The Role of Hierarchy in Response Surface Modeling of Wind Tunnel Data

Richard DeLoach*
NASA Langley Research Center, Hampton, Virginia, 23681

This paper is intended as a tutorial introduction to certain aspects of response surface modeling, for the experimentalist who has started to explore these methods as a means of improving productivity and quality in wind tunnel testing and other aerospace applications. A brief review of the productivity advantages of response surface modeling in aerospace research is followed by a description of the advantages of a common coding scheme that scales and centers independent variables. The benefits of model term reduction are reviewed. A constraint on model term reduction with coded factors is described in some detail, which requires such models to be "well-formulated", or "hierarchical". Examples illustrate the consequences of ignoring this constraint. The implication for automated regression model reduction procedures is discussed, and some opinions formed from the author's experience are offered on coding, model reduction, and hierarchy.

I. Introduction

A popular experimental method for aerospace ground testing, and especially for experimental aeronautics, suffers from a weakness that severely and adversely impacts productivity. Practitioners of this method, known as One Factor At a Time (OFAT) testing, attempt to obtain the information necessary to characterize how changes in test article responses depend on changes in various combinations of independent variable levels (also called factor levels). They do this typically by making all such changes of interest and recording for each the corresponding changes in a number of response variables.

The productivity issues associated with OFAT testing stem from the very large number of factor combinations that are likely to be of interest, the cost and time required to physically set each combination, and the fact that each factor combination is examined individually, one at a time. This latter, defining attribute of the method is due to a widely held but erroneous assumption that if multiple factors were to be changed simultaneously, the resulting response changes could not be accurately partitioned among the factors responsible for them. One would assume by this prevailing wisdom that if angle of attack and Mach number were both changed before the next data point in a wind tunnel test, for example, it would not be possible to know how much of the resulting change in forces and moments was due to the change in angle of attack and how much was due to the change in Mach number. This is true for an individual data point. However, if a suitably selected sample of simultaneous factor combinations is acquired, the partitioning of effects can in fact be accomplished, and rather easily, about which more presently.

Returning to the OFAT productivity question, consider a typical wind tunnel test for which there might be a half dozen independent variables, each to be set at 10 levels. This would not represent a particularly ambitious test design; in actual practice many tests are considerably more elaborate. Nonetheless, this test features a million (10^6) possible factor level combinations. While results vary over a wide range from test to test, the maximum volume of data that resource constraints permit in a wind tunnel test is typically measured in the thousands of points, not the millions. The result is that OFAT wind tunnel tests often generate information on only a few tenths of 1% of all the possible factor combinations.

Not all possible factor combinations are of interest, to be sure, and experienced OFAT practitioners rely upon a combination of skill and subject matter expertise to guide them in selecting subsets of the design space that are of particular interest to them. Nonetheless, the boundaries between design space regions of greater and lesser interest are often difficult to define clearly, especially in the most interesting cases in which relatively little a priori information exists about the test article. It seems unlikely that there would be little to learn from an examination of

* Senior Research Scientist, Aeronautical Systems Engineering Branch, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Associate Fellow.

99+ percent of the design space that must go unexplored in a typical OFAT wind tunnel test due to resource constraints.

The inefficiency of OFAT testing is responsible for a strong focus on data acquisition speed, since there is time to examine so little of the design space in a typical wind tunnel test. This need for speed forecloses options that would otherwise be available to implement quality assurance and quality assessment tactics such as replication, randomization, and blocking. The inherent inefficiency of OFAT testing therefore adversely impacts quality as well as productivity.

An integrated system of formal experiment design, execution, and analysis methods, which has been called the Modern Design of Experiments (MDOE), was introduced to the experimental aeronautics community of Langley Research Center in the mid-1990s to improve upon the quality and productivity weaknesses of OFAT testing. The MDOE method is an application of industrial experiment design methods for product and process improvement that began to emerge early in the 20th century[1-3], and have been adapted in small ways for the special requirements of aerospace ground testing. The basic principles of MDOE are documented in the references[4-7], and some representative examples of its application in aerospace research are also provided[8-19].

A key element of MDOE testing in certain broad classes of experiments is its reliance upon Response Surface Methods (RSM), also called Response Surface Modeling. Response surface models are mathematical relationships expressing various system responses of interest in terms of the independent variables (factors) that influence them. It is by developing such response models that the MDOE practitioner is able to determine how much of the change in some system response of interest is caused by changes in one independent variable, and how much is caused by changes in another, notwithstanding the fact that multiple factor changes may have been made simultaneously for each data point.

Productivity enhancements in response surface modeling derive from the fact that the experimenter is not required to acquire data at every site of interest within the design space in order to estimate responses at those sites. If one acquires the minimum volume of data necessary to adequately establish a response model by some mathematical fitting process such as regression, then system responses at intermediate design space sites can be predicted using the model. Assuming an adequate response model, it is not necessary to take the time or bear the expense of physically setting those intermediate factor combinations and measuring the corresponding responses, which gives this method a substantial productivity advantage over OFAT testing.

This advantage is leveraged by the fact that an MDOE test matrix features a relatively small number of measurements that entail simultaneous changes in multiple factor levels. It is precisely because so many factors are changed for each MDOE test point that so few points are required; each point contains much more information that a point for which only one factor is changed at a time, which enables the experimenter to traverse the whole design space with many fewer data points than an equivalent OFAT test.

The interpretation of response models can be enhanced, and in some cases the uncertainty associated with response model predictions can be reduced, by transformations of the independent variables that scale and center them, as discussed in Section II of this paper. Section III briefly outlines the rationale for further improving response predictions by selectively rejecting certain terms in the regression model. These methods can be influenced by a property of response models called hierarchy, as will be described in Section IV. There is a discussion of selected topics in Section V, and concluding remarks are provided in Section VI.

II. Factor Coding for Regression

As a practical matter, when we say that we are interested in how some response depends on a number of independent factors, we do not have in mind an unlimited range of factor levels. If we are conducting a wind tunnel test in which we will quantify forces and moments as a function of angle of attack (AoA) among other variables, for example, our interest is likely to focus on a limited range of AoA values. For a commercial jet transport, we might be concerned with angles in a range something like [-4°:+10°]. We would not likely examine response variables at, say, 90° in such a test, although other AoA ranges may be of interest in other tests.

In this example, we could describe the limited range of AoA values that interests us in terms of a mid-range value, α_0=3°, plus or minus an interval half-width, h_α=7°. To facilitate certain calculations, as well as the interpretation of the results they produce, it is convenient to transform the physical variables of an experiment into coded variables. For the angle of attack range illustrated here, we could use this simple coding transformation:

$$x_\alpha = \frac{\alpha - \alpha_0}{h_\alpha} = \frac{\alpha - 3}{7} \tag{1}$$

At the extremes of the range of interest, -4° and +10°, the coded variable assumes values of ±1, and has the value of zero in the middle of the range. All such variables coded in this way span the range of ±1 and are centered at zero. So, for example, if Mach number, M, is another variable in our test that ranges from 0.70 to 0.96, say (M = 0.83±0.13), then the corresponding coded Mach number variable would be

$$x_M = \frac{M - M_0}{h_M} = \frac{M - 0.83}{0.13} \tag{2}$$

Note that the coded variables are simply linear transformations of the physical variables, so the results of any calculation in terms of coded variables can be readily transformed back to physical units. However, the inverse transformation may be problematical unless the response model features a property known as hierarchy, as will be discussed in more detail in Section IV.

