47th AIAA Aerospace Sciences Meeting and Exhibit **AIAA 2009-1344**
5-8 January 2009, Orlando, Florida

Regression Model Optimization for the Analysis of Experimental Data

N. Ulbrich*
Jacobs Technology Inc., Moffett Field, California 94035–1000

A candidate math model search algorithm was developed at Ames Research Center that determines a recommended math model for the multivariate regression analysis of experimental data. The search algorithm is applicable to classical regression analysis problems as well as wind tunnel strain–gage balance calibration analysis applications. The algorithm compares the predictive capability of different regression models using the standard deviation of the PRESS residuals of the responses as a search metric. This search metric is minimized during the search. Singular value decomposition is used during the search to reject math models that lead to a singular solution of the regression analysis problem. Two threshold dependent constraints are also applied. The first constraint rejects math models with insignificant terms. The second constraint rejects math models with near–linear dependencies between terms. The math term hierarchy rule may also be applied as an optional constraint during or after the candidate math model search. The final term selection of the recommended math model depends on the regressor and response values of the data set, the user's function class combination choice, the user's constraint selections, and the result of the search metric minimization. A frequently used regression analysis example from the literature is used to illustrate the application of the search algorithm to experimental data.

Nomenclature

$\mathbf{A}$	= matrix containing regressors of the original data set
$\overline{\mathbf{A}}$	= matrix containing regressors of the modified data set
a_1, a_2	= regression coefficients
$\mathbf{B}$	= vector containing responses of the original data set
$\overline{\mathbf{B}}$	= vector containing responses of the modified data set
b_1, b_2	= regression coefficients
C	= dimensionless contact time (acetylene data example)
$c_1, c_2, c_3, c_4, \cdots$	= regression coefficients
$c'_1, \cdots, c'_4$	= regression coefficients
$\overline{c_1}, \overline{c_2}, \overline{c_3}, \overline{c_4}, \cdots$	= regression coefficients
$\mathbf{e}$	= unit basis vector
e	= component of unit basis vector
H	= dimensionless mole ratio of H_2 to n–heptane (acetylene data example)
i	= data point index
j	= data point index
k	= data point index
m	= total number of regression coefficients
P	= percentage of conversion of n–heptane to acetylene (acetylene data example)
p	= total number of data points
$\mathbf{Q}$	= vector containing regressors of a data point

* Aerodynamicist, Jacobs Technology Inc.

r	= response
r'	= alternate response
Δr	= response residual
δr	= PRESS residual of response
T	= dimensionless reactor temperature (acetylene data example)
$\mathbf{X}$	= vector containing regression coefficients of the original data set
$\overline{\mathbf{X}}$	= vector containing regression coefficients of the modified data set
x	= first regressor
$x_1, x_2, x_3, \cdots$	= regressors
$x_1^*, x_2^*, x_3^*, \cdots$	= centered regressors
y	= second regressor
z	= third regressor
δ_{ij}	= Kronecker delta
μ	= assumed constant shift of the first regressor
ξ	= assumed constant shift of the second regressor
ρ	= fitted response using the original data set
$\overline{\rho}$	= fitted response using the modified, i.e., reduced, original data set
σ_{PRESS}	= standard deviation of PRESS residuals

I. Introduction

During the past 4 years a candidate math model search algorithm for multivariate global regression analysis was developed at Ames Research Center that determines a recommended math model for the analysis of experimental data. This recommended math model is generated using the experimental data itself and some user selections. By design, the recommended math model is an optimized regression model. It fulfills a set of statistical quality metrics that statisticians traditionally use for the assessment of a regression model's predictive capabilities.

Experience showed that the algorithm has made a faster, more efficient, and more reliable regression analysis of experimental data possible. The search algorithm was originally developed at the Ames Balance Calibration Lab for the global regression analysis of wind tunnel strain–gage balance calibration data. Since 2007, however, it is also applicable to classical regression analysis problems. Figure 1 shows the connection between the classical regression analysis process and wind tunnel strain–gage balance calibration data analysis.

Figure 2 shows the basic input and output of the candidate math model search algorithm. The search algorithm uses the regressors and responses of the data set and some user selections in order to determine a recommended math model from all possible math models that could be used for the regression analysis of the data. The algorithm ultimately identifies a recommended math model for the regression analysis of the given experimental data set that is expected to have superior predictive qualities.

In classical regression analysis the final regression coefficients and a simple matrix multiplication are used to predict responses for a given set of measured regressors. In wind tunnel strain–gage balance calibration analysis, however, it is required to predict regressors (aerodynamic loads) from measured responses (electrical output of strain–gages). Therefore, an additional step is required for the analysis of wind tunnel strain–gage balance calibration data. Coefficients of the so–called "data reduction matrix" need to be derived from the original regression coefficients. Then, these coefficients are used in an iterative process in order to predict regressors from responses (see Ref. [1] for a detailed explanation of the iteration process). Strain–gage balance calibration analysis essentially combines results of the classical regression analysis of the strain–gage balance calibration data with an iteration scheme that makes the prediction of regressors (aerodynamic loads) from measured responses (electrical output of strain–gages) possible (see also Fig. 1).

In the first part of the paper key elements of the candidate math model search algorithm are revisited. Then, latest improvements of the algorithm are discussed. Finally, data from a frequently used regression analysis example is used to illustrate benefits of selecting the algorithm's recommended math model for the analysis.

II. Candidate Math Model Search Algorithm

A. General Remarks

The development of the candidate math model search algorithm originally started in 2004 in an effort to improve reliability and efficiency of the analysis of wind tunnel strain–gage balance calibration data at the Ames balance calibration lab (see Ref. [1]). At that time two key ideas were first introduced that are still a central part of the current version of the search algorithm: (i) Singular value decomposition is applied to the regressors in order to identify math models that lead to a singular solution of the global regression problem. (ii) A search metric is used in order to compare the predictive capability of different regression models. The metric is minimized during the search.

At the end of 2006 a comparison of the search algorithm's results with results of an alternate math model building technique was performed (Ref. [2]). This study revealed a major shortcoming of 2004 version of the candidate math model search algorithm. The algorithm was unable to identify math models that have near–linear dependencies between terms (also called collinearity or multicollinearity). These near–linear dependencies can greatly diminish the predictive capability of a regression model. In addition, a math model with massive near–linear dependencies can cause iteration convergence problems if it is used for the analysis of wind tunnel strain–gage balance calibration data (see also remarks in Ref. [3], p. 4). In 2007 a new threshold dependent constraint was introduced in the algorithm that addresses the shortcoming. The constraint compares the maximum of the variance inflation factors of a tested math model with a threshold in order to identify and reject math models with near–linear dependencies.

Two additional improvements of the algorithm were made in 2007 in an attempt to make the algorithm more robust. They are the result of applying additional statistical quality metrics during the search. These improvements can be summarized as follows: (i) Another threshold dependent constraint was introduced that uses the maximum of the *p*–value of the *t*–statistic of the coefficients of the regression model in order to reject math models with insignificant terms. (ii) The standard deviation of the PRESS residuals of the responses was introduced as the new search metric.

At the beginning of 2008 another refinement of the search algorithm was made. It was concluded, after reviewing ideas presented in Ref. [4], that hierarchical math models have the ability to capture a possible constant shift in a regressor. Therefore, an optional third constraint was introduced in the search algorithm that allows the user to enforce the hierarchy rule during or after the search.

Figures 3a and 3b summarizes key elements of the 2008 version of the candidate math model search algorithm. Figure 3a shows the search algorithm if the optional hierarchy constraint is enforced by rejecting non–hierarchical math models during the search. Figure 3b shows the algorithm if the optional hierarchy constraint is applied by adding missing lower order terms after the completion of the search. In the next section of the paper important elements of the 2008 version of the search algorithm are discussed in more detail.

B. Singular Value Decomposition

The search algorithm applies a numerical technique called singular value decomposition to the regressors and the user's initial function class (math term goup) selection in order to define the largest math model that will lead to a non–singular solution of the global regression problem. This feature was first introduced in the candidate math model search algorithm in 2004 (see Ref. [1] for more detail). The application of singular value decomposition to the regressors of the regression model is of great practical importance. Singular value decomposition essentially defines the upper bound of all math models that may successfully be tested during the search without causing a "software crash." Therefore, it makes the automation of the candidate math model search possible.

C. Primary Search Constraint

The test of the statistical significance of individual coefficients of a regression model is used as a constraint during the candidate math model search. This approach has the advantage that both the "dependent" variables of the regression problem, i.e., the responses, and the "independent" variables of the regression problem, i.e., the regressors, are used during the math model search in order to assess the significance of different terms of a math model (for more detail see Ref. [5], pp.84–85 or Ref. [6], pp.31–33).

The selected test of significance looks at the standard error of each regression coefficient of a math model. The standard error is an estimate of the standard deviation of the coefficient. It can be thought of as a measure of the precision with which the regression coefficient is measured. A coefficient should be included in the math model if it is large compared to its standard error.

Traditionally, the t–statistic of a coefficient is used in order to quantitatively compare a regression coefficient with its standard error. The t–statistic equals the ratio between the coefficient value and its standard error. A coefficient is probably "significant" if its t–statistic is greater than the critical value of a Student's t–distribution (see Ref. [6], p.32). This comparison can also be performed using the p–value of the coefficient. The p–value of a coefficient is determined from a comparison of the t–statistic with values in a Student's t–distribution. With a p–value of, e.g., 0.1 (or 10 %) one can say with a 90 % probability of being correct that the regression coefficient is having some effect. Therefore, the largest p–value of the coefficients of a regression model can be compared with a threshold in order to decide if the regression model is significant from a statistical point of view. Now, it is possible to formulate the primary constraint for the candidate math model search:

> **PRIMARY SEARCH CONSTRAINT (REQUIRED)**
> TEST ONLY MATH MODELS DURING THE CANDIDATE MATH MODEL SEARCH THAT HAVE
> A LARGEST P–VALUE OF A COEFFICIENT OF LESS THAN A USER SELECTED THRESHOLD.

The user selected threshold may range from a conservative value of 0.0001 to a liberal value of 0.1. A decrease of the p–value threshold, e.g., from 0.001 to 0.0001, tightens the constraint that is applied during the candidate math model search. The corresponding recommended math model will have fewer terms.

D. Secondary Search Constraint

The search algorithm uses a second threshold dependent constraint in order to reject math models during the search that contain near–linear dependencies (also called collinearity or multicollinearity). These near–linear dependencies could greatly diminish the predictive capability of a regression model and have to be avoided at all cost.

In general, the maximum of the variance inflation factor of the math model may be used as a metric in order to assess near–linear dependencies. This value is compared with a threshold in order to assess the presence of near–linear dependencies between terms of the regression model (see Ref. [5] and [6] for a detailed discussion of the variance inflation factor). The search constraint can be defined as follows:

> **SECONDARY SEARCH CONSTRAINT (REQUIRED)**
> TEST ONLY MATH MODELS DURING THE CANDIDATE MATH MODEL SEARCH THAT HAVE
> A LARGEST VARIANCE INFLATION FACTOR OF LESS THAN A USER SELECTED THRESHOLD.

A threshold of 5 or 10 for the variance inflation factor is suggested in the literature (see Ref. [5] and [6]). The maximum variance inflation factor of a math model should stay below this value in order to avoid near–linear dependencies between terms. The author's experience showed, however, that these traditionally quoted thresholds are often too conservative. In a recent publication his experience was confirmed. A threshold as large as 40 appears to be acceptable (see remarks in Ref. [7]). Therefore, the author suggests to use 5 as a conservative and 40 as a liberal threshold for the detection of near–linear dependencies in the search algorithm. A decrease of the variance inflation factor threshold from, e.g., 10 to 5, tightens the constraint that is applied during the candidate math model search. This change results in a recommended math model with fewer terms.

An interesting observation was made during the implementation of a procedure that computes variance inflation factors as input for the candidate math model search algorithm. It was noticed that at least two different calculation methods are used in commercially available statistics software packages for the determination of variance inflation factors. Consequently, the computed variance inflation factors for the same regression model term may differ if results of the two methods are compared. Table 2 below compares key elements of these two different methods for the calculation of variance inflation factors.

A detailed analysis showed that the two methods differ in steps that are used to generate the non–linear (i.e., squared, cubed, ...) terms for the calculation of the correlation matrix (compare STEP 1 and STEP 2 in Table 1). *Method 1* first centers the linear terms of the regression model (STEP 1) and then computes the non–linear terms from the centered linear terms (STEP 2). *Method 2*, on the other hand, directly computes the non–linear terms from the original linear terms (STEP 2).

Table 1: Comparison of two methods for the calculation of variance inflation factors.

	METHOD 1 (e.g., DESIGN–EXPERT[†])	METHOD 2 (e.g., SAS/STAT)
STEP 1	CENTER ALL LINEAR TERMS OF REGRESSION MODEL BY SUBTRACTING THE ARITHMETIC MEAN OF MINIMUM AND MAXIMUM OF THE SELECTED LINEAR TERM FROM EACH VALUE	– NOT APPLICABLE –
STEP 2	GENERATE NON–LINEAR (SQUARED, CUBED, ...) TERMS FROM THE CENTERED LINEAR TERMS	GENERATE NON–LINEAR (SQUARED, CUBED, ...) TERMS FROM THE ORIGINAL LINEAR TERMS
STEP 3	CENTER ALL LINEAR AND NON–LINEAR TERMS OF THE REGRESSION MODEL BY SUBTRACTING THE ARITHMETIC MEAN OF THE CORRESPONDING LINEAR OR NON–LINEAR TERM FROM EACH VALUE	
STEP 4	APPLY UNIT LENGTH SCALING, I.E., DIVIDE EACH CENTERED LINEAR OR NON–LINEAR TERM BY THE SQUARE ROOT OF THE SUM OF SQUARES OF THE CENTERED LINEAR OR NON–LINEAR TERM	
STEP 5	COMPUTE THE CORRELATION MATRIX USING THE CENTERED & SCALED TERMS OF STEP 4	
STEP 6	COMPUTE THE INVERSE OF THE CORRELATION MATRIX	
STEP 7	VARIANCE INFLATION FACTORS ARE THE DIAGONAL ELEMENTS OF THE INVERSE MATRIX	

[†] The DESIGN–EXPERT software uses regressors in "coded units", i.e, the linear terms of STEP 1 should theoretically be (a) centered and (b) scaled before the non–linear (i.e., squared, cubed, ...) terms are generated. In STEP 4, however, unit length scaling is applied to all centered linear and non–linear terms. Therefore, scaling can be omitted in STEP 1 because the centered & scaled linear and non–linear terms used in STEP 5 are independent of any scaling that is applied in STEP 1.

It is a surprising fact, superficially viewed, that *Method 1* has to center the non–linear terms of the regression model in STEP 3 even though these terms were obtained from the centered linear terms of STEP 1. The additional centering must be performed because the product of two centered regressors in not necessarily centered. This assertion may easily be proven using a simple example. It is assumed that two regressors, e.g., x_1 and x_2, are given. In addition, it is assumed that the data set consist of three data points. The first and second regressor could have the following values written in vector format:

$$1st\ regressor: \quad x_1(1),\ x_1(2),\ x_1(3) \quad \Longrightarrow \quad 1,\ 0,\ 0$$

$$2nd\ regressor: \quad x_2(1),\ x_2(2),\ x_2(3) \quad \Longrightarrow \quad -1,\ 0,\ +1$$

The regressors are centered by subtracting the arithmetic mean of the minimum and maximum of each regressor from each regressor value. Then, we get:

$$Centered\ 1st\ regressor: \quad x_1^*(1),\ x_1^*(2),\ x_1^*(3) \quad \Longrightarrow \quad 0.5,\ -0.5,\ -0.5$$

$$Centered\ 2nd\ regressor: \quad x_2^*(1),\ x_2^*(2),\ x_2^*(3) \quad \Longrightarrow \quad -1,\ 0,\ +1$$

In the next step, the product of the two centered regressors is computed. Then, we get:

$$3rd\ regressor: \quad x_3(1) = x_1^*(1) \cdot x_2^*(1),\ x_3(2) = x_1^*(2) \cdot x_2^*(2),\ x_3(3) = x_1^*(3) \cdot x_2^*(3) \quad \Longrightarrow \quad -0.5,\ 0,\ -0.5$$

We see that the product of the centered regressors is no longer centered. The corresponding centered product of the regressors equals the following value:

$$Centered\ 3rd\ regressor: \quad x_3^*(1),\ x_3^*(2),\ x_3^*(3) \quad \Longrightarrow \quad -0.25,\ 0.25,\ -0.25$$

A natural question emerges: Which variance inflation factor calculation method, i.e., *Method 1* or *Method 2*, should provide the input for the secondary search constraint? The author recommends a compromise. He suggests (i) to compute the maximum of the variance inflation factor of the regression model for each calculation method and (ii) to use the greater of the two maxima as input for the secondary search constraint. This approach guarantees that the candidate math model search algorithm will always generate math models that fulfill the chosen variance inflation factor constraint for both calculation methods. In addition, the compromise favors smaller math models.

