, The Hague, 26–29 May 2002.

Dahlquist, G., Bjork, A., & Anderson, N. (1974). *Numerical methods*. Englewood Cliffs: Prentice-Hall.

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data*. New York: Wiley.

Fearn, T. (1983). A misuse of ridge regression in the calibration of near infrared reflectance instrument. *Applied Statistics 32* (10), 73–79.

Himmelblau, D. M. (1970). *Process analysis by statistical methods*. New York: Wiley.

Kunugi, T., Tamura, T., & Naito, T. (1961). New acetylene process uses hydrogen dilution. *Chemical Engineering Progress 57* (11), 43–49.

Marquardt, D. W., & Snee, D. R. (1975). Ridge regression in practice. *The American Statistician 29* (2), 3.

Neter, J., Wasserman, W., & Kutner, M.H. (1990). *Applied linear statistical models*, Irwin, Burr Ridge.

Setzmann, U., & Wagner, W. (1989). A new method for optimizing the structure of thermodynamic correlation equations. *International Journal of Thermophysics 10* (6), 1103.

Shacham, M., & Brauner, N. (1997). Minimizing the effects of colinearity in polynomial regression. *Industrial & Engineering Chemistry Research 36* (10), 4405–4412.

Shacham, M., & Brauner, N. (1999a). Considering precision of experimental data in construction of optimal regression models. *Chemical Engineering and Processing 38*, 477–486.

Shacham, M., & Brauner, N. (1999b). A general framework for considering data precision in construction of optimal regression models. In: *Proceedings of the 5th International Conference on Foundations of Computer Aided Process Design*, Breckenridge, CO, 18–23 July 1999.

Stewart, G. W. (1987). Colinearity and least squares regression. *Statistical Sciences 2* (1), 68–100.

Wagner, W. (1973). New vapor pressure measurements for argon and nitrogen and a new method for establishing rational vapor pressure equations. *Cryogenics 13*, 470–482.