Coded variables have a number of advantages over physical variables in the generation of a response model. For example, when factors are expressed in physical units, the numerical values of the response model coefficients depend on which units are selected. Consider a study of shape memory alloy wing components that are deformed by heat from applied electrical power. It might be of interest to develop a mathematical relationship between certain lateral stability responses such as yawing moment and rolling moment, and the voltages applied to change the shape of various wing components. The coefficients of the response model regressors will vary substantially if factor levels originally expressed in volts are changed to kilovolts, for example. This is especially true for higher-order terms. The resulting ambiguity can make it difficult to assess the relative contribution of, say, first-order and second-order effects. On the other hand, if the factors are represented by coded variables, the regression coefficients are independent of the units in which the physical variables are expressed, and the relative impact of the various regressors on system response is much easier to see.

Equations (3) illustrate the distinction between coded and physical variables in a quadratic response model in two factors:

$$\begin{aligned} y_{coded} &= 542 - 152x_A - 73x_B + 308x_Ax_B - 207x_A^2 + 160x_B^2 \\ y_{physical} &= 1206 - 5A - 186B + 15AB - 10A^2 + 8B^2 \end{aligned} \tag{3}$$

The same response is expressed in terms of coded variables x_A and x_B in the first equation, and in the corresponding physical variables *A* and *B* in the second. It is clear that *over the ranges tested* (from -1 to +1 in the coded units), the "*A*" factor has about twice as much influence on the response as the "*B*" factor. The relative influence of *A* and *B* are completely obscured in the second equation, because the coefficients depend entirely on the physical units. Likewise, when the response model is expressed in coded units as in the first equation, it is clear that the interaction between the two factors dominates the first-order effects of either factor alone. When the response is modeled in terms of physical factors, this inference is obscured by the dependence of the regression coefficients on the physical units. Completely different inferences could be drawn by expressing the factors in different units. Note also that in coded units it is clear that curvature effects for variable *B* dominate first-order effects, while the model expressed in physical units make this comparison much less clear.

Another advantage of the factor coding transformation illustrated in Eq. (1) and Eq. (2) is that it ensures that the factor range includes zero. This imparts a useful physical interpretation to the intercept of the response model. When zero is within the range of all fitted factors, the intercept of the response model is equal to the average of all the fitted response measurements. This has certain advantages in quantifying block effects, for example.

Consider a sample of data acquired over factor ranges that include zero, and assume that a response model has been constructed in terms of those factors. Now imagine that this sample of data is replicated at some later time, and the same response model is fitted to the second set of data. The coefficients, including the y-intercept, would be expected to differ slightly due to ordinary chance variations in the data, but no statistically significant difference would be anticipated between corresponding regression coefficients absent changes in the measurement environment that are not postulated in this example, or changes in the test article that violate the assumption of a replicated data sample.

Imagine, however, that a statistically significant change is in fact observed in the y-intercepts. That is, assume that the second y-intercept differs from the first to a greater degree than can be attributed to ordinary random error. This implies that the sample mean has shifted from one data sample to the next.

It is common in response surface modeling to organize test matrices in intervals or *blocks* of time. Systematic response shift between blocks such as described here are called *block effects.* Block effects are not uncommon in wind tunnel testing, and often contribute substantially more to the unexplained variance of a sample of data than the random component that is more widely recognized and understood.

Block effects are evidence of systematic changes occurring in the measurement environment over time. In a wind tunnel test, these can be due to instrument drift, temperature effects, operator fatigue and learning effects, geometric changes in the test section or test article, progressive changes in sting bending under applied aerodynamic load, and any of a myriad number of other systematic error sources that conspire against the state of statistical control that is generally assumed in an OFAT wind tunnel test. Because coding the independent variables so that factor ranges include zero makes the y-intercept of fitted regression models equal to the mean of all fitted points, systematic error effects can be detected from trends in the intercept time histories of replicated samples fitted in different blocks of time. Figure 1, presented by the author in Ref. 20 and explained in more detail there, displays a time series of differential y-intercepts for ten ostensibly identical lift polars acquired over a period of 2–3 weeks. Each point represents the difference between the intercept of one polar and the mean of all ten polars. Error bars on each data point indicate the degree of random error. Clearly, a trend of progressively increasing estimates of lift is observed over time, with this systematic error dominating the random error in this test.

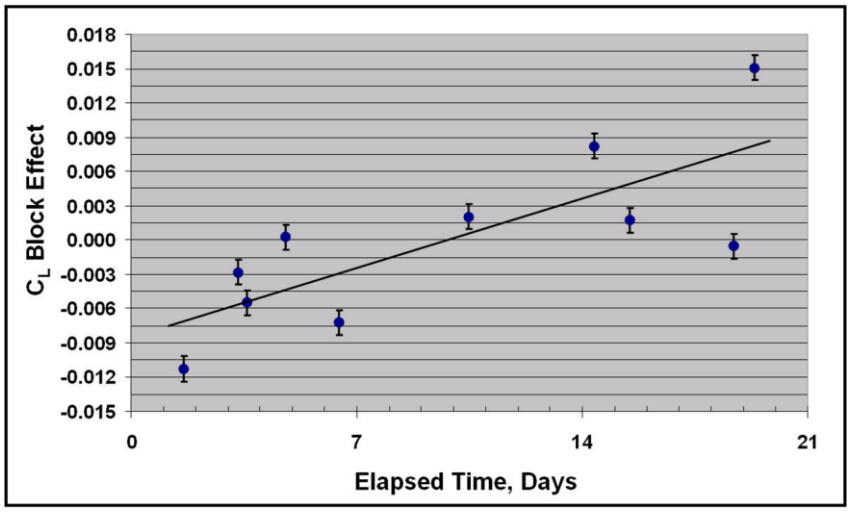


Figure 1. Block effects reveal unexplained systematic variation over time.

The coding transformation described in this section is desirable for response surface modeling from another point of view: it decouples slope and intercept effects. That is, least-squares estimators of the intercept are uncorrelated with the slope estimators, so changes in the slopes of the response surface that induce shape changes do not affect the intercept of the response model and conversely. This ensures that the functional form of the response model is decoupled from problems associated with precisely determining the intercept. Bias errors therefore affect only the intercept, and not the details of how responses depend on the independent variables. Marquardt and Snee[21] describe this result of centering independent variables through coding transformations by saying that it reduces "nonessential ill-conditioning." This then reduces the inflation in variance estimates associated with individual regression coefficients, which improves model prediction, especially at off-design points within the design space[22].

Table 1 provides an example of how the coding transformation illustrated in Eqs. (1) and (2) reduces "nonessential ill-conditioning" for a simple two-level, two-factor experiment to acquire data capable of fitting the following response model in coded factors:

$$y = b_0 + b_1 x_\alpha + b_2 x_M + b_{12} x_\alpha x_M \tag{4}$$

The coded factors x_α and x_M correspond to angles of attack in the range of -4° to +10° and Mach numbers from 0.70 to 0.96 in this example, just as in Eqs. (1) and (2), and the *b*'s are regression coefficients. A similar response model (with different coefficients) could be fitted to the variables in physical units, of course.

For both coded and physical units the factor levels are listed in Table 1 along with the cross-products of each, and the cross products are summed. If the columns of factor levels are regarded as vectors, the sum of cross-products

is proportional to the cosine of the angle between them. Note that for the coded units, the sum of cross-products is zero, implying vectors that are at right angles, or orthogonal. The vectors of the factors in physical units are not orthogonal. This means that if a model is fitted in terms of physical units, the value of the AoA coefficient will depend upon whether Mach is in the model or not (and conversely), but if the model is fitted in terms of coded units, the value of the x_α coefficient will be the same whether x_M is retained or not, and conversely.

When the regressors are not orthogonal as when the response is modeled in terms of factors expressed in physical units, each estimated regression coefficient can be a function of some linear combination of the true coefficients of more than one regressor. Orthogonality is desirable because it ensures that the magnitude of a given coefficient is independent of other terms in the model. This makes the interpretation of regressor effects independent of other terms, which is helpful in understanding the underlying physics. Orthogonality is an especially desirable property when terms are rejected from the model to reduce prediction uncertainty, as will be discussed shortly. The coefficients of regressors that are retained in an orthogonal model are not influenced by decisions to retain or reject other terms in an orthogonal model.