E. Regression Model Search Metric

The candidate math model search algorithm needs a metric that can be used for the comparison of the expected predictive capability of different regression models. This search metric is minimized during the search. Initially, the standard deviation of the response residuals was used for this purpose (see Ref. [1]). In 2007, however, the author identified a better metric for the assessment of the predictive capability of different regression models. It is the standard deviation of the PRESS residuals of a math model (see Refs. [3], [5], and [6] for a discussion of PRESS residuals). This metric is used in the current version of the search algorithm.

For a successful and efficient implementation of the calculation of PRESS residuals it is helpful to have access to a rigorous derivation of the PRESS residuals. Unfortunately, derivations presented in the literature are often abbreviated and have gaps (see, e.g., the derivation for the PRESS statistic given in Ref. [5], pp. 562–564). Therefore, the present author decided to develop a very detailed alternate derivation of the PRESS residuals of the responses. This derivation is given in App. 1 of the present paper. In addition, a new derivation of corresponding PRESS residuals of the regressors is given in App. 4. This type of PRESS residual is of interest during the regression analysis of wind tunnel strain–gage balance calibration data.

F. Math Term Hierarchy Constraint

Sometimes it is important to ensure that the regression model of experimental data is able to represent a constant shift in the regressors. Only hierarchical math models can be used for this purpose. Therefore, an optional math term combination constraint was added to the search algorithm. This third constraint enforces the hierarchy rule during the search or adds missing lower order terms to the recommmended math model after the search. The constraint can be described as follows:

MATH TERM HIERARCHY CONSTRAINT (OPTIONAL)
TEST ONLY "HIERARCHICAL" MATH MODELS DURING THE MATH MODEL SEARCH -OR- ADD MISSING LOWER ORDER TERMS TO THE RECOMMENDED MATH MODEL AFTER THE SEARCH.

What is the difference between a hierarchical and a non–hierarchical regression model? A hierarchical math model allows for regressors to have a theoretical off–set in the regressor value. This off–set must not be confused with a bias error. A simple example can be constructed in order to illustrate the application of a hierarchical math model in regression analysis. It is assumed that a single regressor and a single response variable are used to approximate an experimental data set. The regressor variable is called x and the response variable is called r. Furthermore, it is assumed that the given data behaves approximately like a second order polynomial. Figure 4 shows the regressors and responses of the example. Now, ignoring lower order terms intentionally, a non–hierarchical math model of the responses may have the form:

$$r \quad = \quad a_1 \; + \; a_2 \cdot x^2 \tag{1}$$

where a_1 and a_2 are the regression coefficients. It is also assumed that the regressor variable x was shifted by some unknown amount. In that case, the regression model given in Eq. (1) cannot represent the shift μ of the regressor. An alternate regression model needs to be introduced that has the following form:

$$r' \quad = \quad b_1 \; + \; b_2 \cdot (x - \mu)^2 \tag{2a}$$

where b_1 and b_2 are the new regression coefficients. Equation (2a) can also be written as:

$$r' \quad = \quad \left(b_1 \; + \; b_2 \mu^2\right) \; + \; (-2 b_2 \mu) \cdot x \; + \; (b_2) \cdot x^2 \tag{2b}$$

A new set of regression coefficients can now be introduced. We can write:

$$c_1 \quad = \quad b_1 \; + \; b_2 \mu^2 \tag{2c}$$

$$c_2 \quad = \quad -2b_2\mu \tag{2d}$$

$$c_3 \quad = \quad b_2 \tag{2e}$$

Then, Eq. (2*b*) becomes:

$$r' \quad = \quad c_1 \;+\; c_2 \cdot x \;+\; c_3 \cdot x^2 \tag{3}$$

Equation (1) is a non–hierarchical math model as the lower order term x is missing. Equation (3) is a hierarchical math model as all lower order terms of the regression model are present. In addition, we know that the regression model given in Eq. (3) can represent the constant shift in the regressor. Therefore, the hierarchical regression model r' is more flexible than the non–hierarchical math model r.

A regression model is hierarchical if all lower order terms of each term are present. A systematic approach is needed for the identification of the lower order terms. One approach shifts each regressor of the term by a constant amount and afterwards expands all brackets. An example can illustrate this approach. It is assumed that the term x^2y is selected for a regression model. Now, the first regressor x is shifted by a constant μ and the second regressor y is shifted by a constant ξ. Then, we get the expression:

$$x^2 y \quad \Longrightarrow \quad (x-\mu)^2 \cdot (y-\xi)$$

In the next step, after expanding the brackets of the above term, we get:

$$(x-\mu)^2 \cdot (y-\xi) \quad = \quad (-\mu^2\xi) \;+\; (2\mu\xi)\cdot x \;+\; (-\xi)\cdot x^2 \;+\; (\mu^2)\cdot y \;+\; (-2\mu)\cdot xy \;+\; x^2 y \tag{4}$$

After analyzing the right hand side of Eq. (4), we see that five terms are needed to satisfy hierarchy:

$$\textit{hierarchy rule} \quad \Longrightarrow \quad x,\;\; x^2,\;\; y,\;\; xy,\;\; x^2y$$

Finally, we conclude that the lower order terms of x^2y are given by the following four terms:

$$x^2y \quad \Longrightarrow \quad \textit{lower order terms} \quad \Longrightarrow \quad x,\;\; x^2,\;\; y,\;\; xy$$

All these terms are needed in the regression model so that x^2y can model a constant shift in the regressors.

A list of lower order terms of any regression model term can easily be developed using the approach described above. Table 2 below lists lower order terms of typical terms that are used in regression models.

Table 2: Lower order terms of typical regression model terms.

	MATH TERM	LIST OF LOWER ORDER TERMS
1	x	–
2	x^2	x
3	x^3	$x,\; x^2$
4	x^4	$x,\; x^2,\; x^3$
5	x^5	$x,\; x^2,\; x^3,\; x^4$
6	xy	$x,\; y$
7	x^2y	$x,\; x^2,\; y,\; xy$
8	x^3y	$x,\; x^2,\; x^3,\; y,\; xy,\; x^2y$
9	xyz	$x,\; y,\; z,\; xy,\; xz,\; yz$
10	$\lvert x\rvert$	–
11	$\lvert x\rvert^3 \;=\; \lvert x\rvert \cdot \lvert x\rvert \cdot \lvert x\rvert$	$\lvert x\rvert,\;\; \lvert x\rvert \cdot \lvert x\rvert = x^2$
12	$x \cdot \lvert x\rvert$	$x,\; \lvert x\rvert$
13	$\lvert x \cdot y\rvert \;=\; \lvert x\rvert \cdot \lvert y\rvert$	$\lvert x\rvert,\; \lvert y\rvert$
14	$x \cdot \lvert y\rvert$	$x,\; \lvert y\rvert$
15	$\lvert x\rvert \cdot y$	$\lvert x\rvert,\; y$
16	$(1/x)$	–
17	$(1/x)^2$	$(1/x)$
18	$(1/x)^3$	$(1/x),\;\; (1/x)^2$
19	$(1/x)^4$	$(1/x),\;\; (1/x)^2,\;\; (1/x)^3$

A question remains: Should the hierarchy rule always be applied to a regression model if terms of the model cannot directly be derived from first principles of physics or other discipline knowledge? Different answers are given to this question in the literature. *Peixoto*, e.g., recommends *"... as a general rule, polynomial regression models should be hierarchically well formulated, especially in applications where the origin of the predictor variables is arbitrary or irrelevant. ..."* (from Ref. [4], p. 29). *Montgomery, Peck,* and *Vining*, on the other hand, say *"... We have mixed feelings about this [i.e., hierarchy] as a hard–and–fast rule. ... The best advice is to fit a model that has all terms significant and to use discipline knowledge rather than an arbitrary rule as an additional guide in model formulation. ..."* (from Ref. [5], p. 204).

The present author's opinion is closer to the viewpoint of the latter group of authors. Let us assume, for example, that (i) it is unknown if a given experimental data set requires a hierarchical or non–hierarchical regression model and that (ii) the hierarchy rule is blindly applied to this data set. In this case it is important to remember that the hierarchy rule does not take any information into account that is contained in the experimental data itself. The earlier introduced primary and secondary search constraints and the search metric, on the other hand, use information that is contained in the regressor and response values. Figure 4 is also a reminder that applying a hierarchical regression model to data that in reality needs a non–hierarchical model could be as damaging as applying a non–hierarchical regression model to data that needs a hierarchical model. Therefore, the author suggests to ignore the hierarchy rule during the search if the hierarchy characteristics of the experimental data set are unknown. This approach will allow the candidate math model search algorithm to identify the recommended math model of the data set by using only the minimization of the search metric in combination with the primary and secondary search constraint. This recommended math model may be hierarchical or non–hierarchical depending on the result of the minimization of the algorithm's search metric.

The regression analysis of some experimental data could occasionally require a hierarchical regression model. Therefore, the hierarchy rule has been implemented as an optional third constraint in the search algorithm. A user may select one of two approaches to enforce the hierarchy rule in the search algorithm. The first approach simply rejects non–hierarchical math models during the search (see Fig. 3a). The second approach adds missing lower order terms to the recommended model after the completion of the search (see Fig. 3b). Consequently, the second approach cannot guarantee that the primary and secondary search constraints are still fulfilled as the recommended math model is modified after the completion of the initial search.

III. Discussion of Example

A software application called BALFIT was written that implements the current version of the candidate math model search algorithm (see Ref. [8] for the description of an older version of BALFIT). The software supports multivariate classical regression analysis as well as wind tunnel strain–gage balance calibration analysis applications.

In the past BALFIT was applied to a wide variety of global regression problems that are related to the analysis of wind tunnel strain–gage balance calibration data (see Refs. [9], [10], and [11]). For the present paper it was decided to use a more general regression analysis example that is taken from a textbook. This example has the advantage that it consists only of a few data points. In addition, the data can easily be processed if a reader, for example, would like to verify some of the results that are reported in the present paper.

The selected data set is the acetylene data example that is discussed in Ref. [5] (see pp. 329–335). The data table of this example and the original regression model are depicted in Fig. 5a of the present paper. The response equals the percentage of conversion of n–heptane to acetylene (symbol P). The regressors are three dimensionless quantities that are derived from the reactor temperature, the mole ratio of H_2 to n–heptane, and the contact time (symbols T, H, and C).

In the first step of the present investigations the original regression model was selected to fit the data in the least squares sense. This regression model is given as:

$$\begin{aligned} P \;=\;& c_1 + c_2 \cdot T + c_3 \cdot H + c_4 \cdot C + c_5 \cdot T \cdot H + c_6 \cdot T \cdot C \\ & + c_7 \cdot H \cdot C + c_8 \cdot T^2 + c_9 \cdot H^2 + c_{10} \cdot C^2 \end{aligned} \tag{5a}$$

Three different software applications were selected for the analysis of the dimensionless acetylene data example that the data table in Fig. 5a defines: (i) DESIGN–EXPERT 7.0 (Stat–Ease, Inc., Minneapolis, Minnesota), (ii) SAS/STAT (SAS Institute Inc., Cary, North Carolina), and (iii) BALFIT (Ames Research Center). Regression coefficients and analysis of variance results were obtained for the data after the completion of the regression analysis. Figure 5b shows results for DESIGN–EXPERT, Fig. 5c shows results for SAS/STAT, and Fig. 5d shows results for BALFIT. All results show excellent agreement if the coefficients and other statistical metrics are compared.

The comparison of the three solutions also confirms that different approaches are used for the calculation of the variance inflation factor in DESIGN–EXPERT and SAS/STAT as the calculated variance inflation factors do not match. DESIGN–EXPERT appears to use *Method 1* and SAS/STAT appears to use *Method 2* (see again discussion of secondary search constraint above). BALFIT, on the other hand, has the ability to compute variance inflation factors for both calculation methods. These two sets of variance inflation factors show excellent agreement with the corresponding values that DESIGN–EXPERT and SAS/STAT report. At this point it is important to mention that both sets of variance inflation factors confirm that the selected original regression model of the data has unwanted near–linear dependencies.

In the next step BALFIT's candidate math model search algorithm was applied to the data set. The threshold for the primary search constraint was selected to be 0.0010. Therefore, a regression model is only considered during the search if the largest p–value of its coefficients is below 0.0010. The threshold for the secondary search constraint was selected to be 10. Consequently, near–linear dependencies in a regression model are assumed to be negligible if the largest variance inflation factor of the model is less than 10. It was also decided to ignore the hierarchy rule during the search. The regression model defined in Eq. (5a) was selected as the upper bound for the math model search. As a result, the algorithm compared a total of 45 math models during the search. Figures 6a, 6b, and 6c show some results of the candidate math model search. Figure 6a depicts the search metric, i.e., the standard deviation of the PRESS residuals of the responses, as a function of the number of candidate math model terms. The search metric initially decreases, reaches a local minimum, and increases afterwards as more and more terms are added to the candidate math model. Figure 6b shows the primary search constraint values, i.e, the maximum of the p–value of the t–statistic of the terms of each candidate math model, as a function of the number of candidate math model terms. The values increase as more and more terms are being added during the math model search. In other words – more and more statistically insignificant terms are added to the candidate math models as the search progresses. Figure 6c shows the secondary search constraint, i.e., the maximum of the variance inflation factors of each candidate math model, as a function of the number of candidate math model terms. Near–linear dependencies seem to emerge in the candidate math models as soon as the regression model has more than 6 terms.

Figure 7a shows coefficient values and analysis of variance results for the recommended math model that the search algorithm identified for the data set. The recommended math model is described by the following expression:

$$P \;=\; c_1' \;+\; c_2' \cdot T \;+\; c_3' \cdot H \;+\; c_4' \cdot T \cdot H \tag{5b}$$

It is a surprising result of the search that the third regressor, i.e., the dimensionless contact time C, is not contained anymore in the recommended math model. This result can be investigated in more detail by looking at the sequence of terms that were added from candidate math model to candidate math model. The following sequence was observed:

$$\textit{Intercept} \;\Longrightarrow\; T \;\Longrightarrow\; T \cdot H \;\Longrightarrow\; H \;\Longrightarrow\; T^2 \;\Longrightarrow\; H^2 \;\Longrightarrow\; H \cdot C \;\Longrightarrow\; C \;\Longrightarrow\; C^2 \;\Longrightarrow\; T \cdot C$$

The regressor C is first added from the candidate math model with six terms to the candidate math model with seven terms. This is also exactly the point when massive near–linear dependencies first appear in the candidate math models (see Fig. 6c). The regressor C seems to be responsible for the near–linear dependencies that were detected in the original regression model. Therefore, the final result of the candidate math model search can be interpreted in two ways: (1) the data set of the textbook example does not contain enough information that would support the dimensionless contact time C as an independent variable -or- (2) the dimensionless contact time C is not truely an independent variable of the physical phenomenon that is being described in the experiment.

