ELSEVIER

Chemical Engineering and Processing 38 (1999) 477–486

Chemical Engineering and Processing

www.elsevier.com/locate/cep

Considering precision of experimental data in construction of optimal regression models☆

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 Israel, Tel-Aviv 69978, Israel

Received 26 March 1999; accepted 12 April 1999

Abstract

Construction of optimal (stable and of highest possible accuracy) regression models comprising of linear combination of independent variables and their non-linear functions is considered. It is shown that estimates of the experimental error, which are most often available for engineers and experimental scientists, are useful for identifying the set of variables to be included in an optimal regression model. Two diagnostical indicators, which are based on experimental error estimates, are incorporated in an orthogonalized-variable-based stepwise regression (SROV) procedure. The use of this procedure, followed by regression diagnostics, is demonstrated in two examples. In the first example, a stable polynomial model for heat capacity is obtained, which is ten times more accurate than the correlation published in the literature. In the second example, it is shown that omission of important variables related to reaction conditions prevents reliable modeling of the product properties. © 1999 Elsevier Science S.A. All rights reserved.

Keywords: Collinearity; Data; Noise; Precision; Stepwise regression

1. Introduction

Obtaining experimental data is often very expensive and time consuming. However, the accuracy and reliability of process related calculations critically depend on the accuracy, validity and stability of the regression models fitted to the experimental data.

Regression models used for physico-chemical, thermodynamic or rate data 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 results in an inaccurate model characterized by a large variance. Some independent variables which may have critical effects on the dependent variable under certain circumstances, may be omitted. On the other hand, the inclusion of too many explanatory terms 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. Brauner and Shacham [1–3] have demonstrated some of the ill effects of including too many terms of regression models.

The most frequent causes of inaccuracy and/or ill-conditioning in regression are the following:

1. Non-optimal or inadequate model (not all the influential explanatory variables are included in the model and/or non-influential variables are included).
2. Excessive errors in the data (as in the presence of outlying measurements).
3. Presence of collinearity among the explanatory variables.

☆ Dedicated to Professor Em. Dr-Ing. Dr h.c. mult. E.-U. Schlünder on the occasion of his 70th birthday.

* Corresponding author. Tel.: +972-7-6461481; fax: 972-7-6472916.

E-mail address: shacham@bgumail.bgu.ac.il (M. Shacham)

PII: S0255-2701(99)00044-6

4. The algorithm used to calculate the model parameters is highly sensitive to numerical error propagation.

In view of the possible existence of several different causes for non-optimality, ill-conditioning or statistical invalidity of the regression model, the selection of the most adequate, optimal model should proceed in an iterative manner as follows [4–6]:

1. Suggestion of initial pool of explanatory variables which can potentially be included in the regression model.
2. Carrying out a stepwise regression procedure to identify the variables which should be included in an optimal, statistically valid model. An algorithm, which has a low sensitivity to numerical error propagation must be used for regression.
3. Identification of the cause(s) that limit the accuracy and the number of explanatory variables that can be included in the model (collinearity, outlying measurements, influential variables missing from the initial pool, etc.).

Then, remedial actions are taken and the process is reiterated from step 1 to check whether further improvement of the model is possible.

There are several stepwise regression algorithms and programs available (for details see, for example [7,6]). Most algorithms, however, are appropriate for linear regression models, and models containing a linear combination of non-linear functions of the independent variables (such as $1/x$ or log (x)) are considered as non-linear models in statistical analysis. Furthermore, existing programs may overwhelm engineering users with irrelevant and sometimes conflicting statistical information. Therefore, stepwise regression is rarely used by engineers.

In this paper, a stepwise regression procedure based on orthogonal variables (SROV) is presented. This procedure is well-suited to carry out the iterative process for selection of the optimal regression model, because it has low sensitivity to the harmful effects of collinearity and it provides reliable numerical indicators that help to identify the dominant cause(s) that limit the number of terms and accuracy of a statistically valid regression model. In the next section, the SROV procedure and its incorporation in an iterative framework for selecting the optimal regression model are described.

Two examples are presented. These examples demonstrate that the use of the proposed iterative procedure enables to pinpoint the dominant causes that limit the accuracy and stability of the regression model used. In many cases, the application of this procedure leads to a more accurate and stable regression models then the ones published in the literature.

The calculations related to the examples were carried out using the POLYMATH 4.0[1] and MATLAB 5.2[2] packages.

2. Stepwise regressing using orthogonalized variables (SROV)

A standard linear regression model can be written:

$$\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 \ldots + \beta_n\boldsymbol{x}_n + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{y}$ is an N-vector of the dependent variable, $\boldsymbol{x}_j (j = 1,2,\ldots n)$ are N vectors of explanatory variables, $\beta_0,\beta_1\ldots\beta_n$ are the model parameters to be estimated and ε is an N-vector of stochastic terms (measurement errors). It should be noted that an explanatory variable can represent an independent variable or a function of one or more independent variables.

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

$$\boldsymbol{x}_j = \hat{\boldsymbol{x}}_j + \delta\boldsymbol{x}_j \tag{2}$$

where $\hat{\boldsymbol{x}}_j$ is an N-vector of expected value of $\boldsymbol{x}_j$ and $\delta\boldsymbol{x}_j$ 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:

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\beta} = \boldsymbol{X}^T\boldsymbol{y} \tag{3}$$

where $\boldsymbol{X} = [1,\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots\boldsymbol{x}_n]$ is an $N(n+1)$ data matrix and $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{A}$ is the normal matrix. This method is subject to accelerated numerical error propagation due to collinearity (see for example, [2]). The SROV procedure [8] is much less sensitive to numerical error propagation and as such, is more appropriate to be used in a general purpose stepwise regression program.