Table 1. Orthogonality of coded units

Point	Physical			Coded		
	AoA, α	Mach, M	α x M	x_α	x_M	x_α x x_M
1	-4	0.70	**-2.80**	-1	-1	**-1**
2	-4	0.96	**-3.84**	-1	+1	**-1**
3	+10	0.70	**7.00**	+1	-1	**-1**
4	+10	0.96	**9.60**	+1	+1	**+1**
		Sum:	**$9.96 \neq 0$**		**Sum:**	**0**

We close this section with a mention of two additional advantages of the coding transformation discussed here. While less a problem with today's computers than in the past, mapping each factor into a dimensionless range from -1 to +1 can avoid certain computational errors related to a computer's necessarily finite resolution. These problems can arise when different factors are of extremely different magnitudes when expressed in physical units. Consider a wind tunnel test in which forces and moments are modeled as a function of Reynolds number per foot (typically in the millions), and some measure of surface roughness that might be expressed in micrometers. Regression calculations involve differences in sums of squared values, and finite resolution limits could introduce errors in the calculation of differences between the squares of very large and very small numbers. Scaling all factors to a common range of -1 to +1 addresses this problem.

Decoupling slope and intercept effects by coding the independent factors has one further computational advantage. Montgomery, Peck, and Vining[23] show how it simplifies calculations of the confidence interval associated with model predictions.

Altogether, there are a number of advantages to coding the factors in such a way as to scale and center them in a restricted range that includes zero. However, coding the factors in this way does present some complications when the models are reduced by rejected certain terms, due to the small size of their regression coefficients, for example. We will discuss those complications in Section IV, after first explaining why such model reductions are desirable.

III. Regression Model Reduction

The uncertainty associated with a response model prediction depends on the volume of data used to fit the model, the number of coefficients in the model, the intrinsic variability of the measurement environment, and the location in the design space where the prediction is made (i.e., the combination of independent variable levels). Once the experiment has been executed, however, the only one of these four factors that can influence prediction uncertainty at a given site in the design space is *p*, the number of parameters in the model (including the intercept). We now show that the quality of a response surface model can be improved by eliminating some of the terms in the model.

Rationale for Reducing the Number of Model Terms

We begin with a brief review of response modeling basics, following a condensed version of the appendix in DeLoach and Ulbrich [2007][24]. The general form of a full polynomial model in *K* factors is as follows:

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i \tag{5a}$$

where y_i is the response recorded for the i^{th} data point, x_{ij} is the i^{th} level of the j^{th} regressor, β_j is the coefficient of the j^{th} regressor, and ε_i is an error term, assumed to be drawn from a normal distribution with a mean of 0 and with a constant standard deviation for all responses. The quantity β_0 is the intercept term.

Equation (5a) can be described more succinctly in vector/matrix form as follows:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{5b}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of response measurements, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of coefficients, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of error terms.

$\mathbf{X}$ is the design matrix, consisting of n rows corresponding to the number of data points fitted to the response model, and p columns, one for each term in the model, including the intercept term.

Consider a vector $\mathbf{x}_0 = [1\ x_{01}\ x_{02}\ \ldots\ x_{0K}]'$ representing a data point specified by a given combination of factor settings, where x_{0i} is the level of the i^{th} regressor corresponding to this point. The estimated mean response at this point is

$$\hat{\mathbf{y}}\ \ \mathbf{x_0}\ \ = \mathbf{x_0'b} \tag{6}$$

where $\mathbf{b}$ is a vector of estimated regression coefficients. That is, $\mathbf{b}$ is a best estimate (typically by some least-squares criterion) of the vector of true coefficients, β, in Eq. (5b). The variance in the response prediction at a particular $\mathbf{x}_0$ is computed as follows:

$$\mathrm{Var}\left[\hat{\mathbf{y}}\ \ \mathbf{x}_0\ \right] = \sigma^2 \mathbf{x_0'}\ \ \mathbf{X'X}^{\ -1}\mathbf{x_0} \tag{7}$$

Assume for a moment that Eq. (5) represents the largest response model that can be supported by a given test matrix. That is, we assume that this equation describes the highest-order response model for which non-singular regression results can be obtained for the prescribed test matrix.

We wish to examine the consequences of reducing the model so that fewer than the K regressors of the model described by Eq. (5) are retained. Let r represent the number of regressors that we wish to reject and let p represent the number of terms that will be retained in the model, including the intercept term. We can then express Eq. (5) as follows:

$$\mathbf{y} = \mathbf{X}_p\beta_p + \mathbf{X}_r\beta_r + \varepsilon \tag{8}$$

Here, $\mathbf{X}_\mathrm{p}$ is a $(p \times n)$ matrix with columns corresponding to the retained terms in the model, including the intercept, and β_p is a $(1 \times p)$ vector of the corresponding regression coefficients for this reduced model. The columns of $\mathbf{X}_\mathrm{r}$ represent terms that are deleted from the model, and β_r is a vector of the corresponding regression coefficients.

If $\mathbf{b}$ is a vector of estimated regression coefficients for the unreduced model, and $\mathbf{b}_\mathrm{p}$ corresponds to those coefficients that are retained, it can be shown that the matrix $Var(\mathbf{b}_\mathrm{p}) - Var(\boldsymbol{\beta}_p)$ is positive semidefinite.[25] Therefore, dropping terms from the full model and refitting the data to a subset of the original regressors results in model coefficient estimates with variance that is less than or equal to the variance in the corresponding coefficients of the full model. In other words, with respect to the precision of the regression coefficient estimates there is nothing to lose, and possibly something to gain, by reducing the number of regressors in the math model.

Consider now the impact of such a model reduction on the variance of response predictions. Note that a vector of predicted responses for each point in the test matrix can be generated from the vector of measured responses, $\mathbf{y}$, by means of the "hat matrix," $\mathbf{H}$, as follows:

$$\hat{\mathbf{y}} = \mathbf{Hy} \tag{9}$$

where

$$\mathbf{H} = \mathbf{X}\ \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}' \tag{10}$$

and **X** is the design matrix, as before. The variance in the vector of response estimates is computed as follows:

$$\mathbf{Var}\ \hat{\mathbf{y}}\ = \mathbf{H}'\mathbf{Var}\ \mathbf{y}\ \mathbf{H} = \mathbf{H}'\ \mathbf{I}\sigma^2\ \mathbf{H} \tag{11}$$

The hat matrix is both symmetric (equal to its transpose) and idempotent, meaning that **HH** = **H**. Equation (11) therefore reduces to

$$\mathbf{Var}\ \hat{\mathbf{y}}\ = \mathbf{H}\sigma^2 \tag{12}$$

Note that the variance of the i^{th} response prediction is just the i^{th} diagonal element of $\mathbf{H}\sigma^2$. Following Box and Draper[26], we consider the trace of this matrix, which is just the sum of all the diagonal elements:

$$\text{trace}\ \mathbf{H}\sigma^2\ = \sigma^2\text{trace}\left[\mathbf{X}\ \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\right] = \sum_{i=1}^{n}\text{Var}\ \hat{\mathbf{y}}_i \tag{13}$$

We invoke the following matrix identity: trace(**AB**) = trace(**BA**). Let **A** = **X** and **B** = (**X'X**)$^{-1}$**X'**. Then

$$\text{trace}\left[\mathbf{X}\ \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\right] = \text{trace}\left[\ \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{X}\right] = \text{trace}\ \mathbf{I}_p\ = p \tag{14}$$

since p is the dimension of the square matrix (**X'X**)$^{-1}$ as noted above in the description of the covariance matrix. Combining Eqs. (13) and (14) we have

$$\sum_{i=1}^{n}\text{Var}\ \hat{\mathbf{y}}_i\ = p\sigma^2 \rightarrow \frac{\sum_{i=1}^{n}\text{Var}\ \hat{\mathbf{y}}_i}{n} = \frac{p\sigma^2}{n} \tag{15}$$

That is, for any order polynomial model, the prediction variance averaged over all points in a regression analysis is proportional to the term count in that model. Obviously, there is some potential to reduce the average prediction variance simply by reducing the number of terms in the model.