Finally, it is interesting to compare the regression coefficients, the analysis of variance results, the response residuals, and the PRESS residuals for the recommended math model (Eq. (5b)) and the original math model (Eq. (5a)). Figures 7a, 7b, and 7c show results for the recommended math model. Figures 8a, 8b, and 8c show results for the original math model. The unit [%] of the residuals depicted in Figs. 7b, 7c, 8b, and 8c is the original unit of the response, i.e., the percentage of conversion from n–heptane to acetylene. Comparing the regression coefficients, analysis of variance results, and the variance inflation factors, i.e., Fig. 7a with Fig. 8a, we see that the original math model uses many insignificant terms and has unwanted near–linear dependencies. Nevertheless, the response residuals of the original math model (Fig. 8b) are lower than the response residuals of the recommended math model (Fig. 7b). This observation is not a surprise as the original math model uses 10 terms and the recommended math model uses only 4 terms for the fit. The situation is completely reversed if the PRESS residuals of the original and recommended math model are compared. This time the PRESS residuals of the recommended math model (Fig. 7c) are lower than the PRESS residuals of the original math model (Fig. 8c). Therefore, it is concluded that the original math model is overfitting the experimental data. It should not be used to model the experimental data set in the least squares sense.

IV. Summary and Conclusions

Significant improvements of a candidate math model search algorithm were made that was developed at Ames Research Center for the automated multivariate regression analysis of experimental data. The algorithm determines a recommended math model for the regression analysis that meets strict statistical quality requirements. Terms of the recommended math model are selected by comparing the predictive capability of different regression models during the search using the standard deviation of the PRESS residuals of the responses as a metric. In addition, two threshold dependent constraints are applied during the search. The first constraint makes sure that only math models with significant terms are tested during the search. The second constraint uses the variance inflation factors of a tested regression model in order to reject regression models with unwanted near–linear dependencies during the search. An optional third constraint can also be selected that enforces the hierarchy rule. Data from a frequently used textbook example was chosen in order to illustrate key elements of the candidate math model search algorithm.

Studies presented in the current paper showed that at least two different methods are used in existing commercially available statistics software packages for the determination of variance inflation factors of a regression model. These methods may lead to different numerical values even though the regressor values and math terms selection are identical. Consequently, the exact reporting of variance inflation factors of a regression model requires a description of the method that is used to determine the variance inflation factors. At least, the name of the software application should be listed that was used to determine variance inflation factors for a given regression model.

The description of the variance inflation factor calculation method is less important whenever the variance inflation factor maximum of a regression model is large compared to the threshold that is selected for the detection of near–linear dependencies between regression model terms. This threshold may range from 5 to 40. Sometimes, however, it happens that the variance inflation factor maximum of a regression model is close to the threshold. In that situation the exact reporting of the variance inflation factor calculation method is critical as the difference between the variance inflation factors of different calculation methods may be on the order of the threshold itself.

A detailed study of the hierarchy rule was also presented in the current paper. This analysis concluded that the hierarchy rule should not be applied blindly to a regression model if the model's hierarchy characteristics are not known from first principles of physics or other discipline knowledge. Applying a hierarchical model to data that needs a non–hierachical model can be as damaging as applying a non–hierarchical model to data that needs a hierarchical model. Therefore, the hierarchy rule has only been implemented as an optional third constraint in the candidate math model search algorithm.

V. Acknowledgements

The work reported in this paper was partially supported by NASA's Aeronautic Test Program and the

Wind Tunnel Operations Division at Ames Research Center under contract NNA04BA85C. The author would like to thank Richard DeLoach of NASA Langley Research Center and the technical staff of SAS/STAT for generously providing analysis of variance results and variance inflation factors for the acetylene data example. Thanks also go to Tom Volden of Jacobs Technology for his careful proofreading and checking of the final manuscript.

VI. References

[1] Ulbrich, N. and Volden, T., "Strain–Gage Balance Calibration Analysis Using Automatically Selected Math Models," AIAA 2005–4084, paper presented at the 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, Arizona, July 2005.

[2] DeLoach, R. and Ulbrich, N., "A Comparison of Two Balance Calibration Model Building Methods," AIAA 2007–0147, paper presented at the 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2007.

[3] Ulbrich, N. and Volden, T., "Regression Analysis of Experimental Data Using an Improved Math Model Search Algorithm," AIAA 2008–0833, paper presented at the 46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2008.

[4] Peixoto, J. L., "A Property of Well–Formulated Polynomial Regression Models," *The American Statistician*, Vol. 44, No. 1, pp. 26–30, February 1990.

[5] Montgomery, D. C., Peck, E. A., and Vining, G. G., *Introduction to Linear Regression Analysis*, 4th ed., John Wiley & Sons, Inc., New York, 2006, pp. 84–85, pp. 141–142, pp. 323–341, pp. 562–564.

[6] Myers, R. H. and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 1st ed., John Wiley & Sons, Inc., New York, 1995, pp. 31–33, pp. 45–47, pp. 656–666.

[7] O'Brien, R. M., "A Caution Regarding Rules of Thumb for Variance Inflation Factors," *Quality and Quantity*, Vol. 41, Number 5, Springer Netherlands, October 2007, pp. 673–690.

[8] Ulbrich, N. and Volden, T., "Development of a New Software Tool for Balance Calibration Analysis," AIAA 2006–3434, paper presented at the 25th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[9] Ulbrich, N. and Volden, T., "Application of a New Calibration Analysis Process to the MK–III–C Balance," AIAA 2006–0517, paper presented at the 44th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2006.

[10] Ulbrich, N. and Volden, T., "Analysis of Floor Balance Calibration Data using Automatically Generated Math Models," AIAA 2006–3437, paper presented at the 25th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[11] Ulbrich, N. and Volden, T., "Analysis of Balance Calibration Machine Data using Automatically Generated Math Models," AIAA 2007–0145, paper presented at the 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2007.

[12] Burden, R. L. and Faires, J. D., "Numerical Analysis," 3rd edition, PWS–Kent Publishing Company, Boston, Massachusetts, 1985, p. 334.

Appendix 1: Derivation of the PRESS Residual of a Response

1.1 General Remarks

PRESS residuals of the solution of a linear global regression problem for a given set of experiments data are important metrics that may be used for comparison and assessment of the predictive capabilities of different regression models. In principle, the calculation of PRESS residuals of a given data point requires only a few steps. At first, the data point is temporarily removed from the original data set. Then, the regression model is fitted using the remaining data points. In the next step, the regression model is used to predict the response of the withheld data point. Finally, the difference between the predicted and measured response of the withheld data point is computed. This difference is the PRESS residual of the response of the data point. A very detailed alternate derivation of the PRESS residuals of the responses is given in this appendix. In addition, it is shown that the calculation of the PRESS residuals of the responses only requires a single global regression of the original data set.

The present alternate derivation of the PRESS residual is complex. In principle, it follows a derivation that is outlined in Ref. [5]. However, many intermediate steps and missing proofs of important auxiliary relationships were added to the present alternate derivation of the PRESS residuals. These changes and additions have made the derivation more complete and easier to understand.

It was decided to use a nomenclature for the derivation that makes a clear distinction between scalars, vectors, and matrices. Therefore, symbols representing a scalar are printed as plain text and symbols representing a vector or a matrix are printed using boldface. Occasionally, subscripts are added to a symbol that indicate the number of rows and columns of a vector or a matrix.

The derivation will show that a close connection between the classical response residual and the PRESS residual of a response exists. Therefore, the derivation has two parts. At first, an equation for the classical response residual is derived. Afterwards, it is shown how the equation of the response residual is connected to the equation of the PRESS residual.

1.2 Response Residual

The derivation of the classical response residual starts by defining the regression model of a set of responses. The regression model is given by the following equation:

REGRESSION MODEL OF A RESPONSE

$$r(k) \quad = \quad \underbrace{c_1 \; + \; c_2 \cdot x_1(k) \; + \; c_3 \cdot x_2(k) \; + \; c_4 \cdot x_3(k) \; + \; \cdots}_{total\ number\ of\ m\ regression\ coefficients\ (c_1,\ c_2,\ \cdots,\ c_m)} \quad ; \qquad 1 \leq k \leq p \qquad (6)$$

The regression model of each point of a given experimental data set can be written in a more compact format using a matrix and two vectors. Then, we get the following equation:

$$\underbrace{\begin{bmatrix} 1 & x_1(1) & x_2(1) & \cdots \\ 1 & x_1(2) & x_2(2) & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_1(k) & x_2(k) & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_1(p) & x_2(p) & \cdots \end{bmatrix}}_{\mathbf{A}_{p \times m}} \cdot \underbrace{\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_m \end{bmatrix}}_{\mathbf{X}_{m \times 1}} = \underbrace{\begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(k) \\ \vdots \\ r(p) \end{bmatrix}}_{\mathbf{B}_{p \times 1}} \qquad (7)$$

Equation (7) may be written in abbreviated form using the matrix and vector symbols that were introduced above. Then, the global regression problem of an experimental data set, i.e., of the "original data set", can be described as follows:

GLOBAL REGRESSION PROBLEM
(original data set)

$$\mathbf{A}_{p\times m} \cdot \mathbf{X}_{m\times 1} = \mathbf{B}_{p\times 1} \qquad (8a)$$

Equation (8a) is the classical formulation of a linear global regression problem. The solution of the regression problem is given by the so–called "normal equations." They can be written as:

SOLUTION OF GLOBAL REGRESSION PROBLEM
(original data set)

$$\mathbf{X} = \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{B}\right) \qquad (8b)$$

It is convenient to define a column vector **Q** that can be used to write matrix **A** and its transpose $\mathbf{A}^{\mathbf{T}}$ in a more useful format. This column vector and its transpose can be expressed as follows:

$$\mathbf{Q}(k)_{m\times 1} = \begin{bmatrix} 1 \\ x_1(k) \\ x_2(k) \\ \vdots \end{bmatrix} \qquad (9a)$$

$$\mathbf{Q}^{\mathbf{T}}(k)_{1\times m} = [1 \;\; x_1(k) \;\; x_2(k) \;\; \cdots] \qquad (9b)$$

Now, after comparing Eq. (7) with the definition of vectors **Q** and $\mathbf{Q}^{\mathbf{T}}$, we see that matrix **A** and its transpose $\mathbf{A}^{\mathbf{T}}$ can be expressed as follows:

$$\mathbf{A}_{p\times m} = \begin{bmatrix} \mathbf{Q}^{\mathbf{T}}(1) \\ \mathbf{Q}^{\mathbf{T}}(2) \\ \vdots \\ \mathbf{Q}^{\mathbf{T}}(k) \\ \vdots \\ \mathbf{Q}^{\mathbf{T}}(p) \end{bmatrix} \qquad (10a)$$

$$\mathbf{A}^{\mathbf{T}}{}_{m\times p} = [\mathbf{Q}(1) \;\; \mathbf{Q}(2) \;\; \cdots \;\; \mathbf{Q}(k) \;\; \cdots \;\; \mathbf{Q}(p)] \qquad (10b)$$

Finally, the response residual of a data point with index k can be expressed as the difference between the original response $r(k)$ and the fitted response $\rho(k)$. We get:

RESPONSE RESIDUAL

$$\Delta r(k) = r(k) - \rho(k) \qquad (11a)$$

$$\rho(k) = \mathbf{Q}^{\mathbf{T}}(k) \cdot \mathbf{X} \qquad (11b)$$

The calculation of response residuals of all data points requires a single matrix inversion as **X**, i.e., the solution of the global regression problem given in Eq. (8b), depends only on the product of inverse matrix $\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-\mathbf{1}}$ with vector $\mathbf{A}^{\mathbf{T}}\mathbf{B}$.

1.3 PRESS Residual of Response

The PRESS residual of a data point is closely related to its response residual. Only one key difference exists. The PRESS residual of a data point has to be computed using the global regression solution of the modified original experimental data set that has one fewer data point (i.e., the data point itself). Therefore, vectors and matrices of the global regression problem of the original experimental data set change. They become a function of the omitted data point.

It is assumed that the omitted data point has the index k. Then, the global regression problem of the modified, i.e., reduced, original data set can be described using the following equation:

GLOBAL REGRESSION PROBLEM
(modified, i.e., reduced, original data set)

$$\overline{\mathbf{A}}(k)_{(p-1)\times m} \cdot \overline{\mathbf{X}}(k)_{m\times 1} = \overline{\mathbf{B}}(k)_{(p-1)\times 1} \qquad (12a)$$

where

$$\mathbf{A}_{p\times m} \Longrightarrow \overline{\mathbf{A}}(k)_{(p-1)\times m} = \begin{bmatrix} \mathbf{Q^T}(1) \\ \mathbf{Q^T}(2) \\ \vdots \\ \mathbf{Q^T}(k-1) \\ \mathbf{Q^T}(k+1) \\ \vdots \\ \mathbf{Q^T}(p) \end{bmatrix} \qquad (12b)$$

$$\mathbf{X}_{m\times 1} \Longrightarrow \overline{\mathbf{X}}(k)_{m\times 1} = \begin{bmatrix} \overline{c_1} \\ \overline{c_2} \\ \overline{c_3} \\ \vdots \\ \overline{c_m} \end{bmatrix} \qquad (12c)$$

$$\mathbf{B}_{p\times 1} \Longrightarrow \overline{\mathbf{B}}(k)_{(p-1)\times 1} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(k-1) \\ r(k+1) \\ \vdots \\ r(p) \end{bmatrix} \qquad (12d)$$

The solution of the global regression problem of the modified data set can easily be written down using the normal equations. In this case, we get:

SOLUTION OF GLOBAL REGRESSION PROBLEM
(modified, i.e., reduced, original data set)

$$\overline{\mathbf{X}}(k) = \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right) \qquad (13)$$

The PRESS residual of a response can be computed using an equation that is very similar to the equation of the classical response residual. Again, the PRESS residual of a response of a data point with index k

can be expressed as the difference between the original response $r(k)$ and the fitted response $\overline{p}(k)$ that was computed using the regression solution of the modified original data set. We get:

PRESS RESIDUAL OF A RESPONSE

$$\delta r(k) \;=\; r(k) \;-\; \overline{p}(k) \qquad (14a)$$

$$\overline{p}(k) \;=\; \mathbf{Q^T}(k) \;\cdot\; \overline{\mathbf{X}}(k) \qquad (14b)$$

It is interesting to mention that two metrics may be derived from the PRESS residuals of a given experimental data set. These metrics may be used to assess the predictive capability of different regression models. The first metric is called the PRESS statistic. It is defined by the following relationship:

$$PRESS \;=\; \sum_{k=1}^{p} \left[\delta r(k)\right]^2 \qquad (15)$$

The PRESS statistic is recommended in the literature for the comparison of different regression models (see Ref. [5], p. 142). A related metric is the standard deviation of the PRESS residuals of all data points. It may also be used to compare regression models. This metric is defined as:

$$\sigma_{PRESS} \;=\; \sqrt{\frac{1}{p\;-1}\;\sum_{k=1}^{p}\left[\delta r(k)\right]^2} \qquad (16)$$

The standard deviation of the PRESS residuals is the search metric of the regression model optimization algorithm (i.e., the candidate math model search algorithm) that is discussed in the main body of the text.