A schematic flow diagram of the SROV procedure is shown in Fig. 1. Basically, the procedure consists of successive stages, where at each stage 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 stage, 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.

When progressing from stage k to stage $k+1$ in the stepwise regression procedure, the parameter estimate

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

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

corresponding to the explanatory variable x_p (selected as a basic variable at stage k) is obtained by:

$$\hat{\beta}^{k+1} \equiv \hat{\beta}_p = \frac{\mathbf{y}_k^T \mathbf{x}_p}{\mathbf{x}_p^T \mathbf{x}_p}; \quad \mathbf{x}_p \equiv \mathbf{x}^{k+1} \tag{4}$$

Then, the updated values of the dependent variable y^{k+1}, which represent the residual values that are not explained by the variables included in the basis at stage k are calculated by:

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \beta_p \mathbf{x}_p \tag{5}$$

The model variance at this stage,

$$s^2 = \frac{(\mathbf{y}^{k+1})^T(\mathbf{y}^{k+1})}{\nu} \tag{6}$$

is the sum of squares of errors divided by the degrees of freedom ($\nu = N - (n+1)$). The variance is a measure for the variability of the $\hat{y}$ values predicted by the regression model. Smaller variance indicates a better fit of the model to the data.

The confidence interval, $\Delta\beta_p$ on a parameter estimate can be defined:

$$\Delta\beta_p = t(\nu,\alpha)\sqrt{s^2(\mathbf{x}_p^T \mathbf{x}_p)} \tag{7}$$

where $t(\nu, a)$ is the statistical t distribution corresponding to ν degrees of freedom and a desired confidence level, α and s is the standard error of the estimate. Clearly, if $\hat{\beta}_p$ is smaller (in its absolute value), than the term $\Delta\beta_p$, then the zero value is included inside the confidence interval. Thus, there is no statistical justification to include the associated term in the regression model.

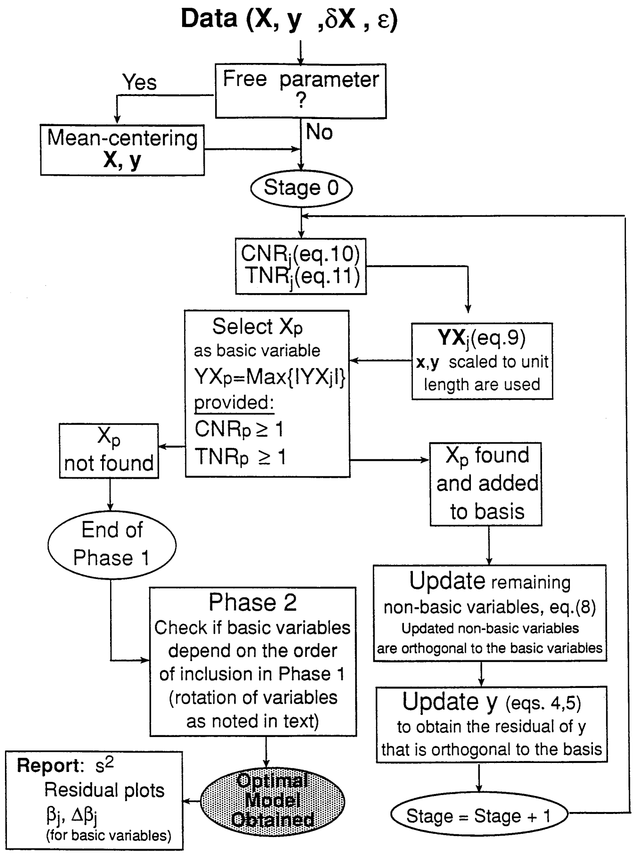
Data (X, y ,δX , ε)
Free parameter ?
Yes
No
Mean-centering X, y
Stage 0
CNRj(eq.10)
TNRj(eq.11)
YXj(eq.9)
x,y scaled to unit length are used
Select Xp as basic variable YXp=Max{IYXjI} provided: CNRp ≥ 1 TNRp ≥ 1
Xp not found
Xp found and added to basis
End of Phase 1
Update remaining non-basic variables, eq.(8)
Updated non-basic variables are orthogonal to the basic variables
Phase 2
Check if basic variables depend on the order of inclusion in Phase 1 (rotation of variables as noted in text)
Update y (eqs. 4,5) to obtain the residual of y that is orthogonal to the basis
Report: s2
Residual plots
βj, Δβj
(for basic variables)
Optimal Model Obtained
Stage = Stage + 1

Fig. 1. Schematic flow diagram of the SROV procedure.

Finally, the orthogonal components (residuals) of the non-basic variables are obtained by:

$$\mathbf{x}_j^{k+1} = \mathbf{x}_j^k - \mathbf{x}_p\left(\frac{(\mathbf{x}_j^k)^T \mathbf{x}_p}{\mathbf{x}_p^T \mathbf{x}_p}\right) \tag{8}$$

2.1. Criteria used for selection of an additional basic variable

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

$$YX_j = \mathbf{y}^T \mathbf{x}_j \tag{9}$$

where $\mathbf{y}$ and $\mathbf{x}_j$ are centered and normalized to a unit length. The value of $|YX_j|$ is in the range [0,1]. In case of a perfect correlation between $\mathbf{y}$ and $\mathbf{x}_j$ (y is aligned in the $\mathbf{x}_j$ direction), $|YX_j| = 1$. In case y is unaffected by $\mathbf{x}_j$ the two vectors are orthogonal), $YX_j = 0$. The inclusion of a variable $\mathbf{x}_p$, which has the highest level of correlation with $\mathbf{y}$, in the basic set (YX_p value is the closest to one) will affect the maximal reduction of the variance of the regression model. Therefore, the criterion $\mathbf{x}_p = \mathbf{x}_j\{\max |YX_j|\}$ is used to determine which of the non-basic variables should preferably be included in the regression model at the next stage, provided that the following CNR and TNR tests are both satisfied. The CNR_j measures the signal-to-noise ratio of YX_j, and is defined by:

$$CNR_j = \left\{\frac{|\mathbf{y}^T \mathbf{x}_j|}{\sum_{i=1}^{N}(|x_{ij}\varepsilon_i| + |y_i \delta x_{ij}|)}\right\} \tag{10}$$

