Assessment of Response Surface Models Using Independent Confirmation Point Analysis

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

This paper highlights various advantages that confirmation-point residuals have over conventional model design-point residuals in assessing the adequacy of a response surface model fitted by regression techniques to a sample of experimental data. Particular advantages are highlighted for the case of design matrices that may be ill-conditioned for a given sample of data. The impact of both aleatory and epistemological uncertainty in response model adequacy assessments is considered.

I. Introduction

Substantial gains in aerospace testing productivity have been achieved in recent years by the application of response surface modeling methods in empirical investigations[1-8]. Response surface models are mathematical relationships expressing various system responses of interest to the independent variables (also called “factors”) that influence those responses. For example, response surface models developed in wind tunnel testing typically describe such system responses as forces, moments, and pressures as a function of such factors as angle of attack, sideslip angle, Mach number, Reynolds number, and the deflection angle of various control surfaces. Qualitative variables, also called “categorical factors,” are also accommodated. These are independent variables constrained to a finite number of discrete states or levels so, for example, the state of the landing gear might be represented by a categorical factor that assumes the value of “0” when the gear is not deployed (“landing gear up”) and a value of “1” when the gear is deployed (“landing gear down”).

The concept of a “design space” (also called “inference space”) is used to graphically represent the factor combinations specified in a response surface modeling experiment. A design space is simply a coordinate system in which each axis corresponds to a different independent variable. Each point in this space, called a “site,” represents a unique combination of factor levels. We use the term “design site” or “model site” to describe factor combinations for which data are acquired to fit a response model. Other sites within the design space may be described variously as “off-design” sites, or, for a particular application that will receive considerable attention in this paper, “confirmation-point” sites.

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. It is not necessary to take the time or bear the expense of physically setting those intermediate factor combinations and measuring the corresponding responses.

The operative word in the preceding paragraph is “adequately.” For a response model to be regarded as “adequate,” it must be capable of predicting system responses for any combination of factor levels of interest, within a prescribed tolerance. As a practical matter, the process of generating a candidate response model from a sample of experimental data is not particularly difficult, as commonly available commercial software packages have automated this task. Most of the effort in developing a response model is not in generating the model in the first place, but in validating it once it is in hand.

There are numerous procedures available to test the adequacy of a candidate response model. Most of these entail an examination in one form or another of “residuals.” Residuals represent the difference between measured and predicted system responses at prescribed sites in the design space. Absent information obtained independent of the test in which the response model is developed, the only available information about the quality of a response model is derived from residuals, so this is a crucial topic for response surface modeling methods.

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

The distributional properties of residuals are among the first properties to be examined. An adequate response model developed from independent measurements is expected to fare through points near the mean of a sample of responses that might be acquired at each site in the design space. Even if only one data point is acquired at a given site, it is regarded having been drawn from a sample of points that could have been acquired there, the mean of which would correspond to the model prediction at that site for a perfectly fitted model. Any departure of a measured response from that mean would be attributed to ordinary random experimental error for such a model, so residuals of well-fitted response models are expected to be drawn from a normal population by the Central Limit Theorem. We therefore invoke various tests for the normality of the residual distribution. The presence of a significant systematic component of the sample variance that remains unexplained by the response model is regarded as a Lack of Fit (LOF) error, and interpreted as evidence of an inadequate response model[9-10].

The residuals are examined for other patterns as well, typically using graphical methods that display them as a function of predicted response, as a function of each factor level, as a function of elapsed acquisition time, and in any other way that the experimenter believes will be revealing. A comprehensive exposition of the types of available tests is beyond the scope of this paper, but standard references on response surface modeling describe such tests in detail[11-13].

Finally, the magnitude of the residuals is assessed, typically by comparing some summary statistic such as the standard deviation of the residuals to a tolerance specification that defines adequate modeling precision. If the model is adequate in other respects but simply lacks the requisite precision, and this assessment is made during the execution of the test, additional replicates can be specified to further cancel experimental error.

Residuals used to assess the adequacy of a response model have been historically estimated at the design space sites where the fitted data were acquired. We refer to these as “model residuals.” Having fitted a candidate response model to a sample of experimental data acquired at specific sites within the design space, model residuals are computed by subtracting response predictions from response measurements at those sites.

It is unfortunate that model residuals have the potential to understate the uncertainty associated with a fitted response model. This is because the regression methods used to develop response models typically generate a vector of regression coefficients that are optimized in a least squares sense. That is, the coefficients are chosen to minimize the sum of squared residuals across all model design sites. This produces the smallest possible average squared residual across all sites used in the regression, which is commonly regarded as an effective metric for the quality of the response model.

