A software toolbox for data analysis and regression, considering data precision and numerical error propagation.

Neima Brauner, School of Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel;
Mordechai Shacham, Chem. Eng. Dept., Ben Gurion University , Beer-Sheva 84105, Israel

Abstract

An algorithm for data analysis and regression by orthogonalized-variable-based stepwise regression (SROV) has been developed and was implemented as a MATLAB toolbox. The program uses QR decomposition based on Gram-Schmidth orthogonalization, which is highly resilient to numerical error propagation, for regression. 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 noise level. The use and benefits of SROV are demonstrated by two examples. The first one involves removing non-influential dimensionless groups from a regression model. In the second one the nonlinear terms that should be included in an optimal thermodynamic property correlation are selected.

1. Introduction

Analysis, reduction and regression of experimental and process data are critical ingredients of various CAPE activities, such as process design, monitoring and control. The accuracy and reliability of process related calculations depend on the accuracy, validity and stability of the regression models fitted to the experimental data. It is usually unknown, a priori, how many explanatory variables (independent variables and/or their functions) should be included in the model. An insufficient number of explanatory variables result in an inaccurate model, where some independent variables that under certain circumstances significantly affect the dependent variable, are omitted. On the other hand, the inclusion of too many explanatory terms renders an unstable model. Shacham and Brauner (1997) and Brauner and Shacham (1998 a, b) provide several examples where regression models published in the chemical engineering literature are grossly inaccurate and/or unstable. Shacham and Brauner (1999) have established the theoretical basis for considering precision of the data in determining how many and which explanatory variables (independent variables and their nonlinear functions) should be included in an optimal regression model. An optimal model is a stable model of highest possible accuracy. An algorithm for carrying out an orthogonalized-variable-based stepwise regression (SROV) process for the construction of optimal regression models have been developed and was implemented as a set of MATLAB script files (toolbox, Shacham and Brauner, 2001). The toolbox contains the SROV programs for fitting linear, quadratic, polynomial and general (linear combination of various nonlinear functions of the independent variables) models to data. The toolbox can be downloaded from the ftp site: ftp://ftp.bgu.ac.il/shacham/SROV.

In this paper the use of SROV for removing non-influential dimensionless groups from a model and for selecting the nonlinear terms that should be included in an optimal regression model for vapor pressure correlation are demonstrated. In the next section the two motivating examples are presented. After that the SROV algorithm is described briefly and the solutions of the motivating examples using SROV are presented. The computations reported in the paper were carried out with MATLAB 5.3 (trademark of The Math Works, Inc., http://www.mathworks.com) and POLYMATH 5.1 (copyrighted by M. Shacham, M. B. Cutlip and M. Elly, http://www.polymath-software.com)

.

2. Motivating Examples

2.1 Fitting a regression model to heat transfer data (Dow and Jacob, 1951; Lapidus, 1962)

Dow and Jacobs (1951) proposed the following dimensionless equation for representing experimental data dealing with heat transfer between a vertical tube and a fluidized air-solid mixture,

$$Nu = a_1 N_1^{a_2} N_2^{a_3} N_3^{a_4} \,\mathrm{Re}^{a_5} \qquad (1)$$

where $Nu = \frac{h_m D_t}{k_g}$; $N_1 = \frac{D_t}{L}$; $N_2 = \frac{D_t}{D_p}$; $N_3 = \frac{1-\varepsilon}{\varepsilon}\frac{\rho_s}{\rho_g}\frac{C_s}{C_g}$; $\mathrm{Re} = \frac{D_t G}{\mu_g}$

with h_m – heat transfer coefficient, D_t – tube diameter, D_p – solid particle diameter, L- heated fluidized bed length, ε – void fraction of fluid bed, G – gas mass velocity, k_g, ρ_g, C_g, μ_g- properties of gas phase and C_s, ρ_s – properties of solid phase.

Dow and Jacob (1951) and Lapidus (1962) tested the appropriateness of this model by regression of the linearized form of Eqn. (1) (linearized by taking logarithm of both sides of the equation) with the data shown in p. 354 in Lapidus (1962). Regression results obtained using the multiple linear regression option of the POLYMATH program for the linearized five-parameter model are shown in Table 1. The results include the parameter values, the 95% confidence intervals on the parameter values, the variance and the linear correlation coefficient (R^2).