Note that the denominator of Eq. (10) represents the error in YX_j as estimated via the error propagation formula. A value of $CNR_j \gg 1$ signals that the correlation between $\mathbf{x}_j$ and $\mathbf{y}$ is significantly larger than the noise level. Thus, an accurate value of YX_j can be calculated. But when $CNR_j \leqslant 1$, the noise in YX_j, as affected by $\delta\mathbf{x}_j$ and $\boldsymbol{\varepsilon}$, is as large as, or even larger than $|YX_j|$. If this is the case, no reliable value for $|YX_j|$ can be obtained and the respective variable should not be included in the regression model.

The TNR_j measures the signal-to-noise ratio in an explanatory variable x_j. It is defined in terms of the corresponding Euclidean norms [1]:

$$TNR_j = \frac{\|\mathbf{x}_j\|}{\|\delta \mathbf{x}_j\|} = \left\{\frac{\mathbf{x}_j^T \mathbf{x}_j}{\delta \mathbf{x}_j^T \delta \mathbf{x}_j}\right\}^{1/2} \tag{11}$$

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

For calculation of CNR_j and TNR_j, $\boldsymbol{\varepsilon}$ and $\delta \boldsymbol{x}_j$ values must be provided by the user. These values are based on the estimated experimental error or precision of the measurement and control devices. In the first stage of the regression, the values provided by the user are used. In subsequent stages, those estimates are updated using numerical perturbation. To this aim, the regression is carried out using two data sets in parallel, the original one and a perturbed data set. The differences between the updated values of the explanatory variables (Eq. (8)) and the dependent variable (Eq. (5)), obtained with the original and perturbed data sets, provide the estimates for the updated values of ε and δx_j at each new stage. (For a more detailed explanation, see [8].)

The selection of new variables (from among the non-basic variables) to be added to the basic variables in the SROV procedure stops when for all the non-basic variables either $CNR_j \leq 1$ or $TNR_j \leq 1$.

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

2.2. Phase 1. Finding an initial solution

In the 0-th stage of phase 1, the dependent and explanatory variables are centered (except for models which do not include a free parameter where centering is avoided).

The first variable selected to enter the basis is determined using Eqs. (9)–(11) and the concepts that were explained in the previous section. Then the calculation of the corresponding parameter value β^1 and the updating of the dependent and the remaining explanatory variables (and the corresponding errors) is carried out. In all subsequent stages of phase 1, the selection of additional variables to enter the basis and the updating of the dependent and explanatory variables is performed the same way as in the 0-th stage. As noted above, the stopping criteria of phase 1 is $CNR_j \leq 1$ or $TNR_j \leq 1$ for all the remaining non-basic variables.

When the correlation between the original explanatory variables is weak (they are nearly orthogonal) the regression model found in phase 1 is the optimal with a minimal variance (and the sum of square errors) value. However, if there is a considerable collinearity 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, rotation of the variables can lead to a solution with a smaller variance.

2.3. Phase 2. Rotation of the basic variables

In this phase, the variables in the basis are rotated so that each of them is tested versus the nonbasic variables and reselected as the last one to enter the basis. Before starting a new 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 a new variable enters the basis during rotation, a new rotation cycle (a new phase) is started.

2.4. Presentation of the optimal model

For the variable included in the optimal regression model the parameter estimate, $\hat{\beta}_j$ confidence intervals, $\Delta\beta_j$ and the ratio of confidence interval to parameter value $\Delta\beta_j/|\beta_j|$ for the orthogonalized variables are reported.

For a statistically valid model all $\Delta\beta_j/|\beta_j|$ must be smaller than 1. However, since the use of the model with orthogonalized variables may not be convenient for practical application, parameter estimates compatible with the original variables are also presented. These are obtained via a back-substitution algorithm.

The presentation of the optimal model includes also a residual plot of the error $\boldsymbol{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ versus $\boldsymbol{y}$ (where $\boldsymbol{y}$ is the vector of measured values of the dependent variables and $\hat{\boldsymbol{y}}$ is the vector of estimated values). The residual plot helps to verify that the regression model represents the data correctly (random error distribution) and helps to identify outlying observations. Standardized residuals: ε_i/s are also presented to assist with the decision concerning removal of outliers.

For the variables, which were not included in the optimal model, the final values of YX_j, TNR_j and CNR_j are presented. These variables can assist in determining the dominant cause that limits the accuracy of the regression model and indicate the possible direction of the actions that should be taken to further improve the regression model. Regression diagnostic using the results of the SROV procedure and the possible actions that can be taken to improve the model are described in the next section.

3. Regression diagnostics and model improvement

After an optimal regression model has been found, the indicators of the SROV procedure can be used for further diagnostic, in order to identify the actions that should be taken for further improvement of the regression model. Three typical cases are considered:

1. All $\Delta\beta_j/|\beta_j| < 1$ for the variables included in the model and all $CNR_j \leq 1$ (or are very close to 1) for variables not included in the model. In this case, a stable and statistically valid model has been obtained. The inclusion of additional explanatory variables in the model is prevented by the level of the noise (i.e. experimental error). The model can be further improved by providing more precise data of y and X.