However, the model must make adequate response predictions at other sites within the design space besides those where the data used to fit the model were acquired. For this reason, a case is made in this paper for evaluating the response model by analyzing residuals from a sample of “confirmation points.” These are measurements made at other sites besides those used to fit the model. They are held in reserve for the express purpose of testing the predictive power of a candidate response model and are not used in the regression analysis or other computations that produce the model. Confirmation points provide a more stringent test of a regression response model. If a sufficient number of confirmation points are selected from randomly-chosen sites within the design space, a more realistic evaluation of the model’s general predictive capability can be made.

Examples of model and confirmation-point residuals for the same test are presented in Section II, illustrating typical differences in the associated standard errors and also certain additional insights afforded by confirmation points. Section III introduces the impact of multicollinearity and describes the special advantages of confirmation point residuals over model residuals when regressors are correlated. Section IV describes critical binomial analysis, a proposed methodology for quantitatively assessing residual magnitudes. Distinctions between aleatory and epistemological uncertainty in response model adequacy assessment are described in Section V. A few miscellaneous topics are selected for further discussion in Section VI, and concluding remarks are presented in Section VII.

II. Comparison of Model-Point and Confirmation-Point Residual

The number of terms, including the intercept, in a full d^th^-order polynomial in *k* factors can be computed as follows:

$$p = \frac{(d+k)!}{d!k!} \tag{1}$$

The term count clearly grows rapidly with both the order of the model and the number of factors. A broad class of six-component force balances used in wind tunnel testing are typically calibrated with a second-order response model, and such models therefore have (2+6)!/2!6! = 28 terms if all terms are retained in the model. (A common analytical process makes tare corrections that drive the intercept through the origin, leaving 27 terms). Each term contributes to the prediction uncertainty in the response model. The unexplained prediction variance averaged over all sites in the design space for which fitted data were acquired is in fact directly proportional to the term count (Ref. 9, Appendix 3E):

$$\overline{Var\ \ \hat{y}} = p\left(\frac{\sigma^2}{n}\right) \tag{2}$$

where the caret over y indicated a predicted response, n is the number of points fitted in the regression, and σ^2 is the unexplained variance in the fitted data. Note that the term in parentheses is just the variance in the distribution of an n-point sample mean, and represents the average contribution of each term in the model to the average unexplained variance in model predictions, the square root of which is the standard error in the prediction (the "one-sigma uncertainty" in a model prediction). Clearly, all else being equal we prefer response models with the fewest terms (smallest p) possible.

Various methods have been devised to test the coefficients of a candidate response model with a view to possibly eliminating some of the model terms, thereby reducing prediction uncertainty. The criterion for retention in the model is typically that the value of the regression coefficient must be sufficiently large compared to the uncertainty in estimating it that it can be regarded as non-zero with some prescribed level of confidence (95%, say).

A. A balance calibration example

The author recently participated in a comparative evaluation of two software systems, each using somewhat different algorithms to objectively reduce the term count in a candidate response model[14]. A 28-term balance calibration model was used as the starting point, representing the best fit by each software system to a set of data from a balance calibration experiment. Each software system then identified subsets of these 28 terms with signal-to-noise ratios small enough to justify dropping them to reduce prediction uncertainty. The standard deviation in a set of 25 confirmation-point residuals was then computed for each resulting reduced model and used to evaluate the two model reduction methods.

The confirmation-point sites were distributed at random throughout the six-dimensional design space of the calibration experiment. Except under one set of circumstances that will be described in more detail in Section III, both model reduction methods resulted in nearly identical reduced models, with identical retained terms for a number of balance force and moment response models, and small differences in other models that nonetheless resulted in response predictions that were nearly identical. We revisit those results here to compare residuals from the 25 randomly selected confirmation points with model residuals estimated in the usual way at sites where the measurements were made that were used to fit the response model.

Figure 1 was prepared by computing the standard deviation ("one sigma") of both the model- and confirmation-point residuals for each of the six force and moment reduced response models developed from the balance calibration data using one of the two software systems under evaluation[15]). Each standard deviation was then doubled to approximate a 95% prediction interval half-width. To facilitate comparisons across multiple response variables with differing sensitivities, each "two-sigma" value was then normalized by the full-scale output of the corresponding response variable.

Figure 1 indicates that uncertainties are less than a quarter percent of full-scale for all six force and moment response models. This is a common criterion for an acceptable level of uncertainty, so we conclude that the response models developed for all six forces and moments feature acceptable levels of uncertainty not only at the model sites where data were acquired to fit the response models, but at other randomly selected sites throughout the design space.