The overall impression is that the model represents the data adequately. The residual plot (not shown) displays a random distribution of residuals with a maximal error of about 2% and R^2 = 0.99. But comparing the parameter values with their confidence intervals indicates potential faults of the regression model. For two of the parameters ($\ln(a_1)$ and a_4) the confidence intervals are larger than the parameter value, meaning that the value 0 (zero) is inside the confidence interval. Such a situation usually arises if non-influential independent variables are included in the model or/and there is collinearity between some of the variables. In this particular case a_1 (=1), and N_3 can be omitted from the fit. In section 4, the SROV program will be used for identifying the cause of the wide confidence intervals and to select the influential dimensionless groups from among the ones that were proposed.

2.2 Stepwise regression of vapor pressure data

Wagner (1973) proposed the following equation to represent vapor pressure data between the triple point and the critical point:

$$\ln P_R = \frac{1}{T_R}(a_1\tau + a_2\tau^{1.5} + a_3\tau^3 + a_4\tau^6) \qquad (2)$$

where a_1, a_2, a_3 and a_4 are adjustable parameters, T_R is reduced temperature (T_R = T/T_C , where T_C is the critical temperature), P_R is the reduced pressure ($P_R = P/P_C$, where P_C is the critical pressure) and $\tau = 1 - T_R$. This equation was fitted to vapor pressure data of nitrogen provided by Wagner (1973). The data set includes 68 data points between the triple point: T = 63.148 K, P= 0.1252 bar and the critical point: T_C = 126.2 K, P_C = 34.002 bar. Regression results obtained using the multiple linear regression option of the POLYMATH program for the Wagner equation are shown in Table 2. The results include the parameter values and their 95% confidence intervals, the variance and the linear correlation coefficient (R^2).

Wagner's equation represents the data very well. There is a random distribution of residuals, the error in $\ln P_R$ is well below 0.1% over most of the region and R^2 = 1.0 Wagner (1973) applied stepwise regression on a model containing 27 linear and nonlinear terms in order to arrive at the model of Eqn. 2. The question arises whether Wagner's model is really unique and optimal, or similar and even better models can be derived. In section 4 the SROV program will be used for identifying the terms that should be included in an optimal model, from among the following linear terms proposed by Wagner:
1. τ^{-1}, 2. $\tau^{-0.5}$, 3. $\tau^{0.5}$, 4. τ 5. $\tau^{1.5}$, 6. τ^2, 7. $\tau^{2.5}$, 8. τ^3, 9. $\tau^{3.5}$, 10. τ^4, 11. $\tau^{4.5}$, 12. τ^5, 13. $\tau^{5.5}$, 14. τ^6, 15. $\tau^{6.5}$, 16. τ^7, 17. $\tau^{7.5}$, 18. τ^8, 19. $\tau^{8.5}$, 20. τ^9, 21. $\ln\tau$, 22. $(1-\tau)^2\ln(1-\tau)$ and 23. $\ln T$

3. The SROV (Stepwise Regression using Orthogonalized Variables) algorithm

The SROV program has been described in detail in Shacham and Brauner (1999 and 2001). Here only a brief explanation, which is necessary in order for understanding the results presented, is given.

A standard linear regression model can be written as:

$$\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 \cdots + \beta_n\mathbf{x}_n + \boldsymbol{\varepsilon} \qquad (3)$$

where **y** is an *N*-vector of the dependent variable, $\mathbf{x}_j$(j = 1,2, … *n*) are *N* vectors of explanatory variables, $\bullet_0$, $\bullet_1 \ldots$ $\bullet_n$, are the model parameters to be estimated and • 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. The vector of estimated parameters: $\hat{\boldsymbol{\beta}}^T = (\hat{\beta}_0, \hat{\beta}_1, \cdots \hat{\beta}_n)$ is calculated via QR decomposition when the orthogonal Q matrix is constructed using the Gram-Schmidth method. This method is highly resilient to numerical error propagation. The QR decomposition 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 are updated to include the residual information that is orthogonal to that represented by the basic variables.