Each term in the model carries with it some contribution to the variance that is explained by the model, and some contribution to the residual variance that remains unexplained by the model. The prediction uncertainty of the model can be reduced by discarding terms with contributions to the explained variance that are too small to justify given their contributions to the unexplained variance. That is, the overall "signal to noise ratio" of the response model can be enhanced by judiciously discarding selected terms in the model. Ways and means are outlined briefly in the next subsection.

Model Reduction Methods

Three common model reduction methods will be described briefly here. They are known as "Forward Selection," "Backwards Elimination," and "Stepwise Regression." Each method assumes some initial model that is to be reduced, which consists of all the terms in the unreduced model. So, for example, if these methods were to be applied to a quadratic calibration model for a six-component force and moment balance, the initial model would contain all 28 terms of a 2$^{\text{nd}}$-order response model in six factors.

1. Forward Selection

If the forward selection method is used, an initial model is constructed consisting of a constant term only, which is equal to the mean of all the regression data. The variance of the regression data sample about this mean is

computed, and since at this stage there are no independent variables in the model, all of this variance is by definition unexplained by any of the factors under study.

The one regressor from the original candidate model that has the greatest correlation with the response is then provisionally added to the model and regression coefficients for the resulting two-term model are computed. An analysis of variance (ANOVA) is conducted to quantify how much of the total variance in the regression data is explained by this two term model and how much remains unexplained. The total variance is computed from residuals consisting of differences between measured responses and the sample mean. This was done in the previous step. The explained variance is computed from residuals consisting of differences between model-predicted responses and the sample mean (predictions made with the current response model for all factor combinations in the regression data sample).

If the ratio of the variance explained by the current model to the variance explained by the previous model is large enough to distinguish the two variance estimates with a suitably high level of confidence, the new term is assumed to have explained a sufficiently large portion of the total variance to be retained in the model. Equivalently, the variance that is unexplained by the older and newer models can be compared to determine if there is a significant difference. If so, the most recent term is retained in the model and a new term is provisionally selected from among those remaining to evaluate. The new term is the one that, when combined with terms already in the model, makes the largest incremental increase in explained variance compared to the previous model. If this increase is large enough to detect with a prescribed degree of confidence, this term is retained, and the process continues until finally the most influential new term increases the explained variance by a degree that cannot be detected with the requisite level of confidence. At that point, the last ineffective term is discarded and the process stops, with a reduced model featuring all of the terms that have been added at this point in the process. If every provisional term increases the explained variance significantly, then all terms are retained and the term count for the model is not reduced.

2. Backward Elimination

Backward elimination is similar to forward selection except that the process attempts to identify terms for the final model by working in the opposite direction. The backward elimination method begins with all terms from the full model included. The term with the weakest correlation with the response is provisionally rejected, and the impact on the explained variance of the model is assessed as with the Forward Selection process. If rejecting this term causes a significant reduction in explained variance, it is retained and the process stops. Otherwise, the process continues until no terms in the model can be rejected without causing a significant reduction in the variance explained by the model, or until the only remaining term is the intercept.

3. Stepwise Regression

Stepwise regression is a combination of forward selection and backward elimination. The process begins as a forward selection process and continues until the model contains the intercept and two regressors. Backward elimination is then applied to the three-term model, provisionally eliminating each regressor in turn to assess the corresponding reduction in the model's explained variance, starting with the one that has the weakest correlation with the response. If the rejection of any term causes an insignificant reduction in explained variance, it is rejected. Otherwise, it is placed back in the model and the forward selection process is resumed with whichever remaining candidate term causes the most significant increase in the explained variance. If such a term increases the explained variance significantly, it is retained and backward elimination is again initiated on the new model. The process continues until no candidate regressors increase the explained variance significantly upon entry and none already in the model are so weak that they can be eliminated with no significant effect.

If two regressors are highly correlated, adding one of them to the model may render the first one superfluous, and therefore susceptible to rejection in the backward elimination phase. The backward elimination component of stepwise regression provides some protection against multicollinearity in this way, although it does not guarantee that the resulting model will be completely free of it.

It is common to apply more than one stepwise procedure to the same data set. For example, the author often follows a common convention of applying backward elimination first, to give all model terms a chance to be included. He then applies stepwise regression to the surviving model terms, and finishes with one more application of backward elimination to reject high-order terms that might survive the first two steps without contributing significantly to the model. The model is then manually inspected for terms that might still be candidates for rejection based on multicollinearity metrics, coefficient magnitudes, or other factors, including experience and subject matter expertise. Each candidate term is provisionally eliminated, and changes in the explained variance are assessed as well as numerous other quality assessment metrics.

IV. The Role of Hierarchy

To recap, we have made the case for certain advantages that accrue from fitting experimental data to factors that have been coded by scaling and centering them on intervals that contain zero. We have also demonstrated that the quality of the response model can be enhanced by eliminating unnecessary terms, thereby reducing the uncertainty in predictions that are made with the model. In this section we discuss a potential conflict between these two procedures. That is, we review how coding the factors can place certain constraints on which ones can be eliminated from a response model without certain unintended consequences that will be described.

Eq. (4), presented earlier, is a factor-interaction model in coded variables x_α and x_M that is reproduced here for convenience, with a subscript on y to emphasize that the factors are coded:

$$y_{coded} = b_0 + b_1 x_\alpha + b_2 x_M + b_{12} x_\alpha x_M \tag{4}$$

We are able to infer a number of things about the behavior of the response just from the form of this model. For example, we can tell that the response surface tilts, with mean slopes in the x_α and x_M directions dictated by the magnitudes of b_1 and b_2. We can also infer something about the relative strengths of the angle of attack and Mach number dependences *over the ranges tested*, from the magnitudes of these coefficients. We can tell that angle of attack and Mach number are not independent in this case. That is, the effect that a change in angle of attack has on the response depends on the Mach number, and conversely. We can see this explicitly by rearranging the terms in Eq. (4) as follows:

$$y_{coded} = b_0 + \left(b_1 + b_{12} x_M \right) x_\alpha + b_2 x_M \tag{16}$$

The response can be regarded as a first-order function of x_α and x_M, with a slope for x_α that changes as a linear function of x_M. The response surface is therefore a tilted plane that is twisted, with the slope of y_{coded} vs. x_α changing as a linear function of Mach number. Collecting terms in x_M reveals symmetrical behavior; the slope of y_{coded} vs. x_M also changes, and as a linear function of angle of attack.

Considerable insight can often be gained by a consideration of the response function's geometry, even (one might say *especially*) when the response model is more elaborate than in this simple example. One can usually tell at a glance which factors interact with which other factors and which are independent, what the relative strengths are of the interactions, which first-order responses dominate, where there is curvature, whether the degree of response curvature attributable to changes in one factor is a function of the level of another factor, whether the rate of curvature is changing and where, etc, etc.

These insights can inform further stages of the investigation. For example, if two factors are found to be independent and resources are constrained, then one can justify independent investigations of the two factors in two separate tests, with the second test postponed to save current resources. However, if a strong interaction exists between the two factors, then a study of how the system response variables depend on one of them while the other is held constant is likely to produce results of a meager and unsatisfactory kind. Everything that is learned about the factor that is varied will apply only to the case in which the second variable is at the specific level to which it was held constant during the first test. If there is a strong interaction, the behavior of the variable that was examined initially could be entirely different in another test, in which the second factor is set at another level. We conclude that it is useful to know about the basic functional form of the model.