We note also that as expected, the uncertainty estimated at randomly selected confirmation-point sites is somewhat greater than the uncertainty estimated at the model regression sites where the regression algorithm minimizes the sum of squared residuals. The larger uncertainty estimates based on the confirmation-point residuals are believed to be more representative of the true uncertainty, and represent a more conservative uncertainty estimate.

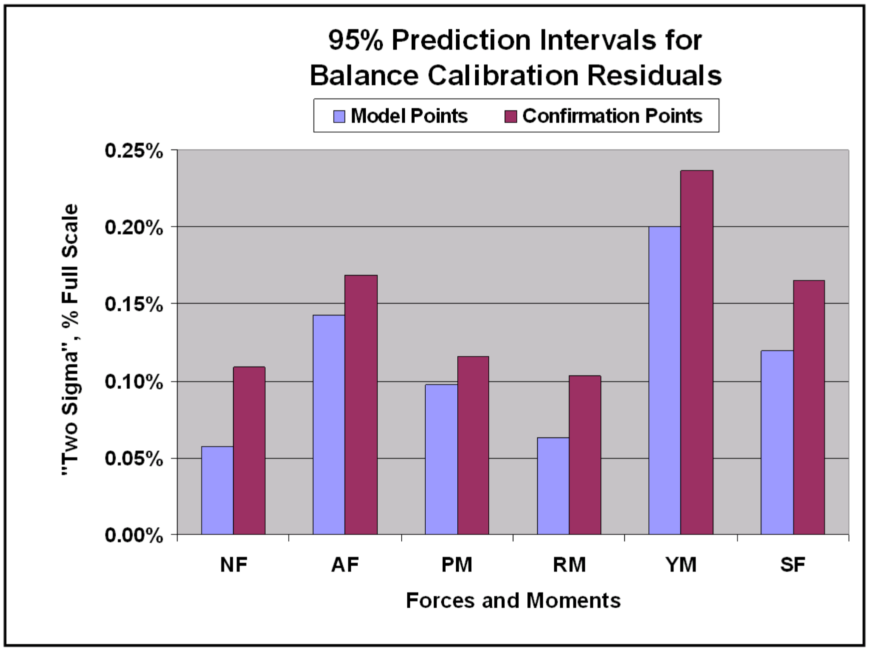


Figure 1. Comparison of model- and confirmation-point residuals in a balance calibration experiment featuring manual dead-weight loading.

It is not clear from Fig. 1 whether the larger confirmation-point variance is attributable to a few isolated outliers, or if it is a general phenomenon that model prediction error is larger at off-design points than at the sites where the regression data were acquired. Figure 2 displays all model residuals and all confirmation-point residuals for the normal force model of Fig. 1, revealing that larger errors at off-design points does indeed appear to be a general phenomenon, at least for this particular system response.

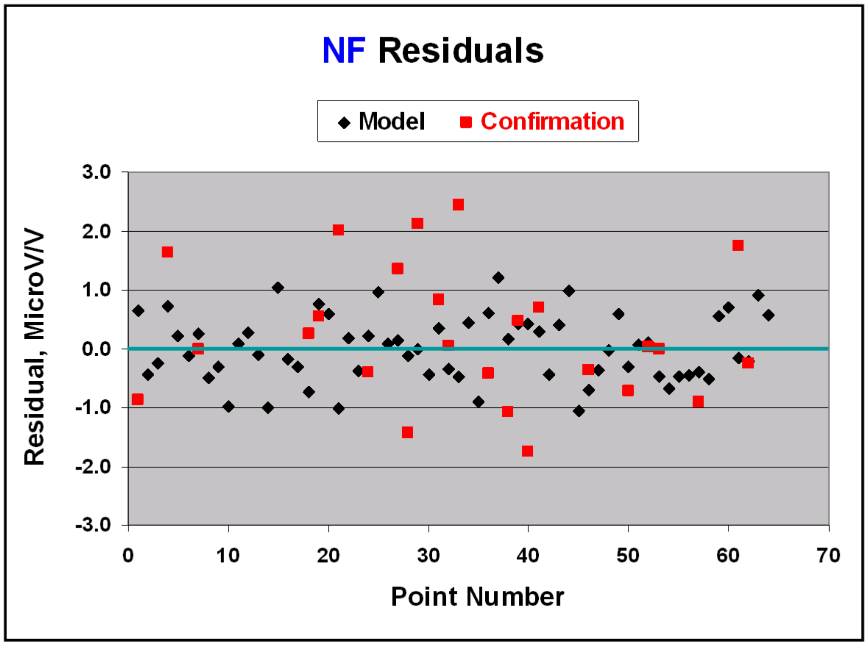


Figure 2. Model and confirmation-point residuals for the normal force response model in a balance calibration.