It seems, superficially viewed, that the calculation of the PRESS residuals of all data points of a data set is time consuming as a matrix inversion, i.e.,

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right)^{-1}$$

has to be performed for each data point with index k of the data set (see also the right hand side of Eq. (13)). However, it can be shown that the PRESS residual of a data point can directly be computed using only (1) the original response residual, (2) the inverse matrix used in the global regression solution of the original data set, and (3) two matrix multiplications. The proof of this surprising fact takes advantage of an important relationship between the inverse matrix used in Eq. (8b) and the inverse matrix used in Eq. (13). The relationship is derived in App. 2 of the present paper. It is given by the equation:

RELATIONSHIP BETWEEN INVERSE MATRICES
(see App. 2 for a derivation of the relationship)

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right)^{-1} \;=\; \left(\mathbf{A^T A}\right)^{-1} \;+\; \frac{\left(\mathbf{A^T A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q^T}(k) \cdot \left(\mathbf{A^T A}\right)^{-1}}{1 \;-\; h(k)} \qquad (17a)$$

$$h(k) \;=\; \mathbf{Q^T}(k) \cdot \left(\mathbf{A^T A}\right)^{-1} \cdot \mathbf{Q}(k) \qquad (17b)$$

Consequently, after inserting Eq. (17a) into Eq. (13), i.e., the regression solution of the modified data set, and after inserting the result into Eq. (14b), i.e., the equation of the fitted response $\overline{p}(k)$, we get the following expression:

$$\bar{\rho}(k) \;=\; \mathbf{Q}^{\mathbf{T}}(k) \cdot \left[\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} + \frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}}{1 - h(k)} \right] \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right) \qquad (18)$$

The brackets on the right hand side of Eq. (18) can be expanded. Then, after rearranging the result, we get for the fitted response the following equation:

$$\begin{aligned} \bar{\rho}(k) \;=\; & \frac{[\,1 - h(k)\,] \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \\ & + \frac{\left[\mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k)\right] \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \end{aligned} \qquad (19a)$$

Now, after using Eq. (17b) to simplify the numerator of the second fraction on the right hand side of Eq. (19a), we get:

$$\begin{aligned} \bar{\rho}(k) \;=\; & \frac{[\,1 - h(k)\,] \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \\ & + \frac{h(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \end{aligned} \qquad (19b)$$

Equation (19b) can be simplified further. After some algebra we can write:

$$\begin{aligned} \bar{\rho}(k) \;=\; & \frac{\mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} - \frac{h(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \\ & + \frac{h(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \end{aligned} \qquad (19c)$$

Then, simplifying the right hand side of Eq. (19c), we get:

$$\bar{\rho}(k) \;=\; \frac{\mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \qquad (19d)$$

In the next step, after using Eq. (19d) to replace the fitted response on the right hand side of Eq. (14a), we get for the PRESS residual of the response the following expression:

$$\begin{aligned} \delta r(k) \;&=\; r(k) - \frac{\mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \\ &=\; \frac{r(k) - h(k) \cdot r(k) - \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \end{aligned} \qquad (20)$$

Now, after using Eq. (17b) in order to replace $h(k)$ in the numerator of Eq. (20), we get for the PRESS residual of the response:

$$\delta r(k) \;=\; \frac{r(k) - \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot r(k) - \mathbf{Q}^{\mathbf{T}}(k) \cdot \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} \cdot \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right)}{1 - h(k)} \qquad (21)$$

Equation (21) can be simplified significantly. From Eq. (7) we know that

$$\mathbf{B}_{p\times 1} \quad = \quad \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(k) \\ \vdots \\ r(p) \end{bmatrix} \tag{22}$$

Then, after combining Eq. (10b) with (22), we can write the product $\mathbf{A^T B}$ in the following form:

$$\mathbf{A^T}\,\mathbf{B} \quad = \quad \sum_{i=1}^{p} \mathbf{Q}(i) \cdot r(i) \tag{23a}$$

Similarly, by inspection, we can write

$$\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k) \quad = \quad \left[\sum_{i=1}^{p} \mathbf{Q}(i) \cdot r(i) \right] \; - \; \mathbf{Q}(k) \cdot r(k) \tag{23b}$$

Then, after inserting Eq. (23a) into Eq. (23b), we get the following relationship:

$$\boxed{\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k) \quad = \quad \mathbf{A^T}\,\mathbf{B} \; - \; \mathbf{Q}(k) \cdot r(k) \qquad (24)}$$

Rearranging terms in Eq. (24), we get:

$$\mathbf{Q}(k) \cdot r(k) \quad = \quad \mathbf{A^T}\,\mathbf{B} \; - \; \overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k) \tag{25}$$

Now, after inserting Eq. (25) into Eq. (21) and simplifying terms, we get:

$$\delta r(k) \quad = \quad \frac{r(k) \; - \; \mathbf{Q^T}(k) \cdot \left(\mathbf{A^T A}\right)^{-1} \cdot \mathbf{A^T}\,\mathbf{B}}{1 \; - \; h(k)} \tag{26}$$

We also know, after inserting Eq. (8b) into Eq. (11b), that the fitted response of the original global regression problem can be written in the following form:

$$\rho(k) \quad = \quad \mathbf{Q^T}(k) \cdot \mathbf{X} \quad = \quad \mathbf{Q^T}(k) \cdot \left(\mathbf{A^T A}\right)^{-1} \cdot \mathbf{A^T}\,\mathbf{B} \tag{27}$$

Inserting Eq. (27) into Eq. (26), we get:

$$\delta r(k) \quad = \quad \frac{r(k) \; - \; \rho(k)}{1 \; - \; h(k)} \tag{28}$$

Finally, after (1) using Eq. (11a) in order to replace the numerator of Eq. (28) and (2) using Eq. (17b) in order to replace $h(k)$ in Eq. (28), we get for the PRESS residual of the response the equation:

PRESS RESIDUAL OF A RESPONSE

$$\delta r(k) \quad = \quad \frac{\Delta r(k)}{1 \; - \; h(k)} \quad = \quad \frac{\Delta r(k)}{1 \; - \; \mathbf{Q^T}(k) \cdot (\mathbf{A^T A})^{-1} \cdot \mathbf{Q}(k)} \tag{29}$$

Equation (29) shows that the PRESS residual of the response of a data point is essentially a scaled value of the corresponding classical response residual of the data point. The scaling factor depends on (1) the inverse of matrix $\mathbf{A^T A}$ that is needed to obtain the solution of the original regression problem and (2) a vector $\mathbf{Q}(k)$ (see also Eq. (9a)) that is a function of the regressor values of the data point.

Appendix 2: Derivation of Relationship between Inverse Matrices

In App. 1 an important relationship between two inverse matrices was used in order to derive a formula for the PRESS residual of a response. The relationship is given by the following set of equations:

RELATIONSHIP BETWEEN INVERSE MATRICES

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right)^{-1} \;=\; \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} + \frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}}{1\,-\,h(k)} \tag{17a}$$

$$h(k) \;=\; \mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k) \tag{17b}$$

The above relationship needs to be proven rigorously. An abbreviated proof is published in Ref. [5]. This proof takes advantage of another relationship between products of sets of matrices that are needed for the calculation of the PRESS residual of the response of a data point. This relationship is given by the following expression:

RELATIONSHIP BETWEEN PRODUCTS OF MATRICES
(see App. 3 for a derivation of the relationship)

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right) \;=\; \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)\,-\,\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k) \tag{30}$$

Unfortunately, no rigorous proof of Eq. (30) is given in Ref. [5] even though the relationship is urgently needed for the proof of Eqs. (17a) and (17b). Therefore, in order to address the shortcoming of Ref. [5], the author developed his own proof of Eq. (30) that can be found in App. 3 of the present paper. For the time being it is assumed that Eq. (30) is a valid relationship so that the proof of Eqs. (17a) and (17b) can be completed.

The proof of Eq. (17a) and (17b) starts by realizing that the following relationship between a matrix and its inverse applies:

$$\mathbf{I} \;=\; \left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right)^{-1}\;\cdot\;\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right) \tag{31}$$

Therefore, we conclude that Eq. (17a) and (17b) can be proven by simply showing that the product of the right hand side of Eq. (17a) and the right hand side of Eq. (30) equals the identity matrix. The corresponding matrix product can be expressed as follows:

$$\mathbf{P} \;=\; \left[\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} + \frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}}{1\,-\,h(k)}\right]\cdot\left[\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)\,-\,\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\right] \tag{32}$$

Brackets on the right hand side of Eq. (32) can be expanded. Then, we get:

$$\begin{aligned}\mathbf{P} \;=\;& \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)\,-\,\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\\ &+\frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)}{1\,-\,h(k)}\\ &-\frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)}{1\,-\,h(k)}\end{aligned} \tag{33}$$

We also know that

$$\mathbf{I} \;=\; \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right) \tag{34}$$

Then, after using Eq. (34) in order to simplify Eq. (33), we get:

$$\begin{aligned}\mathbf{P} \;=\; \mathbf{I} \;-\; \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^\mathbf{T}(k) \;+\; \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} \\ -\; \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^\mathbf{T}(k) \cdot \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)}\end{aligned} \tag{35}$$

Terms in Eq. (35) can be rearranged. Then, Eq. (35) becomes:

$$\begin{aligned}\mathbf{P} \;=\; \mathbf{I} - \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \;\cdot\; \mathbf{Q}(k) \cdot [\;1 \;-\; h(k)\;] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} + \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} \\ -\; \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot \left[\; \mathbf{Q}^\mathbf{T}(k) \cdot \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \;\right] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)}\end{aligned} \tag{36}$$

We also know, by inspection, that the following relationship applies:

$$\mathbf{Q}(k)_{m\times1} \cdot \mathbf{Q}^\mathbf{T}(k)_{1\times m} \;=\; \mathbf{Q}(k)_{m\times1} \cdot [1]_{1\times1} \cdot \mathbf{Q}^\mathbf{T}(k)_{1\times m} \tag{37a}$$

In addition, we know that Eq. (17b) defines the scalar $h(k)$. We get:

$$[h(k)]_{1\times1} \;=\; \mathbf{Q}^\mathbf{T}(k)_{1\times m} \cdot \left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1}_{m\times m} \cdot \mathbf{Q}(k)_{m\times1} \tag{37b}$$

Then, after using Eq. (37a) and Eq. (37b) in order to simplify the numerator of the second and third fraction on the right hand side of Eq. (36), we get

$$\begin{aligned}\mathbf{P} \;=\; \mathbf{I} - \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \;\cdot\; \mathbf{Q}(k) \cdot [\;1 \;-\; h(k)\;] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} \\ +\; \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot [1] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} \;-\; \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot [h(k)] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)}\end{aligned} \tag{38}$$

Finally, after combining the last two fractions of the right hand side of Eq. (38), we get:

$$\mathbf{P} \;=\; \mathbf{I} - \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \;\cdot\; \mathbf{Q}(k) \cdot [\;1 \;-\; h(k)\;] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} + \frac{\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)^{-1} \cdot \mathbf{Q}(k) \cdot [1 \;-\; h(k)] \cdot \mathbf{Q}^\mathbf{T}(k)}{1 \;-\; h(k)} \tag{39}$$

Now, after further simplifying the right hand side of Eq. (39), we get the final result:

$$\mathbf{P} \;=\; \mathbf{I} \tag{40}$$

Therefore, it is proven that Eq. (17a) and (17b) are valid as the product of the right hand side of Eq. (17a) with the right hand side of Eq. (30) equals the identity matrix.

Appendix 3: Derivation of Relationship between Products of Matrices

In this appendix a relationship between two sets of matrices is proven that is urgently needed for the proof of Eq. (17a) in App. 2. The relationship has the following form:

RELATIONSHIP BETWEEN PRODUCTS OF MATRICES

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right) \quad = \quad \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right) \; - \; \mathbf{Q}(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \qquad (30)$$

The proof of Eq. (30) can be performed by showing that the following alternate formulation of Eq. (30) is a valid relationship:

$$\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right) \quad = \quad \sum_{k=1}^{p} \mathbf{Q}(k) \cdot \mathbf{Q}^{\mathbf{T}}(k) \qquad (41)$$

Matrix $\mathbf{A}$ above was defined in Eq. (10a). We know, by inspection, that the following relationship for this matrix is valid:

$$\mathbf{A}_{p\times m} \quad = \quad \begin{bmatrix} \mathbf{Q}^{\mathbf{T}}(1) \\ \mathbf{Q}^{\mathbf{T}}(2) \\ \vdots \\ \mathbf{Q}^{\mathbf{T}}(k) \\ \vdots \\ \mathbf{Q}^{\mathbf{T}}(p) \end{bmatrix} \quad = \quad \sum_{k=1}^{p} \mathbf{e}(k)_{p\times 1} \cdot \mathbf{Q}^{\mathbf{T}}(k)_{1\times m} \qquad (42)$$

where the unit basis vector $\mathbf{e}(k)$ can be expressed as

$$\mathbf{e}(k)_{p\times 1} \quad = \quad \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(\mu) \\ \vdots \\ e(p) \end{bmatrix} \quad ; \quad e(\mu) \quad = \quad \begin{cases} 0 & \text{if} \;\; \mu \neq k \\ 1 & \text{if} \;\; \mu = k \end{cases} \qquad (43)$$

and where vector $\mathbf{Q}^{\mathbf{T}}$ is given by Eq. (9b) as

$$\mathbf{Q}^{\mathbf{T}}(k)_{1\times m} \quad = \quad [1 \;\; x_1(k) \;\; x_2(k) \;\; \cdots] \qquad (9b)$$

We also know, that the transpose of the sum of two matrices equals the sum of the transpose of each matrix (from Ref. [12], p. 334):

$$[\, \mathbf{C} \, + \, \mathbf{D} \,]^{\mathbf{T}} \quad = \quad \mathbf{C}^{\mathbf{T}} \, + \, \mathbf{D}^{\mathbf{T}} \qquad (44a)$$

Then, the transpose of matrix $\mathbf{A}$ defined in Eq. (42) may be written as:

$$\mathbf{A}^{\mathbf{T}}{}_{m\times p} \quad = \quad \left[\sum_{k=1}^{p} \mathbf{e}(k)_{p\times 1} \cdot \mathbf{Q}^{\mathbf{T}}(k)_{1\times m} \right]^{\mathbf{T}} \quad = \quad \sum_{k=1}^{p} \left[\, \mathbf{e}(k)_{p\times 1} \cdot \mathbf{Q}^{\mathbf{T}}(k)_{1\times m} \,\right]^{\mathbf{T}} \qquad (44b)$$

The following two theorems for (1) the transpose of a matrix and for (2) the transpose of the product of two matrices apply (from Ref. [12], p. 334):

$$\mathbf{C} \;=\; \left(\mathbf{C}^{\mathbf{T}}\right)^{\mathbf{T}} \tag{45a}$$

$$(\,\mathbf{C}\;\mathbf{D}\,)^{\mathbf{T}} \;=\; \mathbf{D}^{\mathbf{T}}\,\mathbf{C}^{\mathbf{T}} \tag{45b}$$

Therefore, Eq. (44b) can also be written as:

$$\mathbf{A}^{\mathbf{T}}{}_{m\times p} \;=\; \sum_{k=1}^{p} \mathbf{Q}(k)_{m\times 1}\cdot \mathbf{e}^{\mathbf{T}}(k)_{1\times p} \tag{46}$$

Now, after multiplying the right hand side of Eq. (46) with the right hand side of Eq. (42), we get:

$$\begin{aligned}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)_{m\times m} \;&=\; \left[\sum_{i=1}^{p} \mathbf{Q}(i)_{m\times 1}\cdot \mathbf{e}^{\mathbf{T}}(i)_{1\times p}\right]\cdot\left[\sum_{j=1}^{p} \mathbf{e}(j)_{p\times 1}\cdot \mathbf{Q}^{\mathbf{T}}(j)_{1\times m}\right] \\ &=\; \sum_{i=1}^{p}\sum_{j=1}^{p}\left[\,\mathbf{Q}(i)_{m\times 1}\cdot \mathbf{e}^{\mathbf{T}}(i)_{1\times p}\cdot \mathbf{e}(j)_{p\times 1}\cdot \mathbf{Q}^{\mathbf{T}}(j)_{1\times m}\,\right]\end{aligned} \tag{47}$$

In general, the following relationship between the scalar product of unit basis vectors is valid:

$$\mathbf{e}^{\mathbf{T}}(i)_{1\times p}\cdot \mathbf{e}(j)_{p\times 1} \;=\; \delta_{ij} \;=\; \begin{cases}0 & \text{if } i \neq j\\ 1 & \text{if } i = j\end{cases} \tag{48}$$

Therefore, a summation term used on the right hand side of Eq. (47) is only non–zero if both summation indices i and j of the term are identical. Finally, we get:

$$\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)_{m\times m} \;=\; \sum_{i=1}^{p}\left[\,\mathbf{Q}(i)_{m\times 1}\cdot \mathbf{Q}^{\mathbf{T}}(i)_{1\times m}\,\right] \tag{49}$$

Now, after expanding the summation on the right hand side of Eq. (49) and after rearranging terms, we get:

$$\begin{aligned}\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right) \;-\; \mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k) \;&=\; \mathbf{Q}(1)\cdot\mathbf{Q}^{\mathbf{T}}(1) \;+\; \mathbf{Q}(2)\cdot\mathbf{Q}^{\mathbf{T}}(2) \;+\; \mathbf{Q}(3)\cdot\mathbf{Q}^{\mathbf{T}}(3) \;+\; \cdots \\ &\cdots \;+\; \mathbf{Q}(k-1)\cdot\mathbf{Q}^{\mathbf{T}}(k-1) \;+\; \mathbf{Q}(k+1)\cdot\mathbf{Q}^{\mathbf{T}}(k+1) \;+\; \cdots \\ &\cdots \;+\; \mathbf{Q}(p-1)\cdot\mathbf{Q}^{\mathbf{T}}(p-1) \;+\; \mathbf{Q}(p)\cdot\mathbf{Q}^{\mathbf{T}}(p)\end{aligned} \tag{50}$$

We also know that, after applying Eq. (49) to a matrix $\overline{\mathbf{A}}$ that equals matrix $\mathbf{A}$ with the k–th row removed, the following relationship is valid:

$$\begin{aligned}\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right) \;&=\; \mathbf{Q}(1)\cdot\mathbf{Q}^{\mathbf{T}}(1) \;+\; \mathbf{Q}(2)\cdot\mathbf{Q}^{\mathbf{T}}(2) \;+\; \mathbf{Q}(3)\cdot\mathbf{Q}^{\mathbf{T}}(3) \;+\; \cdots \\ &\cdots \;+\; \mathbf{Q}(k-1)\cdot\mathbf{Q}^{\mathbf{T}}(k-1) \;+\; \mathbf{Q}(k+1)\cdot\mathbf{Q}^{\mathbf{T}}(k+1) \;+\; \cdots \\ &\cdots \;+\; \mathbf{Q}(p-1)\cdot\mathbf{Q}^{\mathbf{T}}(p-1) \;+\; \mathbf{Q}(p)\cdot\mathbf{Q}^{\mathbf{T}}(p)\end{aligned} \tag{51}$$

We see, by inspection, that the right hand side of Eq. (50) equals the right hand side of Eq. (51). Therefore, we conclude that the left hand side of Eq. (50) must equal the left hand side of Eq. (51). Then, we get:

$$\boxed{\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right) \;=\; \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right) \;-\; \mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k) \qquad (52)}$$

Consequently, knowing that vector $\mathbf{Q}^{\mathbf{T}}$ is the k–th row of matrix $\mathbf{A}$, we conclude that the right hand side of Eq. (52) equals the matrix product after the k–th row is removed in matrix $\mathbf{A}$.