Now insert Eqs. (1) and (2) into Eq. (4) to express the model in physical variables, α and M. After gathering terms:

$$y_{physical} = \beta_0 + \beta_1 \alpha + \beta_2 M + \beta_{12} \alpha M \tag{17}$$

where

$$\begin{aligned}
\beta_0 &= b_0 - \frac{\alpha_0 b_1}{h_\alpha} - \frac{M_0 b_2}{h_M} + \frac{\alpha_0 M_0 b_{12}}{h_\alpha h_M} \\
\beta_1 &= \frac{1}{h_\alpha}\left(b_1 - \frac{M_0 b_{12}}{h_M} \right) \\
\beta_2 &= \frac{1}{h_\alpha}\left(b_2 - \frac{\alpha_0 b_{12}}{h_\alpha} \right) \\
\beta_{12} &= \frac{b_{12}}{h_\alpha h_M}
\end{aligned} \tag{18}$$

We now ask this question: What if the first-order Mach-number term was discarded from Eq. (4) through some model selection procedure such as backward elimination or stepwise regression, or for other reasons? This is the equivalent of declaring b_2 to be zero in Eq. (4). If we then convert the response model to physical units, does the model predict the same general behavior as it did in coded units? That is, in physical units would we still see a first-order AoA term, an interaction term, and no first order Mach term? The answer, surprisingly, is "no".

From Eq. (18) it is clear that even if $b_2 = 0$, *both* linear terms appear in the model after it is converted back to physical units, because even when $b_2 = 0$, $\beta_2 \neq 0$. So in physical units the model retains the first-order Mach term that was dropped when the model was expressed in coded units. In other words, the model *reacquired* the first-order Mach term that had been deleted, simply by the act of converting from coded units back to physical units! This is unsettling for anyone wishing to extract fundamental insights about the system's behavior from an examination of the form of the response model.

Consider now another example. The following provisional model for lift coefficient as a function of the angles of attack and sideslip was fitted from data acquired in an early wind tunnel test of approach and landing characteristics for a proposed Space Shuttle replacement vehicle:

$$C_{L,\,coded} = 538.4 + 91.0A + 2.0B + 3.2B^2 \tag{19}$$

A and B are the angles of attack and sideslip expressed in coded units. The coding of independent variables in Eq. (19) clearly reveals that changes in angle of attack have a much greater impact on lift than changes in angle of sideslip over the limited factor ranges tested, a result that was not unanticipated.

The principal investigator rejected this model, arguing that while the quadratic dependence of lift on sideslip angle is consistent with aerodynamic first principles, symmetry considerations imply that there must be no linear sideslip term in the model, and that the relatively small linear sideslip coefficient must therefore be attributable to experimental error. The data used to fit Eq. (19) were therefore refitted to the following reduced model, in the form known from subject-matter expertise to be correct:

$$C_{L,\,coded} = b_0 + b_1 A + b_{22} B^2 \tag{20}$$

This model features the quadratic sideslip dependence that is anticipated, but does not predict the linear sideslip dependence known by symmetry considerations to be artificial.

It is instructive to examine what happens when this model is converted from coded to physical units. Let us assume that the angles of attack and sideslip were coded analogously to Eqs (1) and (2), as follows:

$$\begin{aligned}
A &= \frac{\alpha - \alpha_0}{h_\alpha} \\
B &= \frac{\beta - \beta_0}{h_\beta}
\end{aligned} \tag{21}$$

where in this case β represents sideslip angle, not to be confused with the coefficients of Eqs. (17) and (18). The constants α_0 and β_0 are the physical angles of attack and sideslip, respectively, about which these factors are centered, and h_α and h_β are scaling constants that map each physical variable into a range from -1 to +1.

After inserting Eqs. (21) into Eq. (20) and gathering terms:

$$C_{L,\,physical} = b'_0 + b'_1\alpha + b'_2\beta + b'_{12}\alpha\beta \tag{22}$$

where

$$\begin{aligned} b'_0 &= b_0 - \frac{\alpha_0 b_1}{h_\alpha} + \frac{\beta_0^2 b_{12}}{h_\beta^2} \\ b'_1 &= \frac{b_1}{h_\alpha} \\ b'_2 &= -\frac{2\beta_0 b_{22}}{h_\beta^2} \\ b'_{12} &= \frac{b_{22}}{h_\beta^2} \end{aligned} \tag{22}$$

Note the curious behavior with respect to the linear sideslip term. After a simple linear transformation from coded units back to physical units, the linear sideslip term has reappeared in the model, despite the fact that we discarded it from the coded-units model. Eq. 22 predicts in physical units a linear sideslip dependence believed to be nonexistent from physical first principals, and therefore rejected from the original model in coded units.

The two examples treated here illustrate one of the negative consequences of eliminating what are called "hierarchically inferior terms" from a polynomial model, resulting in a class of model first described by Kempthorne[27] as not "well-formulated". One polynomial term is defined as hierarchically inferior to another if the other term can be constructed by multiplying the inferior term by another polynomial term. For example, in the model of Eq. (19), the linear *B* term is hierarchically inferior to the quadratic *B* term because the quadratic term can be constructed by multiplying the linear term by another polynomial term (namely, in this case, itself). In other words, hierarchically inferior terms are the "components" or "building blocks" from which higher-order terms in the model are constructed. To be considered well formulated, all of the inferior terms of every term in the model must be present. If a compound term is included to preserve hierarchy, all of the component terms of *that* term must also be present. Eq. (20) is not well formulated because B^2 is in the model, but the *B* sub-element of that term is not. Likewise, Eq. (4), which contains a two-way interaction term, is only well-formulated when *both* of the first-order terms comprising that interaction are also present.

The columns in the design matrix, **X**, of Eq. (5b) comprise what is called the "estimation space" of that matrix[28]. Peixoto[29] shows that the estimation space of a polynomial regression model is invariant under coding transformations if and only if it includes all hierarchically inferior terms and is therefore well-formulated. It is because Eq. (20) is not well-formulated that under coding transformation the estimation space of the design matrix changed. There is one column in the design matrix for every term in the response model, and a coding transformation converted the three-column estimation space of the design matrix for Eq. (20) into the four-column estimation space of the design matrix for Eq. (22).

Likewise, from Eqs. (4), (17), and (18) it is clear that reducing a four-term response model in coded units (Eq. (4)) to a *three*-term model by setting to zero the coefficients of either the hierarchically inferior linear Mach term (b_2) or linear alpha term (b_1), produces a *four*-term model in physical units with linear terms for both variables. This is because, by Eqs. (18), setting b_1 and b_2 to zero does not drive β_1 or β_2 to zero. Note, however, that if one dropped the two-way interaction term from Eq. (4) the estimation space would remain unchanged. That is because, by Eq. (18), setting b_{12} to zero drives β_{12} (and only β_{12}) to zero, resulting in the same model regressors in coded and physical units. Only the values of the coefficients would be different. The estimation space remains unchanged because the interaction term is not hierarchically inferior to any other terms in the model.

Another consequence of rejecting hierarchically inferior model terms is that various goodness of fit metrics are affected by coding transformations if the model is not well-formulated, complicating the adequacy assessment of response models. Peixoto[29] gives examples in which third-order models of average January minimum temperatures as a function of North-American longitudes and latitudes are reduced by eliminating hierarchically inferior terms. He considers two such models with regressors (in addition to intercept terms) of [a, ab, b^3] for one model and [ab, b^3, ab^3] for the other, where a and b are latitude and longitude, respectively. Note that these two models are in fact missing certain hierarchically inferior terms. The first model lacks the linear b term required by its ab and b^3 terms, and it lacks the quadratic b term required by its b^3 term. The second model also requires linear and quadratic b terms to be well-formulated, as well as a linear a term.