Table 1
Selected information for heat capacity of solid 1-propanol [12]

	T (K)	Cp (J/K mol)	v	z	$y=(Cp/Cp_{\max})$
Minimal value	10.83	2.397	0.057308	−1	0.01798
Maximal value	188.98	133.3	1	1	1
No. of data points	50				
Avg. δT (K)	0.003				
Avg. δCp (J/deg mol)	0.03				
Avg. δv	1.59E-05				
Avg. δz	3.37E-05				

2. All $\Delta\beta_j/|\beta_j| < 1$, there are still variables not included in the model for which $CNR_j > 1$ but $TNR_j < 1$ (or very close to 1). In this case, collinearity among the explanatory variables prevents the inclusion of additional variables for increasing the model precision. Extending the range of the experiments and/or improving the precision of the independent variables data should be considered for eliminating collinearity. In polynomial or quadratic models, data transformations (such as the z-transformation, see Eq. (13) below) can often alleviate the ill-effects of collinearity and enable addition of more explanatory terms to the model ([2,9]).
3. One or more $\Delta\beta_j/|\beta_j| > 1$, and there are still variables not included in the basis for which $TNR_j > 1$, and $CNR_j > 1$. In this case, the variance is being inflated by either the use of an inappropriate model (the structure is incorrect and/or important explanatory variables have not been considered) or due to unexpected excessive error in y (as in the presence of outlying measurements).

Outlying measurements can be identified in the residual plot. Removing the outliers from the data set (for reducing the variance) can be considered. There are statistical tools for identifying suspected outliers. However, it is always recommended to re-confirm outlying points by repeated experiments or physical justification.

An inappropriate model structure and/or omission of important explanatory terms can sometimes be identified in the residual plots $(\boldsymbol{y}-\hat{\boldsymbol{y}})$ versus $\boldsymbol{y}$ and/or versus the explanatory variables.

An inappropriate linear model can be a result of neglection of non-linear effects. Theoretical consideration can sometimes be used to improve the model (for example, minimizing the relative error instead of the absolute error, [10]) or to identify the explanatory variables that should be added to the model. In the framework of empirical models, non-linear effects can be accounted for by extending the linear model to a quadratic model (inclusion of second order terms) or by including higher order polynomial terms. A quadratic model is defined by:

$$\boldsymbol{y} = \beta_0\boldsymbol{x}_0 + \sum_{j=1}^{n}\beta_j\boldsymbol{x}_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_{ij}\boldsymbol{x}_i\boldsymbol{x}_j + \boldsymbol{\varepsilon} \qquad (12)$$

The use of the z-transformation [9]

$$z = \frac{2\boldsymbol{x} - x_{\max} - x_{\min}}{x_{\max} - x_{\min}} \qquad (13)$$

can be beneficial when polynomial or quadratic models are used. This transformation yields a variable distribution in the range: $-1 \le z \le 1$. The v-transformation is defined by $v_i = x_i/x_{\max}$, where $x_{\max}$ is the largest x (in absolute value). This transformation yields a variable distribution in the region $v_{\min} \le v_i \le 1$. The benefits of extending a linear model to a quadratic model (or to higher polynomial terms) and employing the z-transformation (instead of no transformation or v-transformation) will be demonstrated in the examples.

Two examples are shown, which demonstrate the advantages of using the SROV procedure for obtaining the optimal model and for regression diagnostics.

4. Examples

4.1. Example 1. Heat capacity of solid 1-propanol

Heat capacity versus temperature date are usually correlated by polynomials. Daubert and Danner [11], for example used a 3rd order polynomial to correlate heat capacity (Cp) data versus temperature for solid 1-propanol, as published by Counsel et al. [12]. Selected information from this data is shown in Table 1.

Fitting a 3rd order polynomial (as recommended by Dauber and Danner [11]) to the normalized Cp ($y = Cp/Cp_{\max}$) versus temperature (T in K) yields the following polynomial model (95% confidence intervals are shown in parenthesis):

$$\begin{aligned} \mathrm{y} = {} & -0.127417(0.0197) + 1.20074(0.0931)\times 10^{-2}T \\ & -7.30919(1.099)\times 10^{-5}T^2 + 2.08259(0.365) \\ & \times 10^{-7}T^3 \end{aligned}$$

with a variance estimate $s^2 = 2.41\times 10^{-4}$. This model is stable (all confidence intervals are significantly

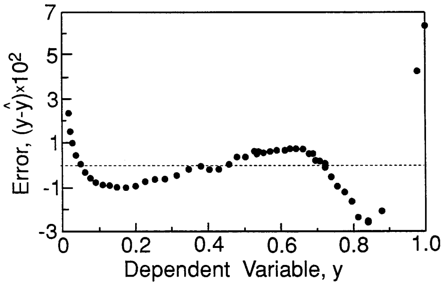

Fig. 2. Residual plot for regression of normalized *Cp* versus temperature data by a 3rd order polynomial.

smaller than the respective parameter values), but not very accurate. In Fig. 2, the residual plot of the error $\varepsilon = y - \hat{y}$ versus y is shown. It can be seen that the *Cp* curve has significant curvature, that cannot be explained by a 3rd order polynomial. There are two data points that can be marked as potential outliers. These two points correspond to the highest temperature and *Cp* values. It is well-known (see [13]) that *Cp* data near the melting point can be very inaccurate due to premelting. Removing the two potential outliers reduces the variance of the 3rd order polynomial representation to $s^2 = 3.53 \times 10^{-5}$, but the unexplained curvature of the residual plot remains.

To find a model that can represent better the *Cp* data, polynomials of various orders were considered as explanatory variables and the v and z-transformation were used. The full data set, as well as the reduced data set (with the two outliers removed), were considered. Using z-transformation for the reduced data set yielded a compact seven parameters model, while the other options (full data set and/or with v-transformation) lead to optimal models comprised of many more terms and parameters. For the sake of brevity, only the results obtained using z-transformation with the reduced data set will be given in some detail.