Appendix 4: Calculation of the PRESS Residual of a Regressor

Strain–gage balances are used in wind tunnel testing in order to measure forces and moments that act on a wind tunnel model during a test. A strain–gage balance has to be calibrated so that measured electrical outputs of the strain–gages can be related to aerodynamic loads that the wind tunnel model experiences during the test. Typical load combinations are applied during the calibration of the balance and corresponding strain–gage outputs are recorded. Then, a regression analysis of the calibration data is performed that fits the measured electrical strain–gage outputs (responses) as a function of the applied calibration loads (regressors). In a final step, the result of the regression analysis is used to construct an iteration equation that makes it possible to directly compute loads, i.e., the value of the regressor variable, from the measures electrical outputs during a wind tunnel test. This iteration process is unique to the analysis and use of wind tunnel strain–gage balance calibration data. A more detailed explanation of the strain–gage balance calibration analysis process and the derivation of an iteration equation that allows for a direct calculation of regressors values (aerodynamic loads) from a given set of responses (measured electrical strain–gage output) is given in Ref. [1].

A calculation of the PRESS residual of an aerodynamic load (regressor) is of great interest to the wind tunnel testing community as aerodynamic loads are often the most important result of wind tunnel tests. Unfortunately, the PRESS residual is traditionally computed for a response value and not for a regressor value (see App. 1). However, the author realized that PRESS residuals of regressors can be computed in a two step process. This two step process combines (1) the regression analysis of the modified original experimental data set with (2) an iteration process for each data point.

The modified experimental data set needed for the calculation of the PRESS residual of an aerodynamic load is identical with the modified data set that is discussed in App. 1. In principle, the calculation of the PRESS residuals of the aerodynamic needs the solution of the global regression problem and the load iteration result for the data point if the data point (index k) itself is withheld from the original data set. Therefore, the modified data set equals the original data set with the data point itself withheld. Consequently, matrix $\mathbf{A}$ of the regression problem has one fewer row and matrix $\mathbf{A^T}$ of the regression problem has one fewer column. The regression solution now becomes a function of the data point (index k) that is omitted for the calculation of the PRESS residuals.

The global regression solution for the reduced calibration data set is needed in order to start the load iteration for a given data point where the PRESS residual of the regressors (aerodynamic loads) is to be computed. It is assumed, similar to Eq. (12a) in App. 1, that the matrix and vector describing the global regression problem of the modified original data set are given as:

$$\overline{\mathbf{A}}(k) \;\cdot\; \overline{\mathbf{X}}(k) \quad = \quad \overline{\mathbf{B}}(k) \tag{53}$$

Then, the global regression solution for the modified calibration data set is simply given by the normal equations:

$$\overline{\mathbf{X}}(k) \quad = \quad \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right)^{-1} \;\cdot\; \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{B}}(k)\right) \tag{54}$$

It can be seen from Eq. (54) that the solution of the global regression problem of the modified data set is a funtion of the following matrices and vectors:

$$\overline{\mathbf{X}}(k) \quad \Longrightarrow \quad \mathcal{F}\Big\{ \left(\overline{\mathbf{A}}^{\mathbf{T}}(k) \cdot \overline{\mathbf{A}}(k)\right)^{-1} \;\; ; \;\; \overline{\mathbf{A}}(k) \;\; ; \;\; \overline{\mathbf{B}}(k) \Big\}$$

It is obvious that the calculation of the inverse matrix on the right hand side of Eq. (54) is a very time consuming operation if the solution of the global regression problem has to be found. It appears, superficially viewed, that the calculation of the global regression solution requires a complete calculation of the inverse matrix for each data point if the PRESS residuals of all data points are to be determined. Fortunately, we can take again advantage of Eq. (17a) and (17b) that are derived in App. 2. These equations have the following form:

RELATIONSHIP BETWEEN INVERSE MATRICES

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right)^{-1} \;=\; \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} + \frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}}{1\;-\;h(k)} \qquad (17a)$$

$$h(k) \;=\; \mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k) \qquad (17b)$$

We get, after inserting Eq. (17a) and (17b) into Eq. (54), the following solution of the global regression problem of the modified data set:

$$\overline{\mathbf{X}}(k) \;=\; \left[\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1} + \frac{\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)\cdot\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}}{1\;-\;\mathbf{Q}^{\mathbf{T}}(k)\cdot\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\cdot\mathbf{Q}(k)}\right]\cdot\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{B}}(k)\right) \qquad (55)$$

Now, the global regression solution of the modified data set depends on the following matrices and vectors:

$$\overline{\mathbf{X}}(k) \;\Longrightarrow\; \mathcal{F}\Big\{\;\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\;;\;\mathbf{Q}(k)\;;\;\overline{\mathbf{A}}(k)\;;\;\overline{\mathbf{B}}(k)\;\Big\}$$

The originally required matrix inversion of the modified data set for each data point $(1 \leq k \leq p)$, i.e.,

$$\left(\overline{\mathbf{A}}^{\mathbf{T}}(k)\cdot\overline{\mathbf{A}}(k)\right)^{-1}$$

has been replaced by simple matrix additions and multiplications that use the inverse matrix of the original regression problem, i.e.,

$$\left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}$$

The matrix of the original regression problem only has to be computed <u>once</u> for all data points. Therefore, the time consuming part of the calculation of the PRESS residuals of the aerodynamic loads is <u>not</u> the calculation of the global regression solution of the reduced data set. Instead, it is the iteration that has to be performed for each data point after the solution of the global regression problem of the reduced data set is obtained. All matrices used in this iteration equation are directly derived for each data point with index k from the global regression solution $\overline{\mathbf{X}}(k)$ of the modified, i.e., reduced, original calibration data set.

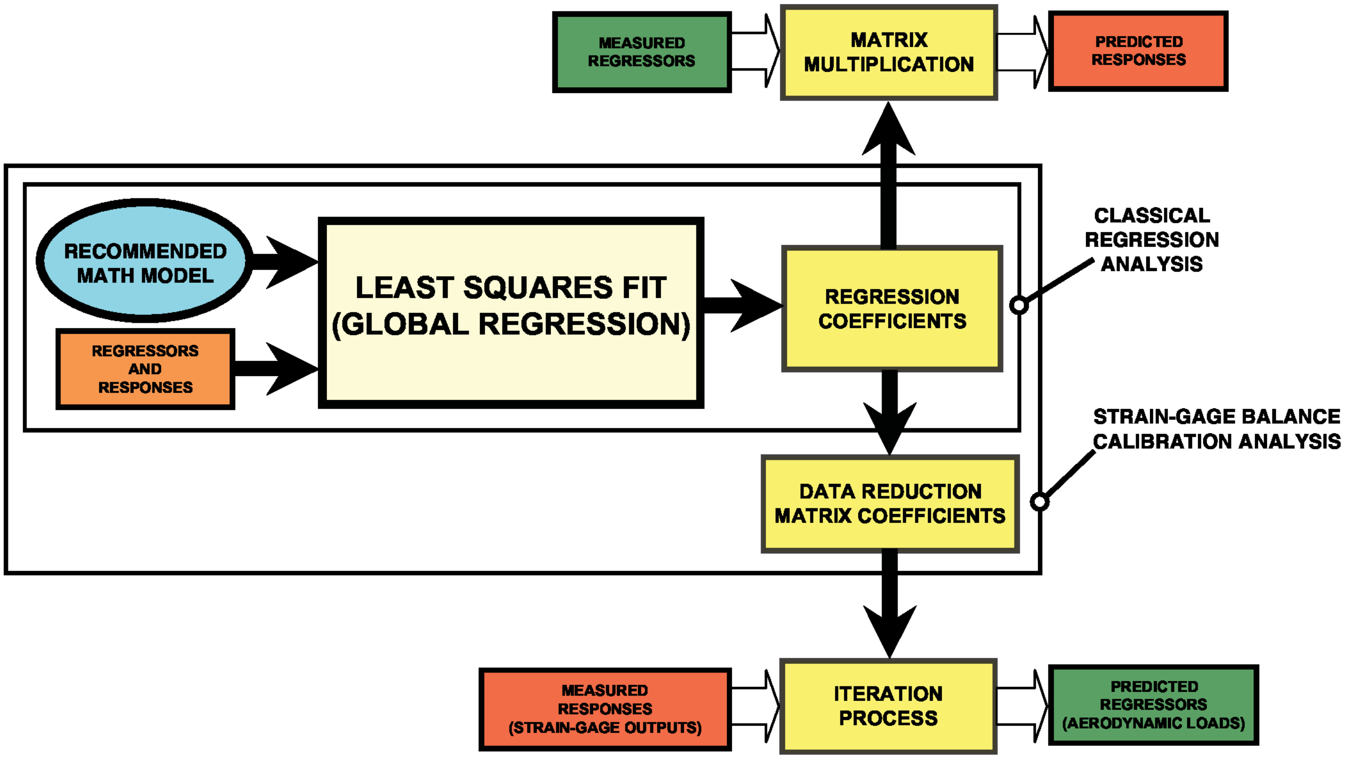

MEASURED REGRESSORS
MATRIX MULTIPLICATION
PREDICTED RESPONSES
RECOMMENDED MATH MODEL
REGRESSORS AND RESPONSES
LEAST SQUARES FIT (GLOBAL REGRESSION)
REGRESSION COEFFICIENTS
CLASSICAL REGRESSION ANALYSIS
STRAIN-GAGE BALANCE CALIBRATION ANALYSIS
DATA REDUCTION MATRIX COEFFICIENTS
MEASURED RESPONSES (STRAIN-GAGE OUTPUTS)
ITERATION PROCESS
PREDICTED REGRESSORS (AERODYNAMIC LOADS)

Fig. 1 Connection between classical regression analysis and wind tunnel strain–gage balance calibration analysis.

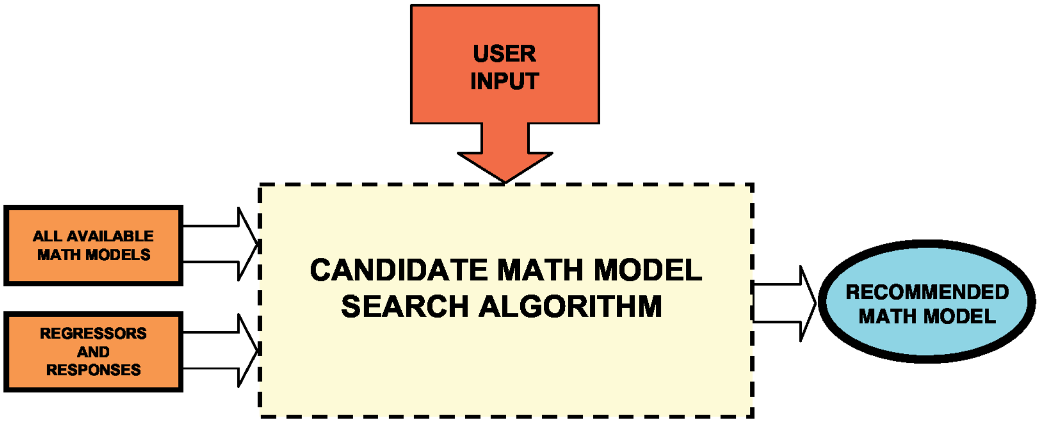

USER INPUT
ALL AVAILABLE MATH MODELS
REGRESSORS AND RESPONSES
CANDIDATE MATH MODEL SEARCH ALGORITHM
RECOMMENDED MATH MODEL

Fig. 2 Basic input and output of candidate math model search algorithm.

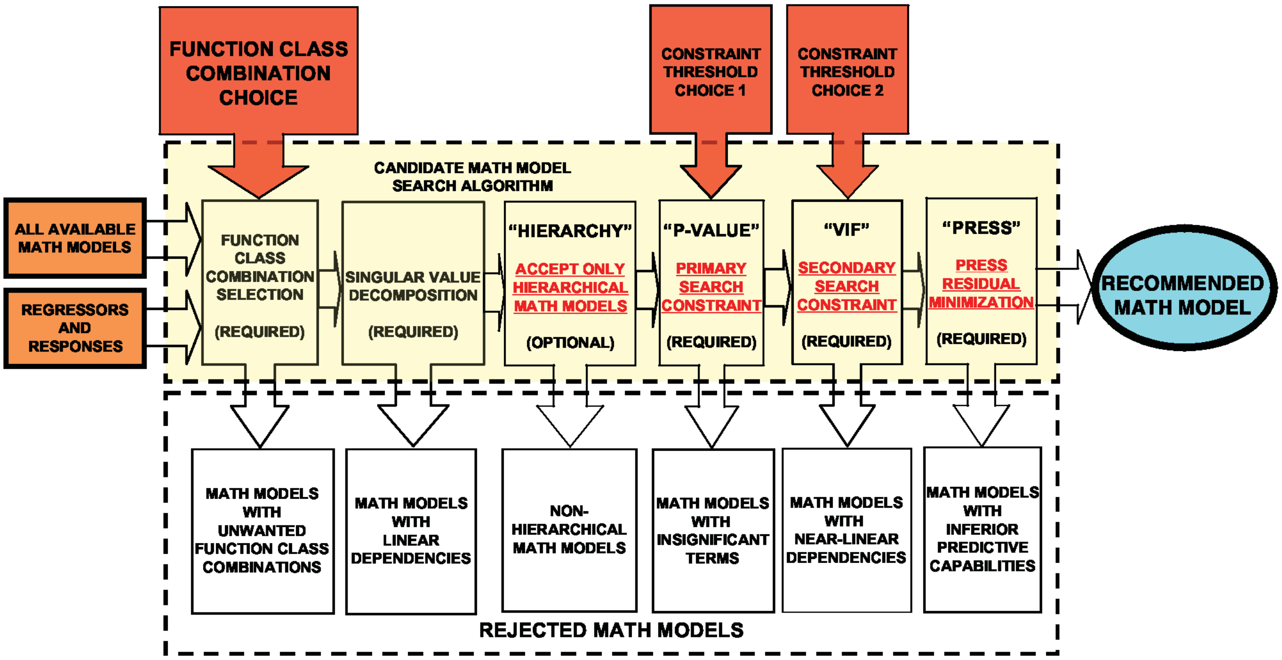

FUNCTION CLASS COMBINATION CHOICE
CONSTRAINT THRESHOLD CHOICE 1
CONSTRAINT THRESHOLD CHOICE 2
CANDIDATE MATH MODEL SEARCH ALGORITHM
ALL AVAILABLE MATH MODELS
REGRESSORS AND RESPONSES
FUNCTION CLASS COMBINATION SELECTION (REQUIRED)
SINGULAR VALUE DECOMPOSITION (REQUIRED)
"HIERARCHY" ACCEPT ONLY HIERARCHICAL MATH MODELS (OPTIONAL)
"P-VALUE" PRIMARY SEARCH CONSTRAINT (REQUIRED)
"VIF" SECONDARY SEARCH CONSTRAINT (REQUIRED)
"PRESS" PRESS RESIDUAL MINIMIZATION (REQUIRED)
RECOMMENDED MATH MODEL
MATH MODELS WITH UNWANTED FUNCTION CLASS COMBINATIONS
MATH MODELS WITH LINEAR DEPENDENCIES
NON-HIERARCHICAL MATH MODELS
MATH MODELS WITH INSIGNIFICANT TERMS
MATH MODELS WITH NEAR-LINEAR DEPENDENCIES
MATH MODELS WITH INFERIOR PREDICTIVE CAPABILITIES
REJECTED MATH MODELS

Fig. 3a Candidate math model search algorithm of 2008 (hierarchy rule applied during search).