Peixoto computed R^2 coefficients of determination for both models before and after applying a coding transformation that centered the longitude variable on 91° west longitude, a coding tactic simply intended to reduce "nonessential ill-conditioning" in the Marquardt and Snee[21] sense described in Section II. He noted that this change was the equivalent of choosing the origin of longitude measurements to pass near St. Louis rather than through Greenwich, which was not expected to have any effect on the functional form of a model describing temperature as a function of coordinate space. Nonetheless, the functional form of the models did change with this transformation, just as in the two examples given earlier.

Furthermore, Peixoto reports an R^2 goodness-of-fit statistic for his first temperature model of 0.776 before the coding transformation and 0.919 afterwards, while R^2 for the second model was 0.921 before the transformation and 0.124 afterwards. The coefficient of determination, R^2, represents the fraction of the total sum of squares that is explained by a given model. This would not be expected to be a function of the units selected for the independent variables of the model. When the same coding transformation was applied to other models that were well-formulated (no missing hierarchically inferior terms), Piexoto reported that the R^2 statistic was the same before and after the coding transformation. He concluded that this and other goodness of fit metrics (he cites Mean Square Error and Mallow's C_p statistic[30] specifically) may be artificially raised or lowered by the origin shifts in a coding transformation if the model is not well-formulated.

V. Discussion

Three specific questions are addressed in this paper:

1) Should the factors in a polynomial regression model be coded? That is, are there benefits to imposing a linear scaling and translation transformation to the independent variables that exceed whatever costs are inherent in doing this?
2) Should the terms in a response model be examined with a view to discarding some of them?
3) Is it necessary to respect hierarchy?

The author has stated his case for answering these three questions in the affirmative. Key points that influence this view are emphasized here.

A. Rationale for Coding, Model Reduction, and Hierarchy

The decoupling of slope and intercept terms that results from coding the independent variables is particularly useful when significant block effects are in play. Block effects, representing systematic bias shifts across data samples acquired in different intervals (or "blocks") of time, are attributable to slowly varying state changes within the test environment that persist for prolonged periods. Such effects may occur more often than they are recognized, and in fact are quite common in wind tunnel testing and other applications, where stringent precision requirements can amplify the consequences of any failure to cope with them effectively.

An assumption that block effects and factor effects are independent is usually justified; one would expect a one-degree *change* in angle of attack to cause the same *change* in pitching-moment whether the data were acquired on Monday or on Tuesday, for example. This is notwithstanding the fact that *absolute levels* of pitching moment might be different on two different days, even if the angle of attack and all other factors were held constant, due to block effects.

The decoupling of intercept and regressor effects that coding facilitates is useful for isolating and detecting block effects because the intercepts used to quantify them are then model-independent. Fitting errors therefore do not complicate the estimation of block effects when the intercept is independent of other regressors.

Specific examples were used in this paper to illustrate the impact of coding transformations on the estimation space of models with rejected terms that are hierarchically inferior to terms retained in a model. Likewise, the impact on goodness of fit was presented for specific models only. However, this behavior characterizes polynomial

models of arbitrary order in any number of factors, which suggests the importance of maintaining hierarchy during the reduction of regression polynomials generally when the factors have been coded.

Even if factors are not coded, a hierarchical model will facilitate coding in the future. This would be desirable if, for example, one might later decide to standardize the regression coefficients by applying a coding transformation to the factors, to more clearly assess how each model term contributes to the explained variance. One might also later decide to invoke a factor coding transformation to reduce the variance inflation in estimates of the regression coefficients. In such cases, the estimation space of the response model can only be expected to remain invariant under coding if hierarchy has been respected in the original response model.

Polynomial term reduction procedures such as those outlined in Section III do not take hierarchy into account, and often generate models that are not well-formulated. For this reason, Draper and Smith[30] suggest that models generated by automated selection procedures should be reviewed and refined to ensure that they are hierarchically well-formulated.

At least one commercial software package (Minitab®[31]) will not even perform an analysis of variance on nonhierarchical models, because the residual sum of squares of a nonhierarchical model includes components due to the missing hierarchically inferior terms. This makes the residual variance less representative of the true experimental error.

B. Two Common Arguments against Imposing Hierarchy

Two practical arguments are commonly offered against imposing hierarchy, notwithstanding the advantages noted above (many of which may simply be unrecognized by those who do not place a high priority on hierarchy). The first is that it can be regarded as difficult to automate. The second is that it forecloses opportunities to reduce prediction variance by rejecting (hierarchically inferior) terms per Eq. (15).

1. Implementation Options

Regarding the first argument, Peixoto has published an algorithm[32] which, for a given model in any number of variables, generates all possible model subsets that respect hierarchy. This algorithm can be included in any software system designed to automate a response surface modeling analysis. As another alternative, Design Expert®[33] is a commercially available software package that offers to impose hierarchy automatically in cases where hierarchically inferior terms are missing from a proposed response model. If the user declines, the software provides a warning and a second prompt. If the user still declines the option to make the model hierarchical, Design Expert analyzes the nonhierarchical model with this disclaimer:

> "Using this non-hierarchical polynomial regression model (it excludes hierarchically inferior terms) is not recommended. Measures of goodness of fit and the predicted response values may not be the same as those from the coded equation. All analysis within Design-Expert software is based on the coded equation."

While Design Expert does permit one to perform an analysis of variance on nonhierarchical models (under duress!), at least one commercially available software package[31], as noted above, will not perform such an analysis on nonhierarchical models under any circumstances. Nonetheless, these software packages are representative of available ways and means to automate the generation of hierarchically well-formulated polynomial response models.

2. Adverse Impact on Prediction Variance

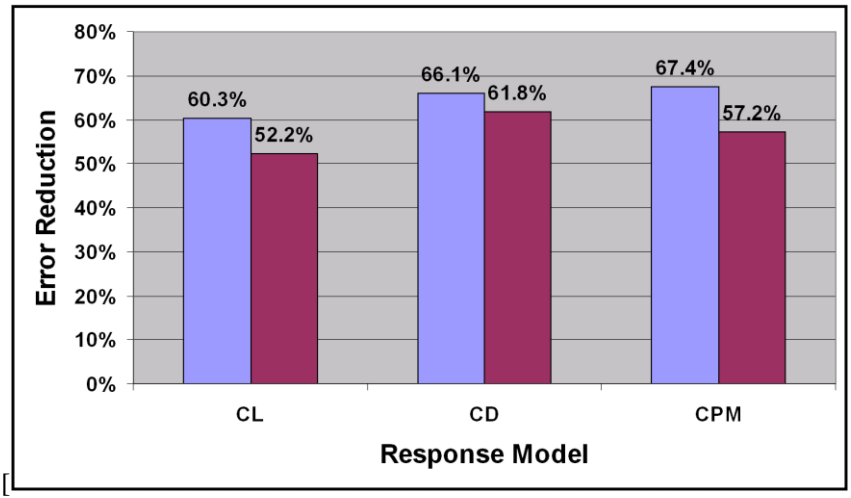
The second practical argument against imposing hierarchy is that reductions in average prediction variance can be achieved by reducing the term count, as described in Section III. Proponents of this argument feel that it is a poor cost/benefit trade to retain a term, especially an insignificant one, for no other reason than to respect hierarchy, and in the process to foreclose an option to achieve some tangible reduction in prediction variance. We have discussed how the benefits of maintaining hierarchy may be underappreciated. We now make the case that the costs of maintaining hierarchy may be overstated.

Experience in wind tunnel tests and other aerospace response surface modeling applications suggests that most of the quality improvement to be derived from a reduction in model terms occurs as a result of eliminating all of the insignificant terms that do *not* conflict with hierarchy. Relatively small incremental improvements result from further eliminating hierarchically inferior terms. The reason is something called "the scarcity of effects principle", which states that higher-order terms are less likely to be significant than lower-order terms in a response model.