The SROV procedure was employed in order to select the terms that should be included in optimal polynomial model, from a pool containing various powers of z, up to z^{15}.

In Fig. 3, mean-cantered and scaled (to unit length) values of various powers of z are plotted versus mean-

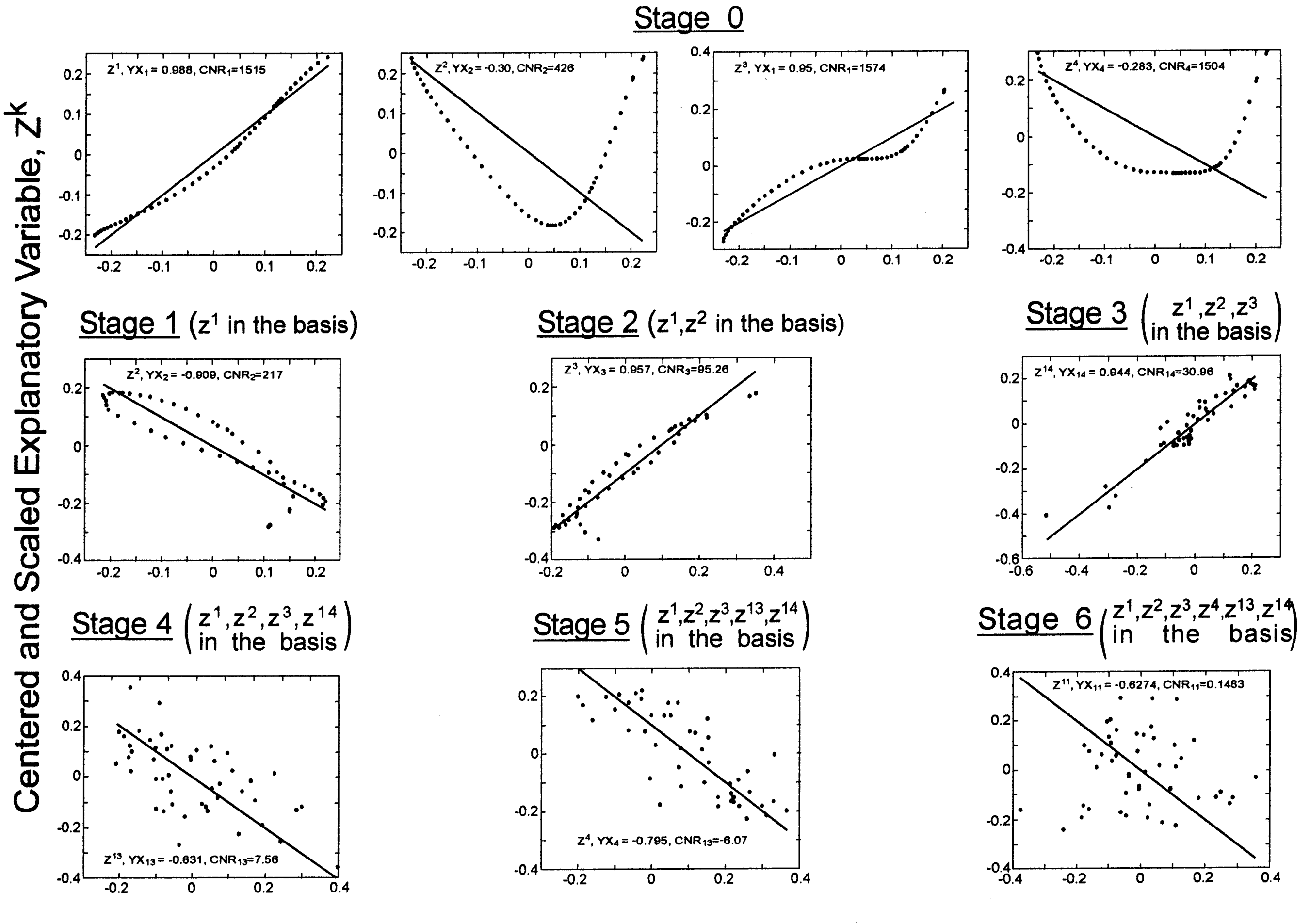

Fig. 3. y versus Z^k at various stages of tire SROV procedure (example 1).

Table 2
Results of the SROV procedure for example 1[a]

Variables included in the regression model				Variables not included in the regression model			
Var. no.: j	β_j	$\Delta\beta_j$	$\Delta\beta_j/\lvert\beta_j\rvert$	Var. no.: j	YX_j	$TNRj$	CNR_j
1	0.472460	0.0004213	0.000892	11	−0.02744	961.5	0.1483
2	−0.125890	0.0008113	0.006444	15	0.02688	796.0	0.1482
3	0.110090	0.001623	0.014746	9	−0.02454	1074.4	0.1322
14	0.040253	0.003065	0.076141	7	−0.01961	1221.3	0.1041
13	−0.011070	0.002195	0.198310	5	−0.01619	1425.6	0.0861
4	0.046730	0.003162	0.067669	8	0.01118	1074.0	0.0532
				6	0.01105	1245.2	0.0521
				12	−0.00737	850.6	0.0360
				10	0.00311	948.4	0.0149

[a] z transformation, orthogonalized variables, outliers removed.

cantered and scaled values of Cp (y) at the various stages of the SROV procedure. The respective values of YX_j and CNR_j are also shown.

At the 0th stage, it can be seen that the data points for z^1 are aligned almost perfectly along the straight line with slope of 1, that represents a perfect correlation. The respective YX_1 value is 0.988. The shape of the curve obtained from the z^3 value is similar to that of z^1, but there is considerably larger curvature. Consequently, the value of YX_3 (0.95) is smaller than the value of YX_1. At this stage, the linear correlation between y and z^2 or z^4 is weak. This is indicated by small absolute values of YX_2 ($=0.30$) and YX_4 ($=$ 0.283). However, at this stage, all TNR_j and CNR_j are much greater than one, thus, stability considerations do not exclude inclusion of anyone of the variables in the regression model.