Variable selection is based on the indicator |*YXj*|, which represents the correlation between **y** and $\mathbf{x}_j$. The value of |*YXj*| is in the range [0,1]. In a case of a perfect correlation between the dependent variable **y** and the explanatory variable $\mathbf{x}_j$ (**y** is aligned in the $\mathbf{x}_j$ direction), |*YXj*| = 1. In case **y** is unaffected by $\mathbf{x}_j$ (the two vectors are orthogonal), |*YXj*| = 0. Two indicators are used for removing explanatory variables from consideration because of an insignificant signal-to-noise ratio. The indicator *TNRj* measures the signal-to-noise ratio in an explanatory variable $\mathbf{x}_j$. A value of *TNRj* >> 1 indicates that the explanatory variable $\mathbf{x}_j$, contains valuable information. On the other hand, a value of *TNRj* ≤ 1 implies that the information included in $\mathbf{x}_j$ is mostly noise, and therefore, it should not be added to the basic variables. The indicator *CNRj* measures the signal-to-noise ratio ratio of |*YXj*|. A value of *CNRj* >> 1 signals that the correlation between $\mathbf{x}_j$ and **y** is significantly larger than the noise level. But when *CNRj* ≤ 1 the noise in |*YXj*| is as large as, or even larger than |*YXj*|. If this is the case, the respective variable should not be included in the regression model. At each step, the variable with the highest |*YXj*| | value is selected to enter the basis (the model), provided that both TNR_j>1 and CNR_j>1. The selection of new variables (from among the non-basic variables) to be added to the basic variables stops when for all the non-basic variables either $CNR_j \leq 1$ and/or $TNR_j \leq 1$.

The SROV algorithm consists of two phases. In the first phase, an initial (optimal or nearly optimal) solution is found. In the second phase, the variables are rotated to change the order of their selection in an attempt to improve the model. In this phase, the variables in the basis are rotated so that each of them is tested versus the non-basic variables for its reselection as the last one to enter the basis. If rotation results in replacement of a basic variable by a non-basic one, a better solution of lower variance, has been found and a new rotation starts. The rotation is terminated when none of the basic variables can be replaced.

4. Application of SROV for the motivating examples

4.1 Example 1

The file that includes the commands for generating the data and error files and for changing default program parameter values for Example 1 is shown in Appendix A. The data for this problem (Lapidus, 1962) is stored in the text file: *heatrans2.dat*. This file is loaded and its contents are transferred to the five vectors: *N1*, *N2, N3*, *Re* and *Nu*. The SROV requires specification of the estimated error for each data point. For a model comprised of general (non-polynomial) functions of the independent variables an error matrix (of the same dimension as the data matrix) has to be provided. The generation of the error matrix is done in several steps. First, an average error level for each variable is specified. If no estimate of the average experimental error is available (as in this case), the assumption that the data are accurate to all reported figures, subject only to rounding error (Stewart, 1987) can be used. Based on this assumption, the expression $0.3*10^{-t}$ is used, where *t* is the digit at which rounding occurs. The error levels, determined according to these assumptions, are stored in the scalars (*N1_el*, *N2_el, N3_el*, *Re_el* and *Nu_el).* Next, a random and normally distributed error, with the mean set at the average error level and a variance of one, is added to all the data points. The resultant values are stored in vectors (*N1_e*, *N2_e,*

N3_e, *Re_e* and *Nu_e)* Finally, the data required by the linearized model are generated by taking the logarithm of the data and error matrices. The results are stored in *xyData0* and *xyErr0*, respectively (both variable names are recognized by the SROV program).

Problem title can be specified for documentation purposes, by entering it into the variable *prob_title*. By specifying *data_file_type* = 1, the program is informed that both data and error matrices are provided. By default, the program carries out standardization of the data, in this case we use the original data without any transformation (*transform*=0). To use a linear (non-quadratic) model, the parameter *model* is set at zero. To obtain both residual and normal probability plots, *plot_level* = 2 is specified (the default is *plot_level* = 1, residual plot only). The second part of Appendix A shows the results that are displayed during the initial base selection phase, when the interactive mode of operation is selected.