A polynomial response model can be regarded as a truncated Taylor series approximation to the true but unknown underlying functional relationship between a response variable of interest and the factors upon which it depends. Higher-order terms in such a series generally explain progressively less and less of the response by the scarcity of effects principle, so that beyond a certain order model they can be ignored altogether with no significant effect on the predictive capability of the model. The absence of significant higher-order terms frees lower order terms to be deleted from the model without violating hierarchy. Therefore, when higher-order terms are not present, model reduction procedures are more likely to target terms for deletion that are not hierarchically inferior than those

that are. Only terms that are "second-order inferior" are targeted for deletion; that is, terms that are "inferior to terms that are inferior" to higher-order terms that have already been rejected.

Figure 2 illustrates how much of the quality improvement to be derived from a reduction in model term count is gained from eliminating terms that are not hierarchically inferior. This figure describes a representative wind tunnel test in which the full, non-reduced models for lift, drag, and pitching moment each featured 55 terms including the intercept. These models were full quartic in three numerical factors, with a fourth categorical factor that could assume one of three defined levels. Backward Elimination was applied to discard from each full response model all terms that were statistically insignificant but not hierarchically inferior. For the model of lift coefficient, for example, this reduced the term count from 55 to 20 terms. Since, by Eq. (15), average prediction variance is proportional to p, the term count including intercept, this would be expected to reduce the average prediction variance ("sigma squared") by the ratio of 20/55, or 36.4%. The average standard error of the prediction ("sigma") would then be reduced by the square root of this, or 60.3% compared to the full model. That is representative of the decrease in uncertainty that is achievable through model reduction generally in these types of experiments, when hierarchy is maintained.



[

Figure 2. Reduction in prediction standard error from eliminating terms in three representative wind tunnel models, with hierarchy imposed (blue bars) and without hierarchy imposed (maroon bars).

A further five hierarchically inferior terms could have been eliminated from the lift model to maximize the reduction in uncertainty, dropping the term count from 55 to 15 instead of 20. This would be have reduced the average prediction variance by the ratio of 15/55, or 27.3%, and the average standard error of the prediction by 52.2% relative to the full model, compared to the 60.3% reduction achieved while maintaining hierarchy. These results are displayed in Fig. 2 for the CL (coefficient of lift) model. Similar results are displayed for the drag and pitching moment models.

In all three models examined, a further reduction in uncertainty was theoretically achievable by eliminating hierarchically inferior terms. However, it is clear that most of the benefit of term-count reduction was achieved before the hierarchically inferior terms were eliminated. The incremental improvement from eliminating those terms is relatively small in this case, and in other cases in which the original term count is large. This is the case in a typical wind tunnel test, in which relatively high-order polynomials (typically up to 4^{th}-order, and occasionally higher) are fitted in several factors.

The author has concluded that it is best to strive for as much reduction in prediction variance as can be achieved by eliminating insignificant terms, while simultaneously respecting hierarchy. A policy of maintaining hierarchy eliminates the potential conflicts between coding and model reduction, allowing the benefits of both coding and model reduction to be realized. This ensures a consistency in model form from coded to physical units as well as a consistency in model adequacy assessment metrics, at the cost of only a slight increase in prediction uncertainty that for practical purposes such as the examples illustrated in Fig. 2, is usually inconsequential.

When there is a term that is both statistically insignificant and hierarchically inferior to a higher order term displaying a substantial level of multicollinearity (as evidenced by a large Variance Inflation Factor, for example), it is often worthwhile to provisionally reject the higher-order, correlated term, refitting the model without it. This can

result in the elimination of the small, hierarchically inferior term that was retained previously only to maintain hierarchy with respect to the higher-order term, so the advantages of term reduction are realized without the need to violate hierarchy. As a bonus, the elimination of higher-order terms simplifies the model, making it less sensitive to experimental error in the estimation of higher-order regression coefficients. However, care must be exercised to ensure that any higher-order terms eliminated in this way are not especially important to the fitting process. Rejecting important high-order terms generally results in a serious degradation of the curve fit that is easy to detect when it occurs.

C. Other Points of View

Three questions were highlighted in the previous subsection: a) should we code, b) should we reject selected model terms, and c) should we maintain hierarchy? While the author answers all three of these questions in the affirmative, there is no consensus in the literature. Other points of view, both pro and con, will be summarized here.

1. Factor Coding

Montgomery, Peck, and Vining[23] note that there is some controversy about the decoupling of the intercept from the rest of the model that coding induces. They point out Brown's[34] argument against centering on the grounds that it impacts only the intercept, and that Belsley, Kuh, and Welch[35] recommend against it because it forecloses certain options to diagnose the role of the intercept in near linear dependencies. However, Montgomery, Peck, and Vining themselves recommend scaling and centering the data to be fitted with a regression model because of the intuitive appeal of the standardized regressors that result from this operation (see discussion above following Eq.(3)). They also cite the reduction in variance inflation in the parameter estimates that results from removing non-essential ill-conditioning associated with arbitrary factor intercepts, and cite other authors who also use coding (Hoerl and Kennard[36-37], Marquardt and Snee[22]). On the other hand, Bradley and Srivastava[38] note that while centering the factors decouples the intercept from other regressors in the model, large correlations can still exist among other regression coefficients, which is the argument made by Brown[34]. Coding therefore cannot be regarded as comprehensive protection against multicollinearity.

2. Model Reduction

Box, Hunter, and Hunter[39] distinguish between causation and correlation in an argument against what they describing as the "cherry-picking" of individual non-significant terms. They claim that replacing a non-zero but statistically insignificant regression coefficient with zero has nothing to recommend it, pointing out that the resulting expression would no longer be a least-squares estimate of the model. (They do not address in their criticism the positive impact that term reduction has on model prediction variance, nor the fact that a reduced model can be refitted to produce a new set of regression coefficients that will in fact represent a least-squares estimate of the new, reduced model.)

They identify two applications of regression analysis; to find out, in their words, "what causes what" on the one hand, and on the other hand to establish a subset of possible regressors that best correlates with some system response of interest. It is in this latter application that model term reduction seems to be better justified in their view, although they emphasize that a model obtained in this way has little to say about causation, notwithstanding the fact that it can be useful for predicting system responses.

3. Hierarchy

Draper and Smith[30] argue against dropping hierarchically inferior terms under a translation of origin and propose the following rule:

> "If a model is to be consistent under a shift in origin, only the highest-order terms can be deleted at first and any chosen deletions must keep the model well-formulated. Moreover, if any of the highest-order terms are retained, all terms of lower order affected by them in a shift of origin must also be retained, whether or not their estimates are significant in the regression fit."

Draper and Smith also provide guidelines for removing terms when a rotational transformation is applied.

Montgomery, Peck, and Vining[23], however, express "mixed feelings" about imposing hierarchy automatically. They concede that it is attractive to have the model form preserved after a coding transformation in which a model fitted in coded units is converted to physical units, but they describe this as "purely a mathematical nicety." (They do not address the fact that model adequacy assessment statistics such as R^2 are not invariant under a coding transformation.) They also note that there are natural laws that are not hierarchical, citing Newton's law of gravity and the magnetic dipole law as examples. They further note that there are circumstances in which subject matter expertise would argue against the inclusion of a hierarchically inferior term, as described in this paper in the discussion surrounding Eq. (20). They recommend generating the full response model, and then relying on discipline knowledge to decide which terms, if any, to eliminate

VI. Concluding Remarks

Substantial productivity improvements in aerospace ground testing can be achieved over conventional OFAT testing methods when response surface modeling methods are adopted. Response modeling benefits in numerous ways by a linear scaling and centering transformation that codes the independent variables by mapping them into a limited range centered on zero and extending, typically, from -1 to +1.