Because of its highest YX_j, z^1 is entered into the model at the 0th stage. In Fig. 3, the updated values of z^2 are plotted versus the updated values of y, as obtained at stage 1. At this stage, there is already a considerable spread of the data points, but the points of z^2 line up nicely along the straight line of slope -1, with $YX_2 = -0.909$. Thus, the residual of z^2, becomes strongly correlated with the residual of y, which is orthogonal to and cannot be explained by z^1. At stage 1, z^2 is added to the model. At stage 2, the residual of z^3 becomes the most collinear with the residual of y (see the plot of stage 2 in Fig. 3), yielding $YX_3 = 0.957$. It is interesting to note that CNR_3 was reduced by more than an order of magnitude after the information which is collinear to z^1 have been subtracted from z^3 and y.

After z^3 has been added to the model, the consecutive power of z (z^4) has lower YX_j value ($=0.715$) than higher powers of z. At this point, z^{14} has the highest correlation with the residual of y ($YX_{14} = 0.944$, see also the plot of stage 3 in Fig. 3). The SROV continues by adding z^{14} to the model at stage 3, z^{13} ($YX_{13} = -$ 0.631) at stage 4 and z^4 ($YX_4 = -0.795$) at stage 5. At stage 6, YX_{11} is the highest, however, CNR_{11} and all other the CNR values become smaller than one, thus, no more variables should be added to the model. The plot of the residual of z_{11} versus y at stage 6 (see Fig. 3) also shows that the correlation is very weak and the distribution of the points is nearly random.

In Table 2, the results of the SROV procedure in terms of orthogonalized variables, are summarized. The optimal model includes z, z_2, z_3, z_4, z_{13} and z_{14}. The $\Delta\beta_j/\lvert\beta_j\rvert$ ratios for all of them are significantly smaller than one, thus, this model is stable. For the variables not included in the model all $TNR_j \gg 1$, indicating that collinearity does not prevent addition of more variables. However, all $CNR_j < 1$, thus combined imprecision of the independent and dependent variables data limits the number of terms in the model.

In Table 3, the parameter values and the variance estimate for the optimal model using non orthogonalized z-transformed variables are shown. The variance estimate is $s^2 = 9.15 \times 10^{-7}$, smaller by almost two orders of magnitude than the variance of the 3rd order polynomial. The residual plot for the optimal model is shown in Fig. 4. It can be seen that, in constrast to the 3rd order polynomial (see Fig. 2), the error is distributed randomly and the maximal errors are of the order of 2.5×10^{-3}, an order of magnitude smaller than for the 3rd order polynomial.

Table 3
Regression results for example 1[a]

Var. no.: j	β_j
0	0.619590
1	0.378440
2	−0.123110
3	0.121200
4	−0.026676
13	−0.009320
14	0.040253
s^2	9.1459E-7

[a] Optimal model, z transformation, outliers removed.

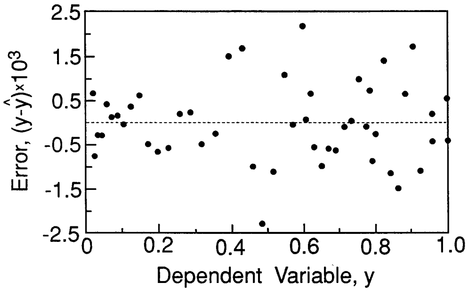

Fig. 4. Residual plot for the optimal regression model obtained for normalized *Cp* versus polynomial of *z*.

Thus, the SROV procedure enabled identifying a much more accurate, stable regression model then the one that is recommended by Daubert and Danner [11].

4.2. Example 2. Precipitating stoichiometric $CaHPO_4 \cdot 2H_2O$

This example was discussed extensively in the chemical engineering literature [14] and the statistical literature [7].

The three independent variables are the mole ratio of the NH_3 to $CaCl_2$ in the calcium chloride solution, addition time (*t*, min) of ammoniacal $CaCl_2$ to $NH_4H_2PO_4$ and starting pH of $NH_4H_2PO_4$ solution. These variables were coded[3] so that $-5/3 < x_j < 5/3$ $(j = 1,2,3)$ using the transformations $x_1 = [(NH_3)/(CaCl_2) - 0.8]/0.09$; $x_2 = (t - 50)/24$ and $x^3 = (\mathrm{pH} - 3.5)/0.9$. The coded values of the independent variables are shown in Table 4. Seven dependent (response) variables were measured, only three of them are used in this work: Y_1, the yield (percent of theoretical), Y_2, Fisher subsieve size (microns) and Y_6, B.E.T. specific surface area (m^2/g). The measured values of Y_1, Y_2 and Y_6 are also shown in Table 4. Aia et al. [14] suggested the use of a full quadratic model for regression of Y_1. Based on analysis of the variance, they concluded that the terms associated with x_2 can be removed from the model, but included all the remaining terms in the model. Table 5 summarizes the values of β_j, $\Delta\beta_j$, $\Delta\beta_j/|\beta_j|$ and s^2 for the model suggested by Aia et al. [14]. It can be seen that for two of the terms (x_1x_3 and x_1^2) $\Delta\beta_j/|\beta_j| > 1$, indicating that these terms can probably be removed from the model.

In Fig. 5a, the residual plot is shown when Y_1 is regressed with the six parameters model which was proposed by Aia et al. [14]. The errors are larger for larger Y_1 values and there is a clear trend in the residuals that cannot be explained with the proposed model.