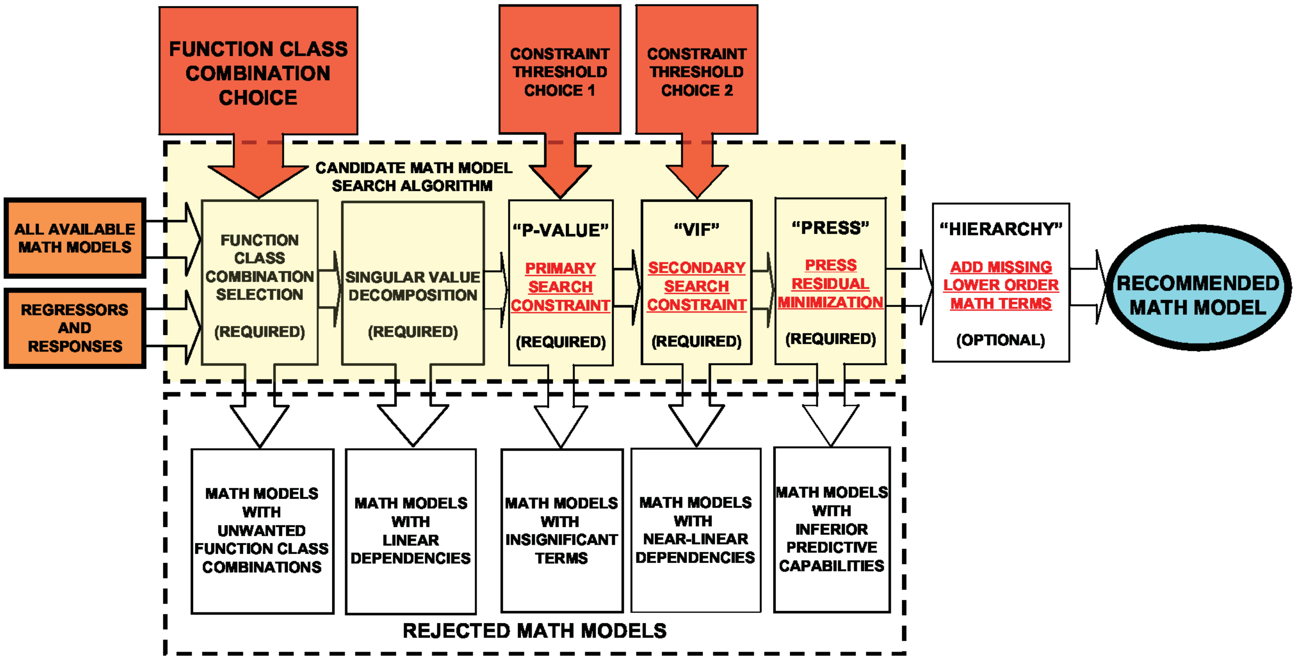

FUNCTION CLASS COMBINATION CHOICE
CONSTRAINT THRESHOLD CHOICE 1
CONSTRAINT THRESHOLD CHOICE 2
CANDIDATE MATH MODEL SEARCH ALGORITHM
ALL AVAILABLE MATH MODELS
REGRESSORS AND RESPONSES
FUNCTION CLASS COMBINATION SELECTION (REQUIRED)
SINGULAR VALUE DECOMPOSITION (REQUIRED)
"P-VALUE" PRIMARY SEARCH CONSTRAINT (REQUIRED)
"VIF" SECONDARY SEARCH CONSTRAINT (REQUIRED)
"PRESS" PRESS RESIDUAL MINIMIZATION (REQUIRED)
"HIERARCHY" ADD MISSING LOWER ORDER MATH TERMS (OPTIONAL)
RECOMMENDED MATH MODEL
MATH MODELS WITH UNWANTED FUNCTION CLASS COMBINATIONS
MATH MODELS WITH LINEAR DEPENDENCIES
MATH MODELS WITH INSIGNIFICANT TERMS
MATH MODELS WITH NEAR-LINEAR DEPENDENCIES
MATH MODELS WITH INFERIOR PREDICTIVE CAPABILITIES
REJECTED MATH MODELS

Fig. 3b Candidate math model search algorithm of 2008 (hierarchy rule applied after search).

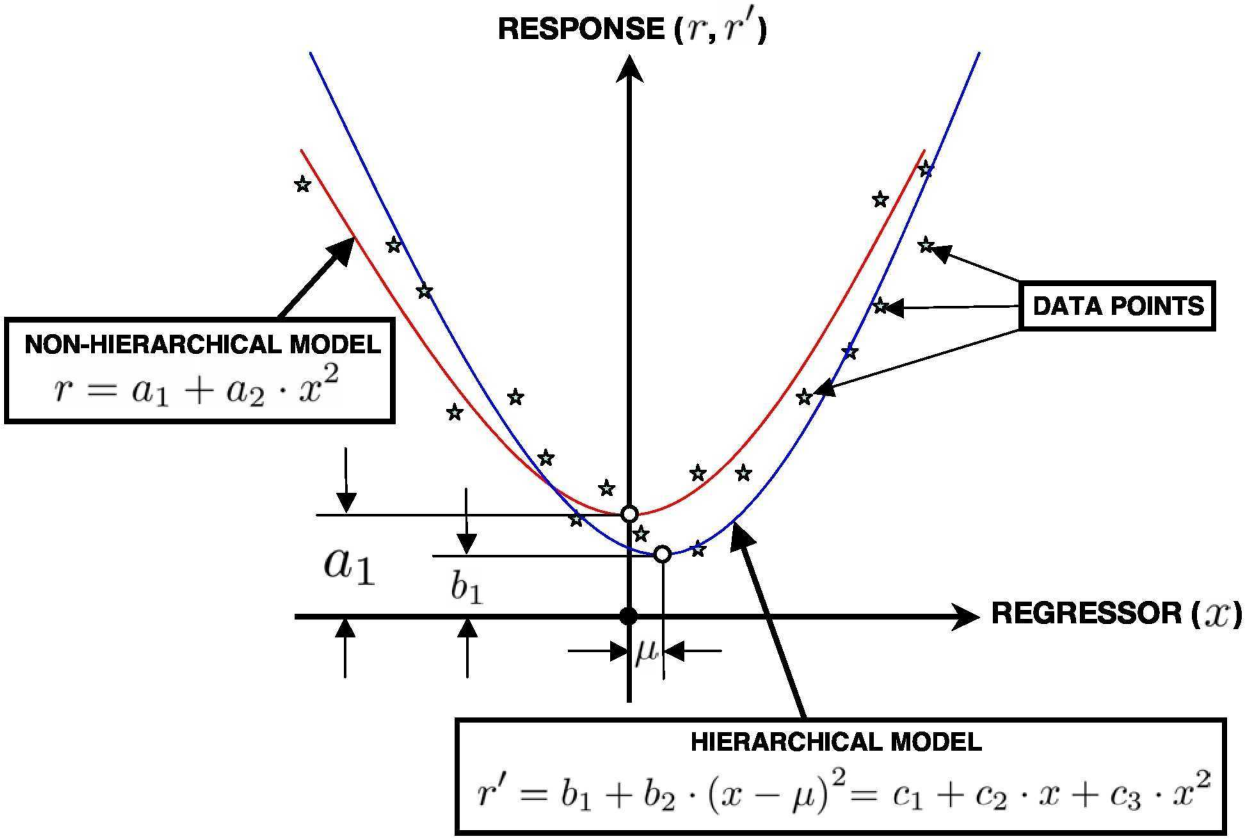
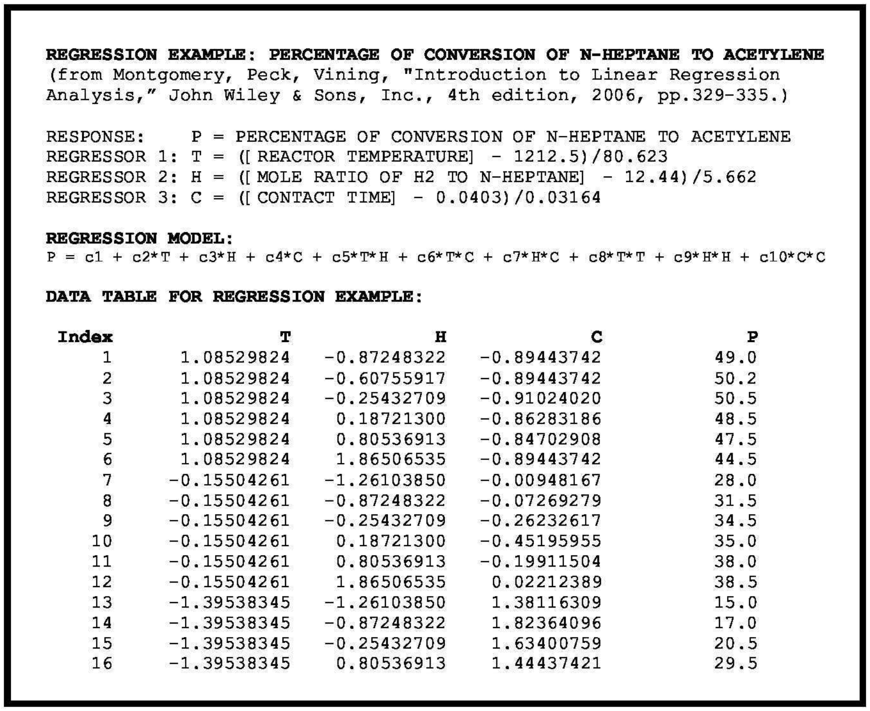
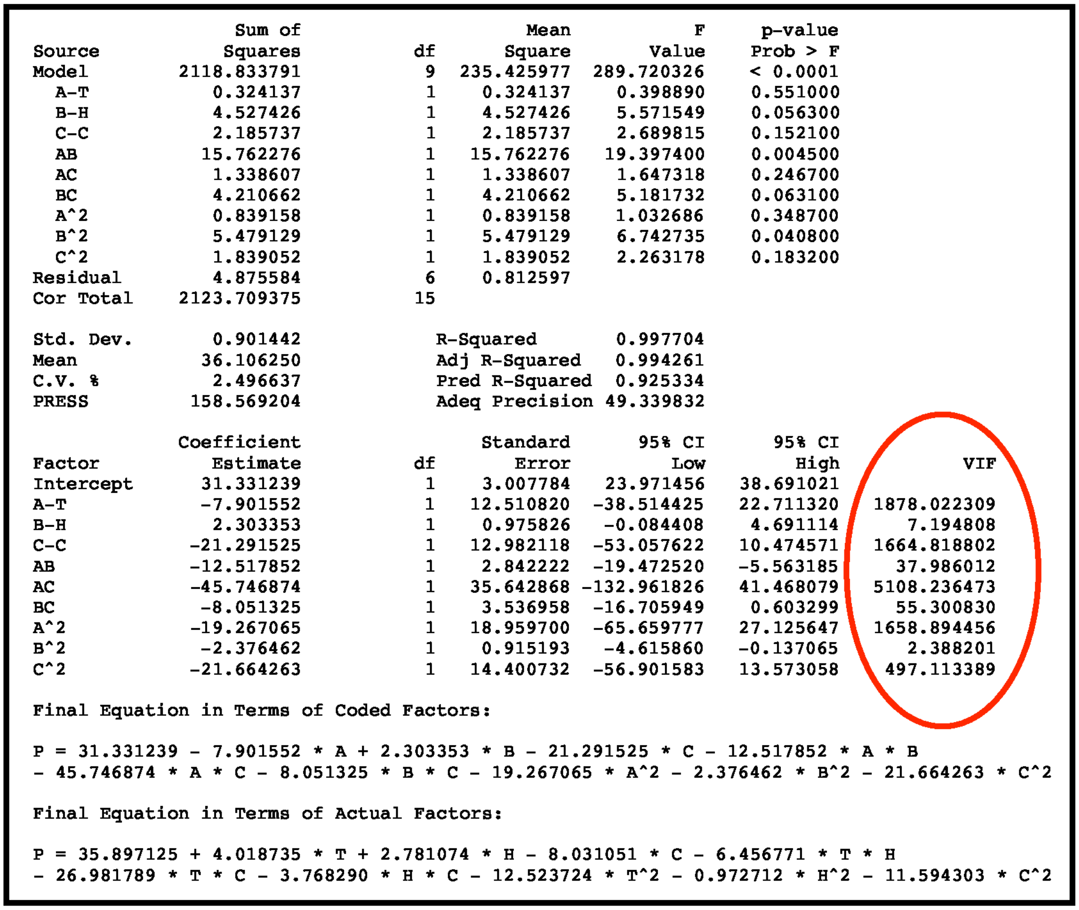


Fig. 4 Hierarchical versus non–hierarchical regression model of responses.