Figure 2 shows that both the model and the confirmation-point residuals are symmetrically distributed about a common mean of zero, suggesting a good fit of the model to the data. The model residuals are approximately contained within a range of ±1 microV/V, while the confirmation-point residuals appear to lie within a somewhat wider range – about ±2 microV/V. Uncertainty estimates in this case that are based on model residuals therefore understate the uncertainty that will be associated with general model predictions by roughly a factor of two.

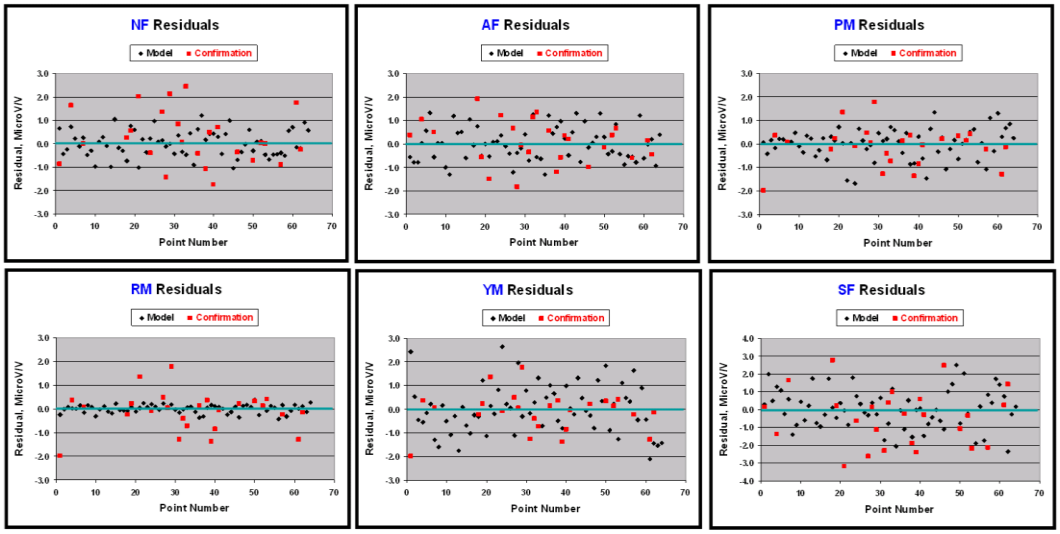


Figure 3. Model and confirmation-point residuals for force and moment models in a balance calibration.

Figure 3 presents residuals estimated at model sites and at randomly-selected confirmation-point sites for all six forces and moments in the calibration analysis described here. Note that same general trends are evident for all the responses. The models for each system response appear to be adequate based on the symmetry of positive and negative errors displayed in this figure, but in general the confirmation-point residuals appear to be uniformly scattered over a range that is somewhat greater than the model residuals, as Fig. 1 indicates.

B. A second balance calibration example

We have contrasted residuals estimated at independent confirmation-point sites within the design space with those estimated in the usual way at sites where the regression data were acquired. Results of a balance calibration experiment suggested that residuals from arbitrarily selected sites within the design space reveal rather more information about the quality of the calibration than conventional regression residuals. These results were based on calibration data acquired with a manual dead-weight loading system. We continue this comparison of residuals with another set of balance calibration data, this one acquired with an automated balance calibration machine.

Dead-weight loading systems rely upon the force of gravity acting upon a certified mass that is hung via a system of pulleys for which the alignment can be established precisely using a high-precision theodolite with a telescopic sight. Automated balance calibration machines use hydraulic actuators to apply a programmed schedule of force and moment loads to a balance mounted in the machine, logging electrical outputs from the balance for each applied load combination.

The relative virtues of manual dead-weight loading methods for balance calibration vs. automated balance machine methods are a topic of sometimes passionate debate among devotees of the each method. The merits of this debate are beyond the scope of the current paper, but the advantages of each method are briefly summarized here: Dead-weight loading proponents cite the reliability of a method that depends on little else for an accurate calibration than the acceleration of gravity, for which corrections for local geological mass concentrations are made to further improve accuracy. They express concern over the requirement for precise component alignment in an automated balance machine. Those who prefer the automated balance machine approach note that alignment issues can be resolved by the careful attention of a skilled practitioner, and point to the large volume of data that can be acquired with relatively little effort and direct operating cost, and in a relatively short period of time, compared to manual loading methods.