At every step of this phase the indicators $|YXj|$ (shown as x*y (norm.)), *TNRj* and *CNRj* are displayed for the three explanatory variables with the highest $|YXj|$ values. The program selects the variable with the highest absolute $|YXj|$ value to enter the basis (the model) next (provided that both $TNR_j>1$ and $CNR_j>1$) but the user can override this selection. After a variable has entered the basis the respective parameter value (Beta) and confidence intervals, as well as the current model variance are displayed. In this particular example, $|YX_1| = 0.7279$ ($|YX_j|$ associated with the independent variable N_1) has the highest valueand is thus the first variable selected for the basis. The order of the first three non-basic variables (according to their YX_j values) after completion of the first step is *Re*, N_2 and N_3 where all $TNR_j>1$ and $CNR_j>1$. *Re* is selected to enter the basis at step two. This addition leads to a significant reduction of the variance and the corresponding parameter value is significantly different from zero. Now, the order of the two remaining non-basic variables (according to their YX_j values) is N_2 and N_3 and still the corresponding $TNR_j>1$ and $CNR_j>1$. After selecting N_2 to enter the basis, the variance is reduced by an order of magnitude. The resultant $|YX_3|<<1$ (≈ 0.002) and $CNR_3<1$, indicating that the residual component of N_3 is non influential (nearly orthogonal to the residual of **y**). Hence, N_3 is not included in the initial basis. The second, rotation phase does not identify a better model or any additional acceptable (sub-optimal) models.

The final results of this analysis are shown in the right part of Table 2. For this four-parameter model, all the parameter values are significant and the variance is even smaller than that of the five-parameter model. Thus, based on the available data, there is no justification to include the N_3 group in the model.

4.2 Example 2.

The average error levels for the various variables in this example are determined using the same assumptions as in Example 1, implying an average error of 0.003 ˚C for the temperature and an average relative error of 0.03% for the pressure. Since this is a large-scale example, the details of the computations are not provided. All calculations are carried out by the SROV automatically, without any user intervention. The results shown in Table 3 include only the final results of the initial model selection and the final optimal and sub-optimal models identified during the second rotation phase.

At the end of the initial phase, a regression model containing five variables: τ, $\tau^{1.5}$, $\tau^{3.5}$, τ^{9} and $\ln T$, is identified with a variance of $s^2 = 5.303\text{e-}8$. Three consecutive rotations yield three additional solutions with consecutively decreasing variances. All four models include five variables with all parameters values significantly different from zero. The solution of the lowest variance is obtained at the completion of the 3rd rotation. This model contains the variables: τ, $\tau^{0.5}$, $\tau^{1.5}$, $\tau^{3.5}$ and τ^{6} with a variance of $s^2 = 4.1534\text{e-}8$. The variance of the other four sub-optimal models are either smaller or slightly larger than that of the Wagner's model ($s^2 = 4.729\text{e-}8$). This example demonstrates the benefits of employing the SROV program for the automatic selection of stable and of highest precision regression models in cases where a large number of potential explanatory variables are to be considered.

5. Conclusions

The use of the new SROV toolbox has been demonstrated for performing stepwise regression and data analysis. It was shown that in modeling of data in terms of dimensionless groups, apparent adequate representation of the data by the model, is not enough for justifying the inclusion of all the groups in the

correlation. In such cases, the SROV can be used to remove the non-influential groups from the model. In modeling of thermodynamic properties with a large bank of potential terms, SROV can provide the optimal model in addition to several sub-optimal models. The sub-optimal models provide flexibility for incorporating additional considerations, besides stability and minimal variance, in the selection of the best correlation to be used.

References

Brauner, N. and M. Shacham, 1998a, AIChE J., 44 (3), 603
Brauner, N. and M. Shacham, 1998b, The Journal of Mathematics and Computers in Simulation, 48 , 77
Brauner, N. and M. Shacham, 1999, Computers chem. Engng., 23. Supplement, S327
Dow, W. M. and M. Jacob, 1951, Chem. Eng. Progr. 47, 637
Lapidus, L., 1962, Digital Computation for Chemical Engineers, McGraw-Hill, New York
Shacham, M. and N. Brauner, 1997, Ind. Eng. Chem. Res., 36, 4405
Shacham, M. and N. Brauner, 1999, Chemical Engineering and Processing, 38, 477
Shacham, M. and N. Brauner, 2001, Submitted for publication
Stewart, G.W., 1987, Statistical Sciences 2 (1) 68
Wagner, W., 1973, Cryogenics 13, 470