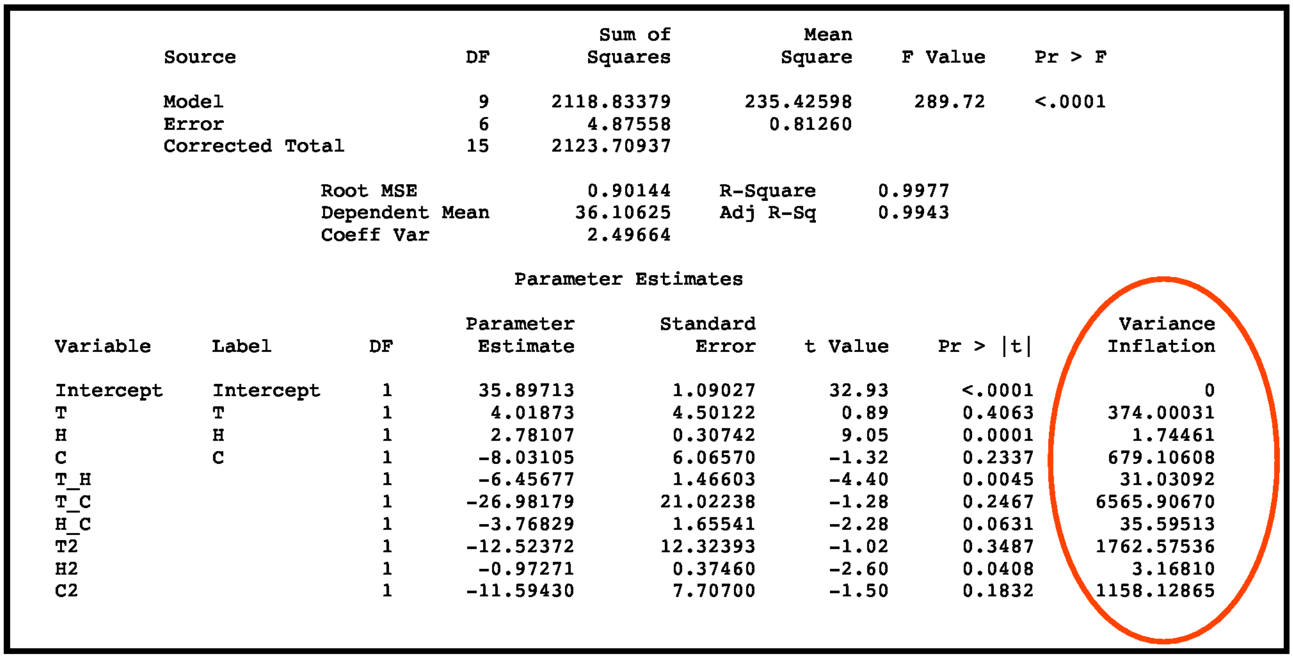
```
REGRESSION EXAMPLE: PERCENTAGE OF CONVERSION OF N-HEPTANE TO ACETYLENE
(from Montgomery, Peck, Vining, "Introduction to Linear Regression
Analysis," John Wiley & Sons, Inc., 4th edition, 2006, pp.329-335.)

RESPONSE:     P = PERCENTAGE OF CONVERSION OF N-HEPTANE TO ACETYLENE
REGRESSOR 1: T = ([REACTOR TEMPERATURE] - 1212.5)/80.623
REGRESSOR 2: H = ([MOLE RATIO OF H2 TO N-HEPTANE] - 12.44)/5.662
REGRESSOR 3: C = ([CONTACT TIME] - 0.0403)/0.03164

REGRESSION MODEL:
P = c1 + c2*T + c3*H + c4*C + c5*T*H + c6*T*C + c7*H*C + c8*T*T + c9*H*H + c10*C*C

DATA TABLE FOR REGRESSION EXAMPLE:

 Index           T            H            C           P
     1    1.08529824  -0.87248322  -0.89443742        49.0
     2    1.08529824  -0.60755917  -0.89443742        50.2
     3    1.08529824  -0.25432709  -0.91024020        50.5
     4    1.08529824   0.18721300  -0.86283186        48.5
     5    1.08529824   0.80536913  -0.84702908        47.5
     6    1.08529824   1.86506535  -0.89443742        44.5
     7   -0.15504261  -1.26103850  -0.00948167        28.0
     8   -0.15504261  -0.87248322  -0.07269279        31.5
     9   -0.15504261  -0.25432709  -0.26232617        34.5
    10   -0.15504261   0.18721300  -0.45195955        35.0
    11   -0.15504261   0.80536913  -0.19911504        38.0
    12   -0.15504261   1.86506535   0.02212389        38.5
    13   -1.39538345  -1.26103850   1.38116309        15.0
    14   -1.39538345  -0.87248322   1.82364096        17.0
    15   -1.39538345  -0.25432709   1.63400759        20.5
    16   -1.39538345   0.80536913   1.44437421        29.5
```

Fig. 5a Textbook regression analysis example (acetylene data, Ref. [5], pp. 329–335).



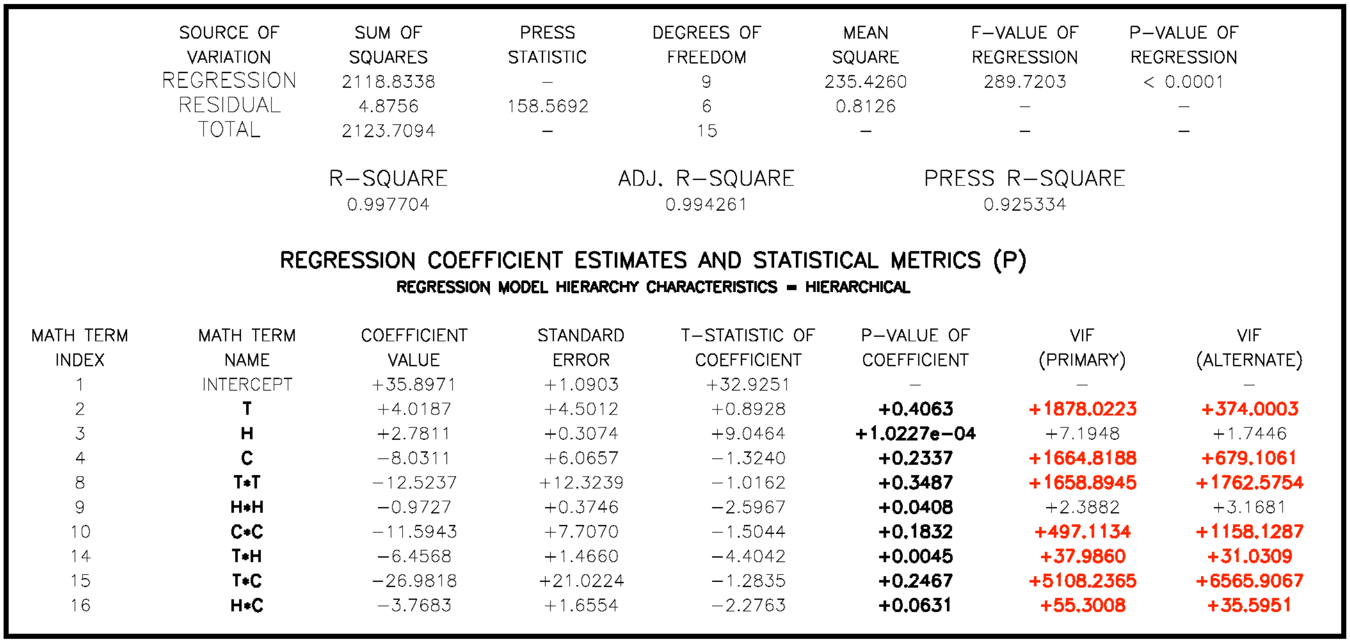
```
                    Sum of                      Mean          F     p-value
Source             Squares          df        Square      Value    Prob > F
Model          2118.833791           9    235.425977 289.720326    < 0.0001
  A-T             0.324137           1      0.324137   0.398890    0.551000
  B-H             4.527426           1      4.527426   5.571549    0.056300
  C-C             2.185737           1      2.185737   2.689815    0.152100
  AB             15.762276           1     15.762276  19.397400    0.004500
  AC              1.338607           1      1.338607   1.647318    0.246700
  BC              4.210662           1      4.210662   5.181732    0.063100
  A^2             0.839158           1      0.839158   1.032686    0.348700
  B^2             5.479129           1      5.479129   6.742735    0.040800
  C^2             1.839052           1      1.839052   2.263178    0.183200
Residual          4.875584           6      0.812597
Cor Total      2123.709375          15

Std. Dev.         0.901442             R-Squared        0.997704
Mean             36.106250             Adj R-Squared    0.994261
C.V. %            2.496637             Pred R-Squared   0.925334
PRESS           158.569204             Adeq Precision  49.339832

               Coefficient                Standard      95% CI      95% CI
Factor            Estimate          df       Error         Low        High           VIF
Intercept        31.331239           1    3.007784   23.971456   38.691021
A-T              -7.901552           1   12.510820  -38.514425   22.711320   1878.022309
B-H               2.303353           1    0.975826   -0.084408    4.691114      7.194808
C-C             -21.291525           1   12.982118  -53.057622   10.474571   1664.818802
AB              -12.517852           1    2.842222  -19.472520   -5.563185     37.986012
AC              -45.746874           1   35.642868 -132.961826   41.468079   5108.236473
BC               -8.051325           1    3.536958  -16.705949    0.603299     55.300830
A^2             -19.267065           1   18.959700  -65.659777   27.125647   1658.894456
B^2              -2.376462           1    0.915193   -4.615860   -0.137065      2.388201
C^2             -21.664263           1   14.400732  -56.901583   13.573058    497.113389

Final Equation in Terms of Coded Factors:

P = 31.331239 - 7.901552 * A + 2.303353 * B - 21.291525 * C - 12.517852 * A * B
- 45.746874 * A * C - 8.051325 * B * C - 19.267065 * A^2 - 2.376462 * B^2 - 21.664263 * C^2

Final Equation in Terms of Actual Factors:

P = 35.897125 + 4.018735 * T + 2.781074 * H - 8.031051 * C - 6.456771 * T * H
- 26.981789 * T * C - 3.768290 * H * C - 12.523724 * T^2 - 0.972712 * H^2 - 11.594303 * C^2
```

Fig. 5b DESIGN–EXPERT: Regression analysis results for acetylene data.



```
                                            Sum of          Mean
        Source                    DF       Squares        Square    F Value    Pr > F

        Model                      9    2118.83379     235.42598     289.72    <.0001
        Error                      6       4.87558       0.81260
        Corrected Total           15    2123.70937

                     Root MSE                0.90144    R-Square     0.9977
                     Dependent Mean         36.10625    Adj R-Sq     0.9943
                     Coeff Var               2.49664

                                   Parameter Estimates

                                      Parameter       Standard                               Variance
Variable      Label         DF         Estimate          Error    t Value    Pr > |t|       Inflation

Intercept     Intercept      1         35.89713        1.09027      32.93      <.0001               0
T             T              1          4.01873        4.50122       0.89      0.4063       374.00031
H             H              1          2.78107        0.30742       9.05      0.0001         1.74461
C             C              1         -8.03105        6.06570      -1.32      0.2337       679.10608
T_H                          1         -6.45677        1.46603      -4.40      0.0045        31.03092
T_C                          1        -26.98179       21.02238      -1.28      0.2467      6565.90670
H_C                          1         -3.76829        1.65541      -2.28      0.0631        35.59513
T2                           1        -12.52372       12.32393      -1.02      0.3487      1762.57536
H2                           1         -0.97271        0.37460      -2.60      0.0408         3.16810
C2                           1        -11.59430        7.70700      -1.50      0.1832      1158.12865
```

Fig. 5c SAS/STAT: Regression analysis results for acetylene data.



SOURCE OF VARIATION	SUM OF SQUARES	PRESS STATISTIC	DEGREES OF FREEDOM	MEAN SQUARE	F-VALUE OF REGRESSION	P-VALUE OF REGRESSION
REGRESSION	2118.8338	–	9	235.4260	289.7203	< 0.0001
RESIDUAL	4.8756	158.5692	6	0.8126	–	–
TOTAL	2123.7094	–	15	–	–	–

R-SQUARE	ADJ. R-SQUARE	PRESS R-SQUARE
0.997704	0.994261	0.925334

REGRESSION COEFFICIENT ESTIMATES AND STATISTICAL METRICS (P)

REGRESSION MODEL HIERARCHY CHARACTERISTICS = HIERARCHICAL

MATH TERM INDEX	MATH TERM NAME	COEFFICIENT VALUE	STANDARD ERROR	T-STATISTIC OF COEFFICIENT	P-VALUE OF COEFFICIENT	VIF (PRIMARY)	VIF (ALTERNATE)
1	INTERCEPT	+35.8971	+1.0903	+32.9251	–	–	–
2	**T**	+4.0187	+4.5012	+0.8928	**+0.4063**	**+1878.0223**	**+374.0003**
3	**H**	+2.7811	+0.3074	+9.0464	**+1.0227e-04**	+7.1948	+1.7446
4	**C**	-8.0311	+6.0657	-1.3240	**+0.2337**	**+1664.8188**	**+679.1061**
8	**T*T**	-12.5237	+12.3239	-1.0162	**+0.3487**	**+1658.8945**	**+1762.5754**
9	**H*H**	-0.9727	+0.3746	-2.5967	**+0.0408**	+2.3882	+3.1681
10	**C*C**	-11.5943	+7.7070	-1.5044	**+0.1832**	**+497.1134**	**+1158.1287**
14	**T*H**	-6.4568	+1.4660	-4.4042	**+0.0045**	**+37.9860**	**+31.0309**
15	**T*C**	-26.9818	+21.0224	-1.2835	**+0.2467**	**+5108.2365**	**+6565.9067**
16	**H*C**	-3.7683	+1.6554	-2.2763	**+0.0631**	**+55.3008**	**+35.5951**

Fig. 5d BALFIT: Regression analysis results for acetylene data.

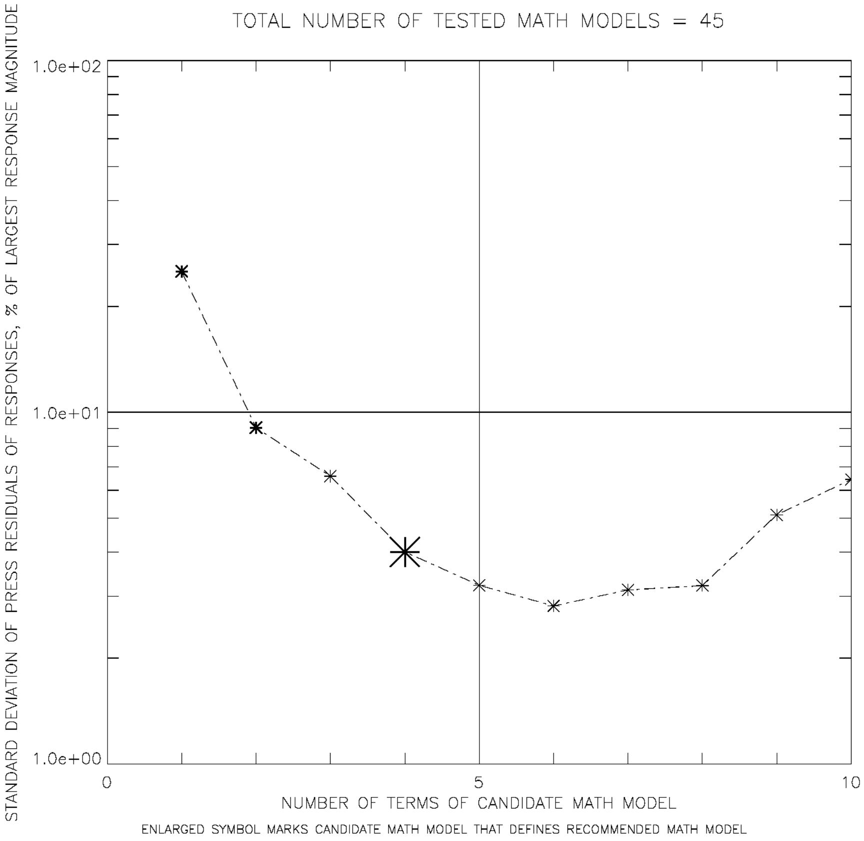


Fig. 6a BALFIT: Search metric for candidate math models of acetylene data.

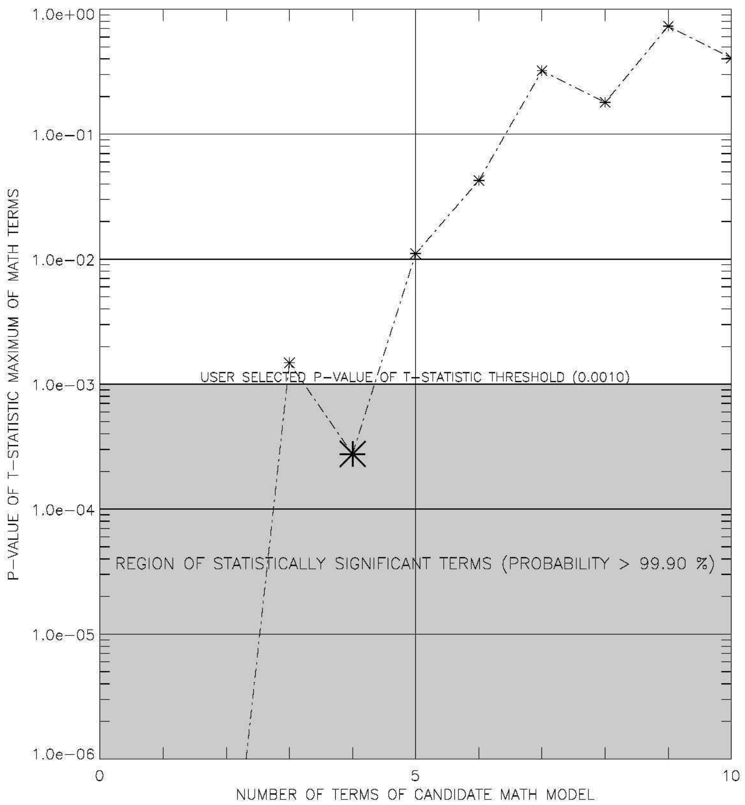


Fig. 6b BALFIT: p–value of t–statistic maximum of candidate math models of acetylene data.