Significant quality improvements can be achieved by rejecting terms in a provisional response model that contribute more to the unexplained variance in a sample of data than to the variance that is explained by the model, but there are special considerations to take into account when terms expressed in coded variables are candidates for deletion. The shift in origin associated with centering factors used in a regression analysis results in a change in the functional form of models with missing hierarchically inferior terms. We say that "the estimation space of the response model is not invariant" under such conditions. Goodness-of-fit statistics such as R^2 are also artificially raised or lowered under a coding transformation when hierarchically inferior terms are missing from the model.

These factors argue in favor of maintaining hierarchy during term reduction, especially since most of the reduction in model prediction variance that is provided by term reduction can be achieved be eliminating terms that are not hierarchically inferior. Further reducing the model by eliminating hierarchically inferior terms, however insignificant, often results in only a relatively small additional decrease in prediction variance. This paper concludes that such a marginal improvement does not justify the estimation space invariance or the artificial increases or decreases in goodness of fit metrics that can accompany the elimination of hierarchically inferior, coded regression factors.

Acknowledgments

This work was supported by the Aeronautics Test Program Office of the National Aeronautics and Space Administration.

References

[1]Fisher, R. A., *Statistical Methods for Research Workers*, 1st Ed., Oliver and Boyd, Edinburgh, 1925.

[2]Fisher, R. A., *The Design of Experiments*, 1st Ed., Oliver and Boyd, Edinburgh, 1935.

[3]Fisher, R. A., *Statistical Methods and Scientific Inference*, 1st Ed., Oliver and Boyd, Edinburgh, 1956.

[4]DeLoach, R., "Applications of Modern Experiment Design to Wind Tunnel Testing at NASA Langley Research Center," AIAA 98-0713, 36th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 1998.

[5]DeLoach, R., "Tailoring Wind Tunnel Data Volume Requirements Through the Formal Design Of Experiments," AIAA 98-2884, 20th Advanced Measurement and Ground Testing Conference, Albuquerque, New Mexico, June 1998.

[6]DeLoach, R., "Improved Quality in Aerospace Testing Through the Modern Design of Experiments (Invited)," AIAA 2000-0825, 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2000.

[7]DeLoach, R., "Tactical Defenses Against Systematic Variation in Wind Tunnel Testing," AIAA 2002-0885, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 14–17, 2002.

[8]DeLoach, R., "A Factorial Data-Rate and Dwell-Time Experiment in the National Transonic Facility," AIAA 2000-0828, 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2000.

[9]DeLoach, R., Hill, J. S., and Tomek, W. G., "Practical Applications of Blocking and Randomization in a Test in the National Transonic Facility" (invited), AIAA 2001-0167, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[10]Morelli, E.A., and R. DeLoach, "Response Surface Modeling Using Multivariate Orthogonal Functions" (invited), AIAA 2001-0168, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[11]Underwood, P., Everhart, J., DeLoach, R., "National Transonic Facility Wall Pressure Calibration Using Modern Design of Experiments" (invited), AIAA 2001-0171, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[12]Parker, P., and R. DeLoach, "Response Surface Methods for Force Balance Calibration Modeling," 19th International Congress on Instrumentation in Aerospace Simulation Facilities, Cleveland, Ohio, August 2001.

[13]Danehy, P. M., DeLoach, R., and Cutler, A.D., "Application of Modern Design of Experiments to CARS Thermometry in a Supersonic Combustor," AIAA 2002-2914, 22nd AIAA Aerodynamic Measurement Technology and Ground Testing Conference, St. Louis, Missouri, June 24–26, 2002.

[14]Morelli, E. A., and R. DeLoach, "Ground Testing Results Using Modern Experiment Design and Multivariate Orthogonal Functions (Invited)," AIAA 2003-0653,. 41st AIAA Aerospace Sciences Meeting & Exhibit, Reno, Nevada, January 6–9, 2003.

[15]Dowgwillo, R. M., and R. DeLoach, "Using Modern Design of Experiments to Create a Surface Pressure Database From a Low Speed Wind Tunnel Test," AIAA 2004-2200, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, Portland, Oregon, June 28–30, 2004.

[16]Albertani, R., Stanford, B., DeLoach, R., Hubner, J. P., and Ifju, P. S., "Wind Tunnel Testing and Nonlinear Modeling Applied to Powered Micro Air Vehicles with Flexible Wings," *AIAA Journal of Aircraft*, Summer 2007.

[17]Erickson, G. E., and R. DeLoach, "Estimation of Supersonic Stage Separation Aerodynamics of Winged-Body Launch Vehicles Using Response Surface Methods," 26th International Council of Aeronautical Sciences, Anchorage, Alaska, Sep 14–19, 2008.

[18]DeLoach, R., Marlowe, J. M., and Yager, T. J., "Uncertainty Analysis for the Evaluation of a Passive Runway Arresting System," AIAA-2009-1156, 47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 5–8, 2009.

[19]DeLoach, R., and K. H. Lyle, "An Airbag Landing Dynamics Experiment Using the Modern Design of Experiments," AIAA-2009-0622, 47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 5–8, 2009.

[20]DeLoach, R., "Analysis of Variance in the Modern Design of Experiments" AIAA 2010-1111, 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 4-7, 2010.

[21]DeLoach, R., "Assessment of Response Surface Models Using Independent Confirmation Point Analysis" AIAA 2010-741, 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 4-7, 2010.

[22]Marquardt, D.W. and R. D. Snee [1975], "Ridge Regression in Practice," Am. Statist., **29**(1),3–20.

[23]Montgomery, D.C., E.A. Peck,, and C.G. Vining, *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Statistics, 3rd ed., John Wiley and Sons, New York, 2001.

[24]DeLoach, R., and Ulbrich, N., "A Comparison of Two Balance Calibration Model Building Methods (Invited)," AIAA 2007-0147, 45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 8–11, 2007.

[25]Montgomery, D.C., and E.A. Peck, *Introduction to Linear Regression Analysis*, 2nd ed., John Wiley and Sons, New York, 1992.

[26]Box, G. E. P., and N. Draper, *Empirical Model-Building and Response Surfaces*, John Wiley and Sons, New York, 1987.

[27]Kempthorne, O., "Classificatory Data Structures and Associated Linear Models," *Statistics and Probability: Essays in Honor of C. R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, Eds., Amsterdam, North Holland, pp. 397–410.

[28]Christensen, R., *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer-Verlag, New York, 1987.

[29]Piexoto, J.L., "A Property of Well-Formulated Polynomial Regression Models," *Amer. Statist*, Feb. 1990, Vol. 44, No. 1.

[30]Draper, N. R., and H. Smith, *Applied Regression Analysis*, 3rd ed., John Wiley and Sons, New York, 1998.

[31]Minitab, Software Package, Ver. 14.2, Minitab, Inc., State College, Pennsylvania, 2003.

[32]Peixoto, J.L., "Hierarchical Variable Selection in Polynomial Regression Models," *Amer. Statist.*, 1987, **41**, pp 311-313.

[33]Design Expert, Software Package, Ver. 7.1.6, StatEase, Inc., Minneapolis, Minnesota, 2008.

[34]Brown, P.J., "Centering and Scaling in Ridge Regression," *Technometrics*, 1977, Vol. 19, pp 35-36.

[35]Belsley, D.A., E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York, 1980.

[36]Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 1970, Vol. 12, pp 55-67.

[37]Hoerl, A.E. and R.W. Kennard, "Ridge Regression: Application to Nonorthogonal Problems," *Technometrics*, 1970, Vol. 12, pp 69-82.

[38]Bradley, R.A. and S.S. Srivastava, "Correlation and Polynomial Regression," *Amer. Statist.*, 1979, Vol. 33, pp 11-14

[39]Box, G. E. P., W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*, 2nd Ed., John Wiley & Sons, New York, 2005.