pp. 587 – 592 in A. Kraslawski and I. Turunen (Eds.), Proceedings of the European Symposium on Computer Aided Process Engineering-13, Elsevier, 2003

A Procedure for Constructing Optimal Regression Models in Conjunction with a Web-based Stepwise Regression Library

N. Brauner[1] and M. Shacham[2]
[1]School of Engineering, Tel-Aviv University, Tel-Aviv 699 78, Israel
[2]Chem. Eng. Dept., Ben-Gurion University, Beer-Sheva 84105, Israel

Abstract

Construction of an optimal, of highest precision and stable regression models is considered. A new algorithm is presented which starts by selecting the independent variables included in a linear model. If such a model is found inappropriate, increasingly more complex, higher precision models are considered. These are obtained by addition of nonlinear functions of the independent variables and transformation of the dependent variables. The proposed algorithm is incorporated in the SROV toolbox (Shacham, M. and N. Brauner, 2002, Computers chem. Engng., in press). Using an example, it is demonstrated that the algorithm generates several optimal models of gradually increasing complexity and higher precision from which the user can select the most appropriate model for his needs.

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 critically depend on the accuracy, validity and stability of the regression models fitted to 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, 1999). Often transformations (such as the Box and Cox, 1964 "maximum likelihood" transformations) of the dependent and/or some of the independent variables should be applied in order to obtain the most accurate and stable regression model.

The presently available stepwise regression programs have several shortcomings for use in CAPE related computations. They do not search for optimal value of the Box-Cox transformation parameters for the dependent and/or the independent variables, thus the search should be conducted manually. They may be highly sensitive to numerical error propagation caused by collinearity among the independent variables and yield inaccurate results without giving any warning concerning the inaccuracy. Most of them do not consider the accuracy of the data available in determining the number of variables to be included in the model, thus may yield an unstable regression model.

Development of better programs for stepwise regression and data reduction is hindered by the lack of data sets, which are large enough and representative to CAPE related

applications, and can be used both for defining the needs for further developments and for testing the software. In order to address this need, we have started developing a web-based library, which includes data such as physical and thermodynamic properties, process monitoring data and data used for estimating properties from descriptors of molecular structure. Some data sets contain over a hundred independent variables and over 200 data points. The library contains the data sets including information concerning the experimental error, pertinent references and optimal models that we have found.

In the course of library development, we have found that in general, applying stepwise regression and/or Box-Cox transformations separately does not yield all the optimal models. These and additional techniques should be applied in a systematic, procedural manner in order to obtain the best results. In the next section, some basic concepts will be reviewed and the proposed algorithm for the selection and identification of optimal regression model will be presented. In section three, the proposed procedure will be demonstrated using refinery data that were extensively discussed in the literature (Daniel and Wood, 1980). All the calculations are carried out with a modified version of the SROV program of Shacham and Brauner (2002).

2. Basic Concepts

A standard linear regression model can be written:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 \ldots + \beta_n \mathbf{x}_n + \boldsymbol{\varepsilon} \qquad (1)$$

where $\mathbf{y}$ is an N-vector of the dependent variable, $\mathbf{x}_j$ (j = 1,2, … n) are N vectors of explanatory variables, $\beta_0, \beta_1, ...\beta_n$ 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 The vector of estimated parameters $\hat{\beta}^T = \hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_n$ can be calculated via the least squares error approach by solving the normal equation:

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} \qquad (2)$$

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 rarely used for actual calculations, since it is subjected to an accelerated propagation of numerical errors in cases of colinearity (see for example, Brauner and Shacham, 1998). The condition number of the normal matrix, $\boldsymbol{\kappa}(\mathbf{A})$, (the ratio of the absolute values of the maximal to minimal eigenvalues) is used as a convenient measure of the ill-conditioning of the regression problem. Alternative methods for least-squares regression are described by Bjorck (1966).

The SROV program combines stepwise regression with QR decomposition to find the optimal regression model. The QR decomposition solves the equation $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}$ by decomposing X into the product of a matrix Q of orthogonal columns, and an upper triangular matrix, R. The SROV algorithm generates the Q matrix using the Gram-Schmidt method (see for example, Bjorck, 1966). 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. Addition of new variables to the model stops when the residual information of all remaining variables gets below their noise level. A detailed description of the SROV algorithm and the criteria used for variables selection and replacement can be found in Shacham and Brauner (1999,2002).

The quality of the regression model is assessed in view of numerical and graphical information, which includes the model variance, confidence intervals on the parameter estimates, the linear correlation coefficient, residual and normal probability plots. The model variance is defined: $s^2 = [(\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}})]/\nu$, where $\mathbf{y}$ and $\hat{\mathbf{y}}$ are the measured and calculated vectors of the dependent variable respectively, ν is the number of degrees of freedom ($\nu = N-(k+1)$) and k is the number of independent variables included in the model. The linear correlation coefficient is defined by $R^2 = [(\hat{\mathbf{y}} - \bar{y})^T (\hat{\mathbf{y}} - \bar{y})]/[(\mathbf{y} - \bar{y})^T (\mathbf{y} - \bar{y})]$, where $\bar{y}$ is the mean of $\mathbf{y}$. The variance and R^2 are used for comparison between various models, where a regression model that yields a smaller variance and the R^2 value closer to 1 is considered superior.