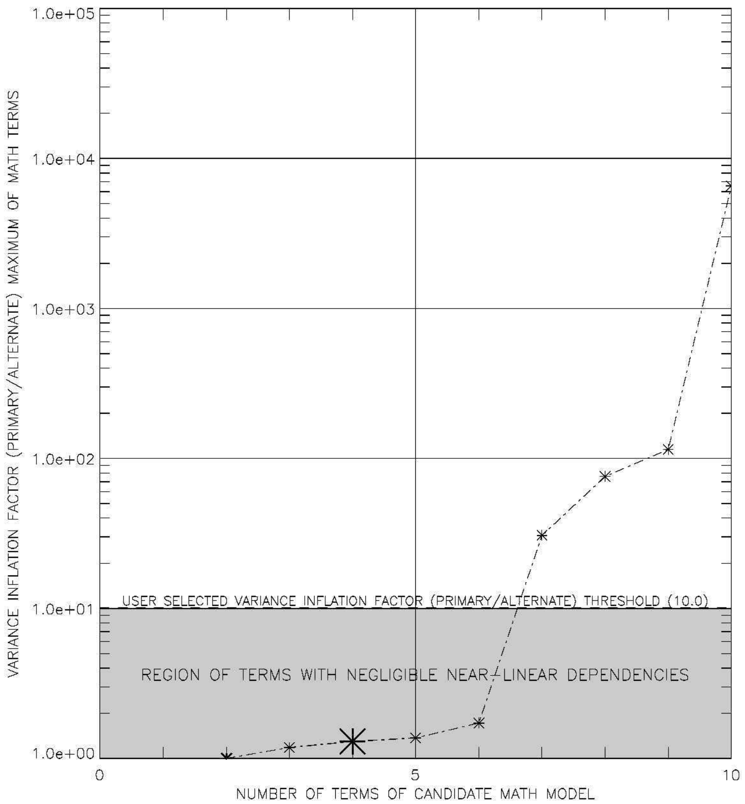
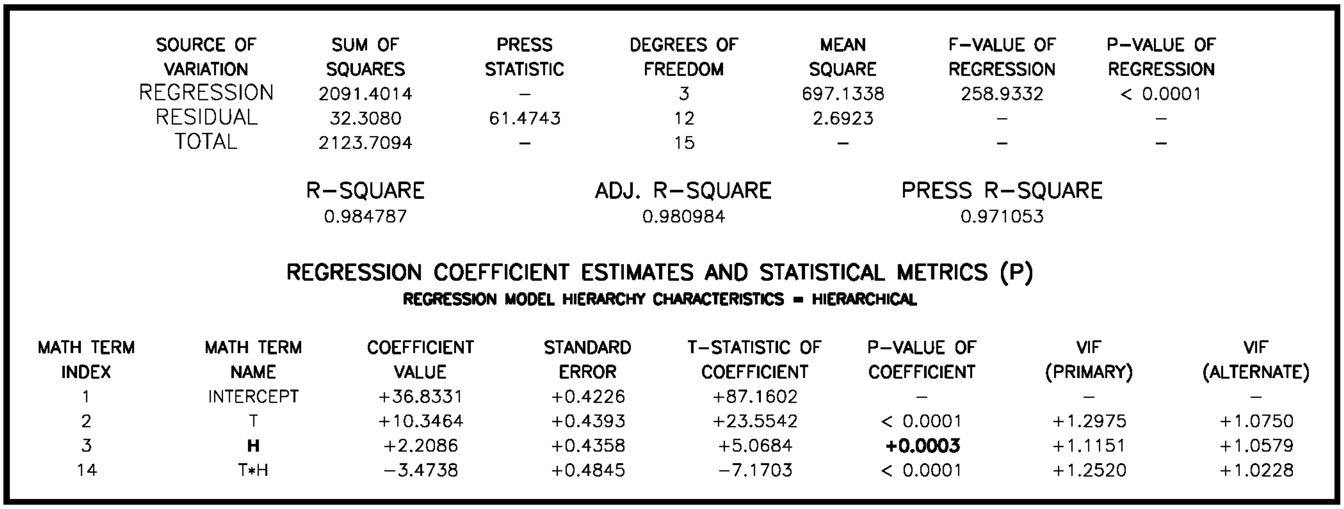


Fig. 6c BALFIT: Variance inflation factor maximum of candidate math models of acetylene data.



SOURCE OF VARIATION	SUM OF SQUARES	PRESS STATISTIC	DEGREES OF FREEDOM	MEAN SQUARE	F−VALUE OF REGRESSION	P−VALUE OF REGRESSION
REGRESSION	2091.4014	–	3	697.1338	258.9332	< 0.0001
RESIDUAL	32.3080	61.4743	12	2.6923	–	–
TOTAL	2123.7094	–	15	–	–	–

R−SQUARE	ADJ. R−SQUARE	PRESS R−SQUARE
0.984787	0.980984	0.971053

REGRESSION COEFFICIENT ESTIMATES AND STATISTICAL METRICS (P)

REGRESSION MODEL HIERARCHY CHARACTERISTICS = HIERARCHICAL

MATH TERM INDEX	MATH TERM NAME	COEFFICIENT VALUE	STANDARD ERROR	T−STATISTIC OF COEFFICIENT	P−VALUE OF COEFFICIENT	VIF (PRIMARY)	VIF (ALTERNATE)
1	INTERCEPT	+36.8331	+0.4226	+87.1602	–	–	–
2	T	+10.3464	+0.4393	+23.5542	< 0.0001	+1.2975	+1.0750
3	**H**	+2.2086	+0.4358	+5.0684	**+0.0003**	+1.1151	+1.0579
14	T∗H	−3.4738	+0.4845	−7.1703	< 0.0001	+1.2520	+1.0228

Fig. 7a BALFIT: Analysis of variance results for recommended math model of acetylene data.

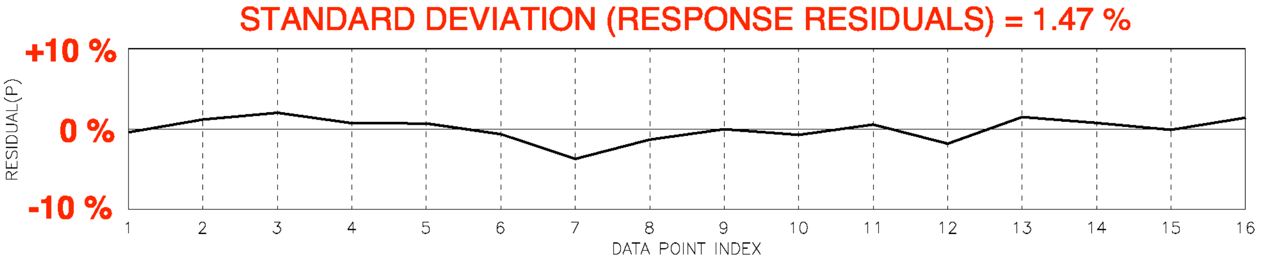


Fig. 7b BALFIT: Response residuals for recommended math model of acetylene data.

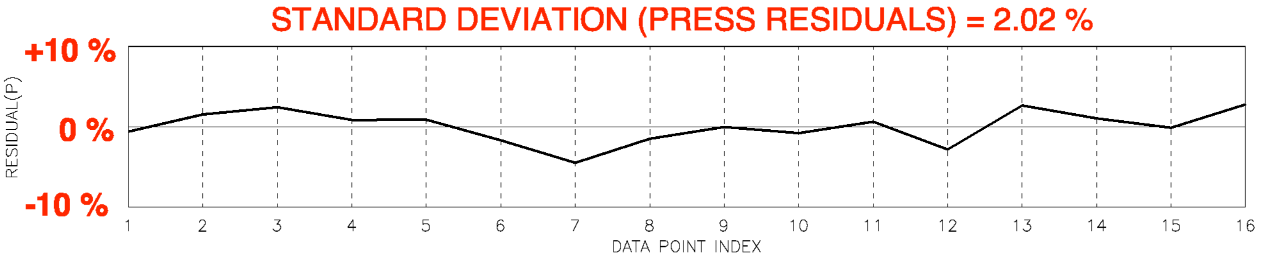
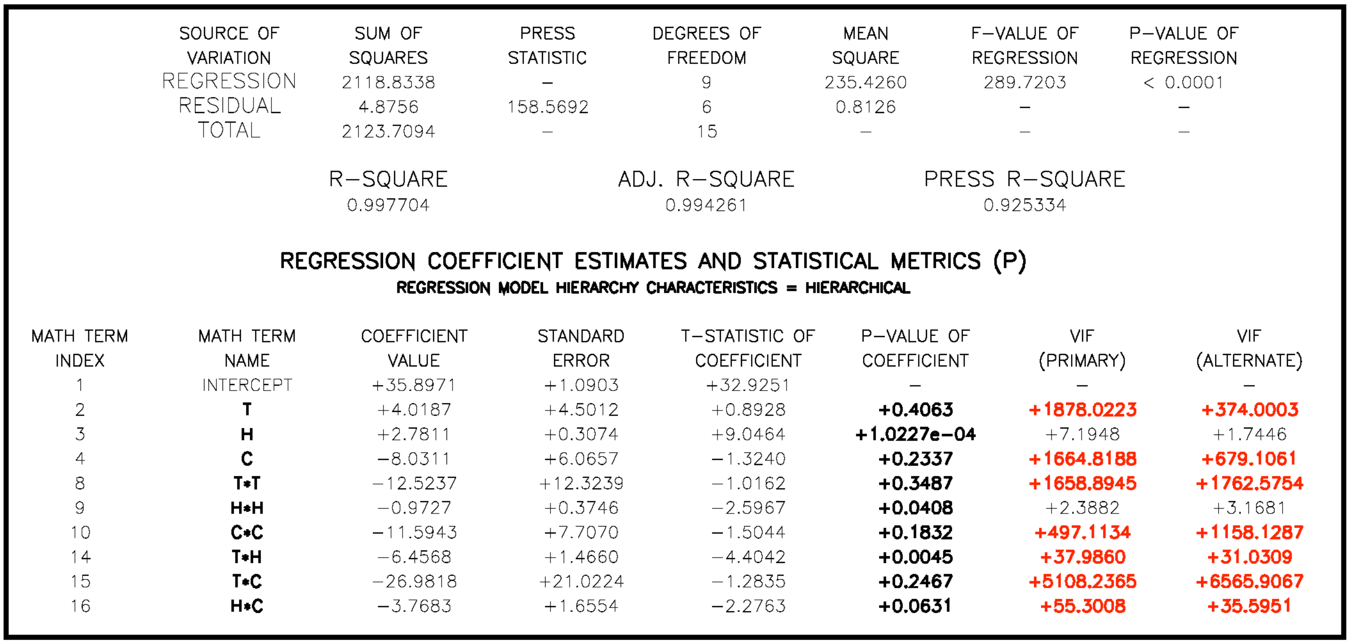


Fig. 7c BALFIT: PRESS residuals for recommended math model of acetylene data.



SOURCE OF VARIATION	SUM OF SQUARES	PRESS STATISTIC	DEGREES OF FREEDOM	MEAN SQUARE	F–VALUE OF REGRESSION	P–VALUE OF REGRESSION
REGRESSION	2118.8338	–	9	235.4260	289.7203	< 0.0001
RESIDUAL	4.8756	158.5692	6	0.8126	–	–
TOTAL	2123.7094	–	15	–	–	–

R–SQUARE	ADJ. R–SQUARE	PRESS R–SQUARE
0.997704	0.994261	0.925334

REGRESSION COEFFICIENT ESTIMATES AND STATISTICAL METRICS (P)

REGRESSION MODEL HIERARCHY CHARACTERISTICS = HIERARCHICAL

MATH TERM INDEX	MATH TERM NAME	COEFFICIENT VALUE	STANDARD ERROR	T–STATISTIC OF COEFFICIENT	P–VALUE OF COEFFICIENT	VIF (PRIMARY)	VIF (ALTERNATE)
1	INTERCEPT	+35.8971	+1.0903	+32.9251	–	–	–
2	**T**	+4.0187	+4.5012	+0.8928	**+0.4063**	**+1878.0223**	**+374.0003**
3	**H**	+2.7811	+0.3074	+9.0464	**+1.0227e–04**	+7.1948	+1.7446
4	**C**	–8.0311	+6.0657	–1.3240	**+0.2337**	**+1664.8188**	**+679.1061**
8	**T*T**	–12.5237	+12.3239	–1.0162	**+0.3487**	**+1658.8945**	**+1762.5754**
9	**H*H**	–0.9727	+0.3746	–2.5967	**+0.0408**	+2.3882	+3.1681
10	**C*C**	–11.5943	+7.7070	–1.5044	**+0.1832**	**+497.1134**	**+1158.1287**
14	**T*H**	–6.4568	+1.4660	–4.4042	**+0.0045**	**+37.9860**	**+31.0309**
15	**T*C**	–26.9818	+21.0224	–1.2835	**+0.2467**	**+5108.2365**	**+6565.9067**
16	**H*C**	–3.7683	+1.6554	–2.2763	**+0.0631**	**+55.3008**	**+35.5951**

Fig. 8a BALFIT: Analysis of variance results for original regression model of acetylene data.

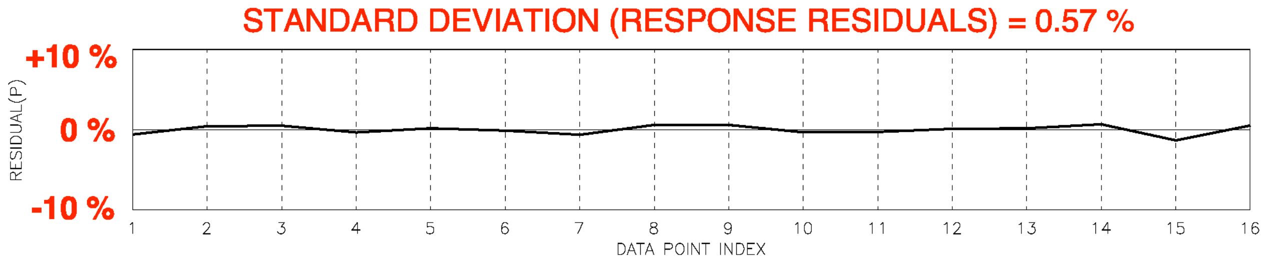


Fig. 8b BALFIT: Response residuals for original regression model of acetylene data.

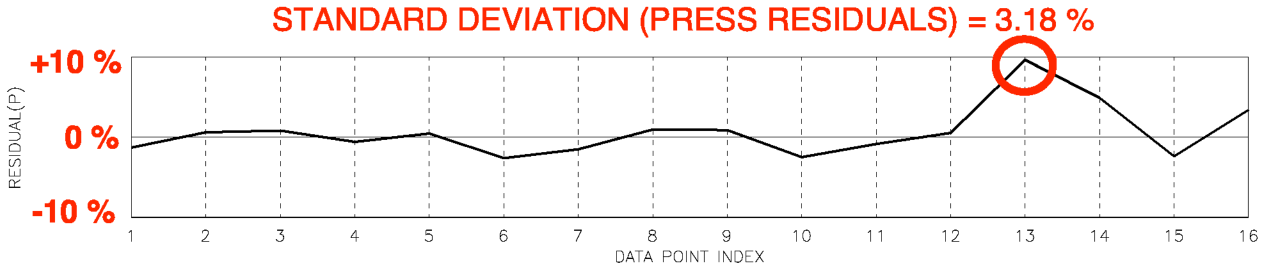


Fig. 8c BALFIT: PRESS residuals for original regression model of acetylene data.