For further investigation of the adequacy of the proposed quadratic model, the SROV procedure was employed for identifying the optimal model. It was assumed that the independent and dependent variables data are accurate up to the number of decimal digits reported (for the uncoded data). Thus, the average error values used are $\delta x_1 = 0.033$, $\delta x_2 = 0.0125$, $\delta x_3 = 0.033$ and $\delta Y_1 = 0.03$ (for details of the error estimation, see [15]).

The results of the SROV procedure are shown in Table 6. The optimal model includes variables x_1, x_3 and x_3^2. All the $\Delta\beta_j/|\beta_j|$ values for the model are smaller than one, thus, the four-parameter (three-variable model) is statistically valid. However, all the TNR_j and CNR_j values for the remaining non-basic variables are greater than one, indicating the possibility that additional variables can be added to the model. But, attempting inclusion of one more variable yields $\Delta\beta_j/|\beta_j|$ value larger than one, thus, a stable model contains only the three variables x_1, x_3 and x_3^2. This situation corresponds to case no. 3 in the regression diagnostic section, where an inappropriate model due to ommission of an important explanatory variable is mentioned as possible cause of inflated variance. Consulting the residual plot for the optimal model (shown in Fig. 5b) substantiates this conclusion. There is a clear trend in the residuals that is not explained by the proposed model.

The ommission of important variables is even more evident when Y_2 and Y_6 are regressed with the quadratic model that contains the same variables. Aia et al. [14] concluded that for regression of Y_2, a linear model containing only the independent variable x_2 is the most appropriate model and there is no justification to add any more terms to the model. The use of the SROV procedure yields the same result. The optimal model obtained is $\hat{Y}_2 = 9.365(0.6279) + 1.06968(0.7627)$. Although CNR_j and TNR_j for all remaining non-basic variables are greater than one, none can be included in the model, because of the excessively large confidence intervals resulting from the large model variance.

In Fig. 6, the residual of Y_2 is plotted versus x_2, when the single variable linear model is used for regression. It can be seen that x_2 alone cannot represent Y_2, but since all the other explanatory variables must be excluded from the model, there must be additional variable(s) (not included in the reported data) that can explain the lack of fit of the linear model.

This conclusion is further reinforced by examining the data of Y_6, the additional dependent variable. Aia et al. [14] could not find any statistically valid model to represent Y_6 with the reported independent variables and concluded that Y_6 is a constant, which is varying within the experimental error. But Y_6 varies over a considerable range $(0.44 \leq Y_6 \leq 1.49)$ and the explana-

[3] Note that these bounds are different than in the paper by Aia et al. [14], where there were apparently typographical errors.

Table 4
Data for example 2 [14]

Point no.	x_1	x_2	x_3	y_1	y_2	y_3
1	−1	−1	−1	52.8	8	1.29
2	1	−1	−1	67.9	6.2	1.11
3	−1	1	−1	55.4	10.4	0.8
4	1	1	−1	64.2	10.6	0.52
5	−1	−1	1	75.1	7.6	1.49
6	1	−1	1	81.6	10.5	0.63
7	−1	1	1	73.8	12	0.69
8	1	1	1	79.5	9.8	0.96
9	−1.66667	0	0	68.1	7.8	0.88
10	1.66667	0	0	91.2	9.4	0.44
11	0	−1.66667	0	80.6	8.8	0.47
12	0	1.666667	0	77.5	11.2	0.58
13	0	0	−1.66667	36.8	10.5	1.2
14	0	0	1.666667	78	9.8	0.52
15	0	0	0	74.6	10.4	0.5
16	0	0	0	75.9	6.3	1.1
17	0	0	0	76.9	9.5	0.8
18	0	0	0	72.3	8	0.77
19	0	0	0	75.9	10	0.95
20	0	0	0	79.8	10.5	1.13

tion that the variation is only due to experimental error, is not at all convincing. More probable is that Y_6 is function of additional variable(s), which are not included in the reported data set, possibly the same variable(s) Y_2 depends on. To test this hypothesis, the SROV procedure has been employed after adding to the quadratic model Y_4 and Y_2 as additional explanatory variables. A statistically valid model $Y_6 = 2.4766(2.007) - 0.01089(0.009912)Y_1 - 0.09102(0.07553)Y_2$ has been obtained.

Since Aia et al. [14] reported only the measurements for the three independent variables, using their data, it is impossible to determine which important variable was omitted. However, it should be pointed out that isothermal operation at 30°C was assumed, although it was noted that a temperature rise of about of about 4°C occurred during precipitation. Since the physical properties of the product are known to widely vary with temperature [14], a possible omitted variable is the temperature rise during precipitation.

5. Conclusions

It has been shown that the use of experimental error estimates can be very beneficial in identifying optimal regression models and in regression diagnostic. The error estimates are considered in the framework of the SROV procedure.

In the first example, the use of the SROV procedure enabled obtaining an optimal regression model for heat capacity of solid 1-propanol, which contains non-con-

Table 5
Regression results, for example 2, independent variable y_1, model of Aia et al. [14]

Var. no.: j	β_j	$\Delta\beta_j$	$\Delta\beta_j/\lvert\beta_j\rvert$
0	76.3886	2.357	0.03086
1 (x_1)	5.48804	1.842	0.3356
2 (x_3)	10.1773	1.842	0.181
3 (x_1x_3)	−1.4625	2.407	1.646
4 (x_1^2)	0.642875	1.784	2.775
5 (x_3^2)	−7.22362	1.784	0.247
s^2	10.073		

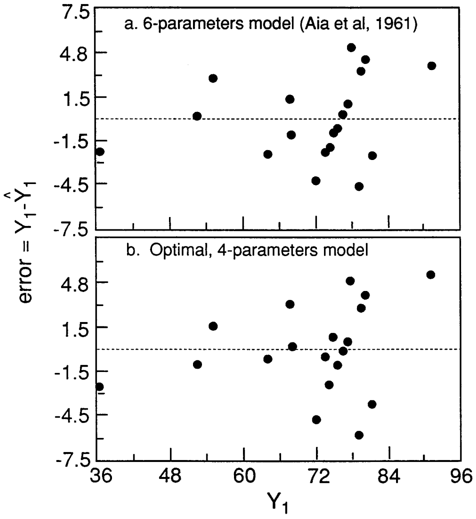

Fig. 5. Residual plot for regression of Y_1 with six parameter and four parameter models (example 2).