The confidence interval on parameter j ($\Delta\beta_j$) is defined by $\Delta\beta_j = t(\nu,\alpha)\sqrt{s^2 a_{jj}}$, where a_{jj} is the diagonal element of the inversed normal matrix, and $t(\nu,\alpha)$ is the statistical t distribution corresponding to ν degrees of freedom and a desired confidence level, α. A model where one or more of the confidence intervals are greater in absolute value than the associated parameter values (parameter value is not significantly different from zero) is considered unstable (or even ill-conditioned). Therefore, it is usually considered as unacceptable. The signal-to-noise ratio indicators, which are used by the SROV program for variable selection, usually remove from the model variables associated with insignificant parameter values. However, the removal of the free parameter (β_0) is the user's responsibility. This may be required based on theoretical considerations or due to excessive confidence intervals on this parameter.

If the distribution of the errors in the residual plot (plot of $\mathbf{y} - \hat{\mathbf{y}}$ versus $\mathbf{y}$) is random, so that no clear trend can be identified, the model can be considered as appropriate representation of the data. Otherwise, the use of Box-Cox transformation and/or addition of nonlinear functions of the independent variables should be considered. The Box-Cox transformation is a power transformation of the dependent variable: $y' = y^{\lambda}$, where the parameter λ is selected so as to minimize the variance of the resultant correlation. In order to enable a meaningful comparison of the variances of regression models obtained with different λ values, the dependent variable must be standardized. The standardized variables employed for the search are : $w = K_1(y^{\lambda} - 1)$ for $\lambda \neq 0$ and $w = K_2 \ln y$ for $\lambda = 0$, where $K_2 = \sqrt[n]{\prod y_i}$ and $K_1 = 1/(\lambda K_2^{\lambda-1})$. In case the Box-Cox transformation with linear terms of the independent variables does not yield a random distribution of the residuals, adding nonlinear functions of the independent variables should be considered. It is customary to use a full quadratic model (containing nonlinear functions of the form $x_i x_j$ and x_i^2) or polynomial model (including higher powers of x^i) as an initial bank of explanatory variables for carrying out the stepwise regression, unless theoretical considerations suggest different functional forms. It is worth noting that quadratic and polynomial models tend to be ill-conditioned if many terms or high powers of the independent variable are included in the model. Ill-conditioning of the model is indicated by a very large value of the condition number of the normal matrix. Such ill-conditioning can be prevented by transformation of the independent variables to the [-1, +1] (or similar) range. One of such transformation

supported by the SROV program is the standardization, where the transformed variable is defined by $z = (x - \bar{x}) / st.dev(\mathbf{x})$.

3. The Procedure for Constructing Optimal Regression Models

For the sake of brevity, only cases where there is no prior information (from theoretical considerations and/or from experience) on nonlinear functions of the independent variables to be included in the regression model, are considered. For the dependent variable, the commonly used function ln$y$ can be used as starting point for the search instead of y. Based on the principles outlined in the previous section, the search procedure for the optimal regression model can be outlined:

1. The Box-Cox parameter is set at $\lambda = 1$ for using y as dependent variable or $\lambda = 0$ for using ln$y$ as dependent variable. A search for the independent variables to be included in the optimal linear model is carried out using SROV. The search is repeated with β_0 set at zero and the best model is selected according to the variance and R^2 values. The residual plot for the selected model is examined. If the errors are randomly distributed finish, otherwise proceed to step 2.

2. A search for the Box-Cox parameter value that yields a minimal variance is carried out, using the SROV in an internal loop to select the independent variables to be included in the model with each value of λ. The residual plot for the selected model is examined. If the errors are randomly distributed finish, otherwise proceed to step 3.

3. Step 2 is repeated using a quadratic model (in case of several independent variables) or a polynomial model (one independent variable). The condition number of the normal matrix is checked. If it is not much larger than that of the linear model, finish. Otherwise proceed to step 4.

4. Step 3 is repeated using transformed independent variables. Note that transformation of the variables requires adding a free parameter to the model even if it was omitted in previous steps. The most appropriate model is selected by comparing the models obtained in steps 1 – 4 on the basis of the variance, R^2, the residual plots and additional practical considerations (complexity of the model, derivatives etc.)

4. Operation of a Petroleum Refining Unit – An Example

This example was first introduced by Gorman and Toman (1966) and since then was extensively used in the statistical literature. The data set contains 36 data points of 10 dependent variables and one independent variable, where each row in the data set represents one day of operation of a petroleum refining unit. The complete data set of this example is given in Daniel and Wood (1980). They have also carried out a stepwise regression analysis of the data set using a linear model that includes a free parameter and the transformation ln y for the dependent variable. This corresponds to Step 1 of the proposed algorithm with $\lambda = 0$. The optimal solution obtained for this case by SROV is shown in Table 1. Note that the range of the dependent variable, the parameter values and the variance correspond to $w = K_2 \ln y = 128.68 \ln y$. Six out of the ten variables are included in the regression model. In terms of stability, this is a borderline case, since the confidence interval on β_0 is very close to the parameter (absolute) value itself. The

residual plot for this model is shown in Figure 1, indicating a clear trend in the error distribution. For low values of the dependent variable (w), the residuals tend to be negative and for high values, the residuals are mostly positive. Thus, it is necessary to proceed to the following steps of the proposed algorithm.