Appendix A

Commands file and partial results of the SROV program for Example 1

1. Commands file

```
load heatrans2.dat
N1=heatrans2(:,1);
N2=heatrans2(:,2);
N3=heatrans2(:,3);
Re=heatrans2(:,4);
Nu=heatrans2(:,5);
randn('state',0);
R=randn(size(heatrans2));
N1_el=0.0003; %absolute error in N1
N2_el=0.3; %absolute error in N2
N3_el=0.3; %absolute error in N3
Re_el=0.3; %absolute error in Re
Nu_el=0.3; % absolute error in Nu
N1_e(:,1)=N1(:,1)+R(:,1)*5*N1_el/3;
N2_e(:,1)=N2(:,1)+R(:,1)*5*N2_el/3;
N3_e(:,1)=N3(:,1)+R(:,1)*5*N3_el/3;
Re_e(:,1)=Re(:,1)+R(:,1)*5*Re_el/3;
Nu_e(:,1)=Nu(:,1)+R(:,1)*5*Nu_el/3;
xyData0=[log(N1) log(N2) log(N3) log(Re) log(Nu)];
xyErr0=xyData0-[log(N1_e) log(N2_e) log(N3_e)
        log(Re_e) log(Nu_e)];
prob_title=(['Heat transfer in fluidized bed  ']);
data_file_type=1;
transform=0;
model=0;
plot_level=2;
```

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

Var. No.	x*y (norm.)	TNR	CNR
1	0.7279	4.15E+02	3.61E+02
3	0.67858	7.50E+02	5.19E+02
2	0.65617	574.87	435.3

The new base variable selected is var. No.1. Press enter to accept or type in a different number.

Step No.	Beta	Variance	Conf. interval
1	1.0754	0.14938	0.51835

Var. No.	x*y (norm.)	TNR	CNR
4	0.88594	470.52	424.88
2	0.46837	918.74	369.33
3	0.3728	1275.5	416.41

The new base variable selected is var. No.4. Press enter to accept or type in a different number.

Step No.	Beta	Variance	Conf. interval
2	0.7988	0.034143	0.22163

Var. No.	x*y (norm.)	TNR	CNR

2	0.95123	748.5	217.04
3	0.52642	1060.5	143.18

The new base variable selected is var. No.2. Press enter to accept or type in a different number.

Step No.	Beta	Variance	Conf. interval
3	0.34522	0.003466	0.061618
Var. No.	x*y (norm.)	TNR	CNR
3	-0.0021259	872.22	0.17183

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

Table 1. Regression results for Example 1, five and four parameters models

	Five parameters model		**Four parameters model**	
	Parameter Value	**95% confidence interval**	**Parameter Value**	**95% confidence interval**
$\ln a_1$	0.164374	1.028335	0.161584	0.634396
a_2	0.740198	0.118044	0.739988	0.097800
a_3	0.345383	0.080566	0.345224	0.064185
a_4	-5.70E-04	0.160737	-	-
a_5	0.786533	0.078299	0.786489	0.073993
variance	0.003999		0.003714	

Table 2. Regression results for Example 2, Wagner and minimal variance models

	Wagner's model		**Minimal variance model**	
Term	**Parameter value**	**95% confidence interval**	**Parameter value**	**95% confidence interval**
$\tau^{0.5}$	-	-	-0.009526	0.002735
τ	-6.101728	0.004510	-6.020998	0.017165
$\tau^{1.5}$	1.147282	0.012326	0.970093	0.027056
τ^{3}	-1.056454	0.027035	-	-
$\tau^{3.5}$	-	-	-1.267259	0.147817
τ^{6}	-1.888232	0.087268	-1.274936	0.057905
variance	4.729E-08		4.156E-08	

Table 3. Results of the SROV program for Example 2

Term	**Parm. value (base solutn.)**	**Parm. value 1st rotation**	**Parm. value 2nd rotation**	**Parm. value 3rd rotation**
$\tau^{0.5}$	-	-	-0.007189	-0.009526
τ	-6.077219	-6.072797	-6.039963	-6.020998
$\tau^{1.5}$	1.060295	1.048628	1.006335	0.970093
$\tau^{3.5}$	-1.531823	-1.479547	-1.425211	-1.267259
τ^{6}	-	-	-	-1.274936
$\tau^{7.5}$	-	-1.842981	-2.093281	-
τ^{9}	-3.868830	-	-	-
$\ln T$	-6.307E-05	-7.477E-05	-	-
variance	5.306E-08	5.069E-08	4.423E-08	4.156E-08