Table 6
SROV procedure and regression results for example 2 (independent variable y_1)

a. Centered-orthogonalized variables							
Variables included in the regression model				Variables not included in the regression model			
Var. no.: j	β_j	$\Delta\beta_j$	$\Delta\beta_j/\lvert\beta_j\rvert$	Var. no.: j	YX_j	$TNRj$	$CNRj$
1	0.38511	0.1257	0.3265	2	−0.2041	84.03	5.8063
3	0.7143	0.3404	0.4765	4 (x_1x_2)	−0.1951	31.68	5.0093
9 (x_3^2)	−0.53073	0.2279	0.4295	5 (x_1x_3)	−0.3215	35.84	8.1766
				6 (x_2x_3)	−0.0632	22.64	1.5029
				7 (x_1^2)	0.195	37.98	3.3587
b. Unaltered variables				8 (x_2^2)	0.126	38.52	2.2376
0	76.9095	1.977	0.02571				
1	5.50328	1.852	0.3365				
3	10.2074	1.852	0.1814				
9 (x_3^2)	−7.39851	1.807	0.2442				
s^2	10.34						

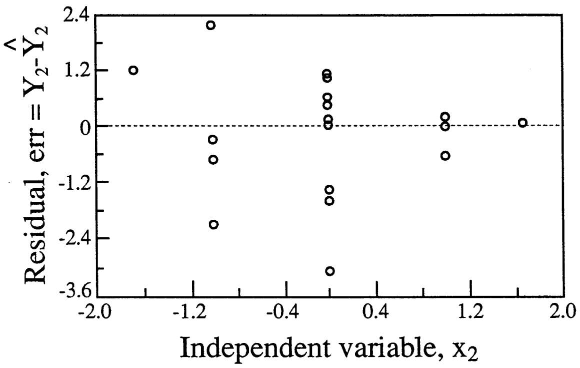

Fig. 6. Residual plot for regression of Y_2 versus x_2 (example 2).

secutive powers of the independent variable. The optimal model is a stable model, which is an order of magnitude more accurate than the 3rd order polynomial model that was recommended for this property in the literature.

In the second example, data precision considerations and the SROV procedure enabled identifying the omission of important variables as the main cause for preventing accurate and stable modeling of various properties of calcium hydrogen orthophosphate as function of reaction conditions. This cause was not detected previously, in spite of the extensive discussion of the problem in the chemical engineering and statistical literature.

Engineers and experimental scientists can usually obtain good estimates of the experimental errors. Using the various indicators and the SROV procedure presented in the paper, they can utilize the experimental error estimates to extract the maximal valuable information from the data. The outcome is either obtaining a satisfactory, most accurate and stable optimal regression model or identifying the main cause that limits the accuracy and stability of the optimal regression model.

References

[1] N. Brauner, M. Shacham, Role of range and precision of the independent variable in regression of data, AIChE J. 44 (3) (1998) 603–610.

[2] N. Brauner, M. Shacham, Identifying and removing sources of imprecision in polynomial regression, J. Math. Compt. Simul. 48 (1) (1998) 77–93.

[3] N. Brauner, M. Shacham, Considering numerical error propagation in modeling and regression of data, Proceedings of the ASEE Annual Conference, Seattle, Washington, 27 June–1 July, 1998.

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

[5] U. Setzmann, W. Wagner, A new method for optimizing the structure of thermodynamic correlation equations, Int. J. Thermophys. 10 (6) (1989) 1103.

[6] J. Neter, W. Wasserman, M.H. Kutner, Applied Linear Statistical Models, Irwin, Burr Ridge, 1990.

[7] N.R. Draper, H. Smith, Applied Regression analysis, 2nd edition, Wiley, New York, 1981.

[8] M. Shacham, N. Brauner, A stepwise regression procedure based on data precision and collinearity considerations, (1999) submitted for publication.

[9] M. Shacham, N. Brauner, Minimizing the effects of collinearity in polynomial regression, Ind. Eng. Chem. Res. 36 (10) (1997) 4405–4412.

[10] N. Brauner, M. Shacham, Statistical analysis of linear and non-linear correlation of the Arrhenius equations constants, Chem. Eng. Process. 36 (1997) 243–248.

[11] T.E. Daubert, R.P. Danner, Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation, Hemisphere Publishing Co, New York, 1989.

[12] J.F. Counsell, E.B. Lees, J.F. Martin, Thermodynamic properties of organic oxygen compounds. Part XIX. Low-temperature heat capacity and entropy of propan-1-ol, 2-methyl-propan-1-ol and pentan-1-ol, Inorg. Phy. Theor. J. Chem. Soc (A), (1968) 1819–1823.

[13] J. Timmermans, Physico Chemical Constants of Pure Organic Compounds, 2nd edition, Elsevier, New York, 1965.

[14] M.A. Aia, R.L. Goldsmith, R.W. Moone, Precipitating stoichiometric $CaHPO_4{\cdot}2H_2O$, Ind. Eng. Chem. 53 (1) (1961) 55–57.

[15] G.W. Stewart, Collinearity and least squares regression, Stat. Sci. 2 (1) (1987) 68–100.