The optimal results obtained by SROV in the various steps of the algorithm are summarized in Table 2. The insignificant value of β_0 obtained in step 1 implies that it can be removed from the model (step 1.2). Indeed, using a linear model with $\lambda = 0$, but without a free parameter improves the results in several respects. The variance decreases, R^2 gets closer to one, there are only six parameters in the model and all the parameter values are significantly different from zero. The residual plot, however, still indicates an opportunity for further model refinement. Proceeding to step 2 with a search for an optimal Box-Cox parameter ($\lambda_{optimal} = 0.3$) results in only a marginal reduction of the variance.

Introducing a quadratic model in step 3 increases the number of the potential explanatory variables to 65. As shown in Table 2, the inclusion of quadratic terms leads to significant improvements, where a stable model with 9 parameters and with a variance of about half the value obtained with the previous linear models. Obviously,

introducing non-linear effects of the independent variables may change the optimal value of the Box-Cox parameter. Thus, a search for the optimal value of λ is carried out simultaneously with the search for the variables to be included in the model. However,

Table 1 Optimal regression model for $\lambda = 0$, (K_2 = 128.68) linear model including a free parameter.

Parameter (variable)	**Value**	**Confidence interval**
β_0	758.0123	267.1427
β_1 (x_1)	-11.4787	11.4697
β_2 (x_3)	-223.1336	175.5181
β_3 (x_5)	-1078.4844	755.9046
β_4 (x_6)	136.0463	85.1292
β_5(x_8)	9.855	7.5778
β_6 (x_{10})	5.4129	3.6777
no. of parm.	7	
variance	1322.05	
R^2	0.851540	
κ(A)	5.4228E+05	

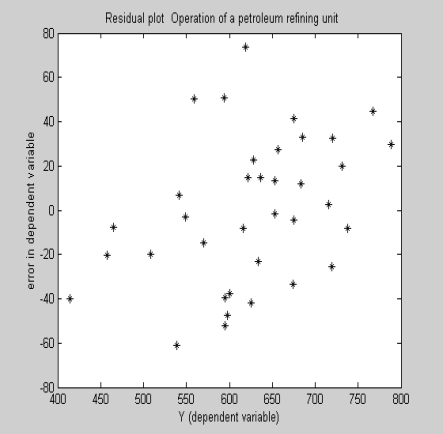

Figure 1. Residual plot for optimal regression model: $\lambda = 0$, linear model including a free parameter.

Table 2 Results summary of optimal models of the various stages of the algorithm.

Algorithm stage	1.1	1.2	2	3	4
λ	0	0	0.3	0.55	0.618
model	linear	linear	linear	quadratic	quadratic
transformation	no	no	no	no	yes
free parm.	yes	no	no	no	yes
no. of parm.	7	6	6	9	13
variance	1322.05	1190.50	1060.72	588.54	344.51
$\mathbf{R^2}$	0.85154	0.86266	0.86799	0.94024	0.97098
κ(A)	5.4228E+05	5.7534E+07	5.7534E+07	1.7786E+13	99.2685

since the condition number of the resulting quadratic model is greater by several orders of magnitude compared to those of the linear models, a transformation of the independent variables is advisable (step 4).

The optimal model obtained by SROV, using the quadratic model with transformed variables, includes 12 explanatory variables (13 parameters), where all the confidence intervals are significantly different from zero. The variance is the smallest of all the models tested and R^2 is the closest to one. The condition number is very small in comparison to the other models, thus this solution can be considered highly accurate. The linearity of the normal probability plot for this case indicates normal distribution of the residuals.

5. Conclusions

It has been demonstrated that the proposed algorithm suggests a considerable progress in modeling and regression of data, especially in cases where there is no a-priori information on the model structure neither from theory nor from experience. The algorithm starts by identifying the independent variables of an optimal (of the lowest variance and stable) linear model and gradually progresses to models of increasing complexity and precision as necessary. Along the route, the algorithm generates several optimal models from which the user can select the one that is most appropriate for his needs, while considering the model complexity and precision.

6. References

Bjorck A.,1996, Numerical Methods for Least Squares Problems, SIAM, Ph.,PA

Box, G. E. P. and Cox, D. R., 1964, J. of the Royal Statistical Society B, **26**, 211.

Brauner, N. and M. Shacham, 1998, J. of Math. and Computers in Simulation, 48, 77.

Daniel, C. and F. S. Wood, 1980, Fitting Equations to Data – Computer Analysis of Multifactor Data, 2nd Ed, John Wiley, New York.

Gorman, J. W. and R. J. Toman, 1966, Technometrics, 8, 27-51.

Shacham, M. and N. Brauner, 1999, Chem. Eng. Process. 38, 477.

Shacham, M. and N. Brauner, 2002, Computers chem. Engng. (in press).