For the purposes of this paper we are less interested in the merits of the two calibration methods than an exploration of what additional information is provided when the calibration model is evaluated with confirmation-point residuals in addition to design-point residuals. To explore this question, design-point residuals were computed for reduced quadratic balance calibration models developed from a calibration experiment in which 1906 loading combinations were applied with an automated balance calibration machine. An equal number of confirmation points were selected at random from a universe of 10,121 available check loads that were set during this experiment, covering various combinations of low and high loading sequences.

Figure 4 compares the unexplained variance of design-point and confirmation-point residuals for calibration models developed for the six component response outputs of the balance. Note that unlike Fig. 1, these results show no systematic difference between confirmation and model residuals.

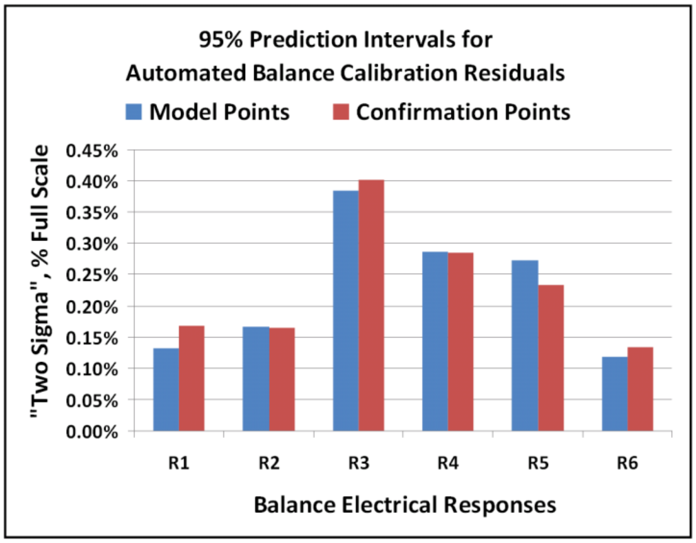


Figure 4. Comparison of model- and confirmation-point residuals from an automated balance calibration experiment.

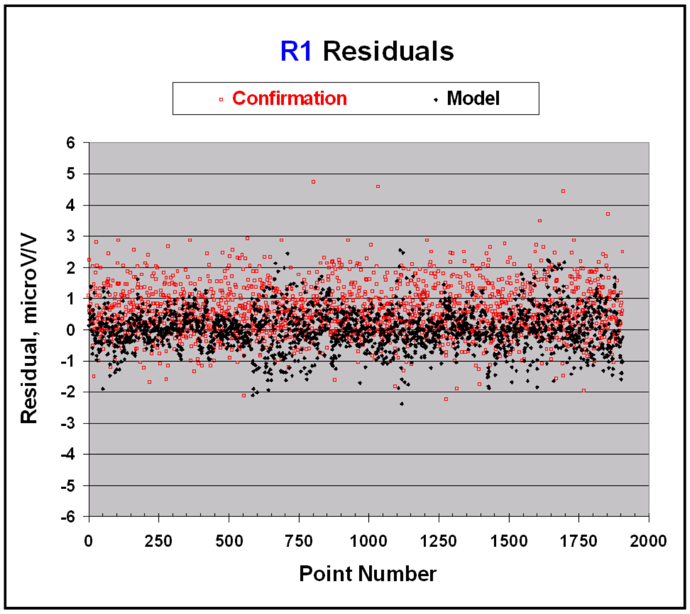


Figure 5. Confirmation point and conventional model design-point residuals for response model R1. Calibration data from automated balance calibration machine.

One explanation for the difference may be the density with which the design space is covered with calibration and confirmation points in an automated balance calibration. For the high-density coverage made possible by an automated calibration machine, the smaller difference between the variance in confirmation-point and conventional design-point residuals is in harmony with an intuitive expectation that variance differences depend on the average distance between confirmation and design point sites in the design space.

Just as in the case of the manual dead-weight loading calibration described in the previous subsection, we present in Fig. 5 all model residuals and all confirmation-point residuals for one of the individual response models of Fig. 4; namely, R1. Figure 5 confirms the result displayed in Fig. 4, that there appears to be no significant difference in the variance of confirmation point residuals and conventional model design-point residuals for this response. However, Fig. 5 does reveal an apparent bias shift of the confirmation point residuals with respect to the model residuals.

While design-point residuals are guaranteed to have a mean of zero due to the nature of the regression calculations, the mean of confirmation point residuals are not constrained to be zero. Nonetheless, for a substantial number of residuals acquired from sites distributed more or less uniformly throughout the design space, one would expect confirmation point residual means to be very nearly zero. We see in Fig. 6 that bias shifts appear in other responses besides R1. The shift for R3 is especially evident.

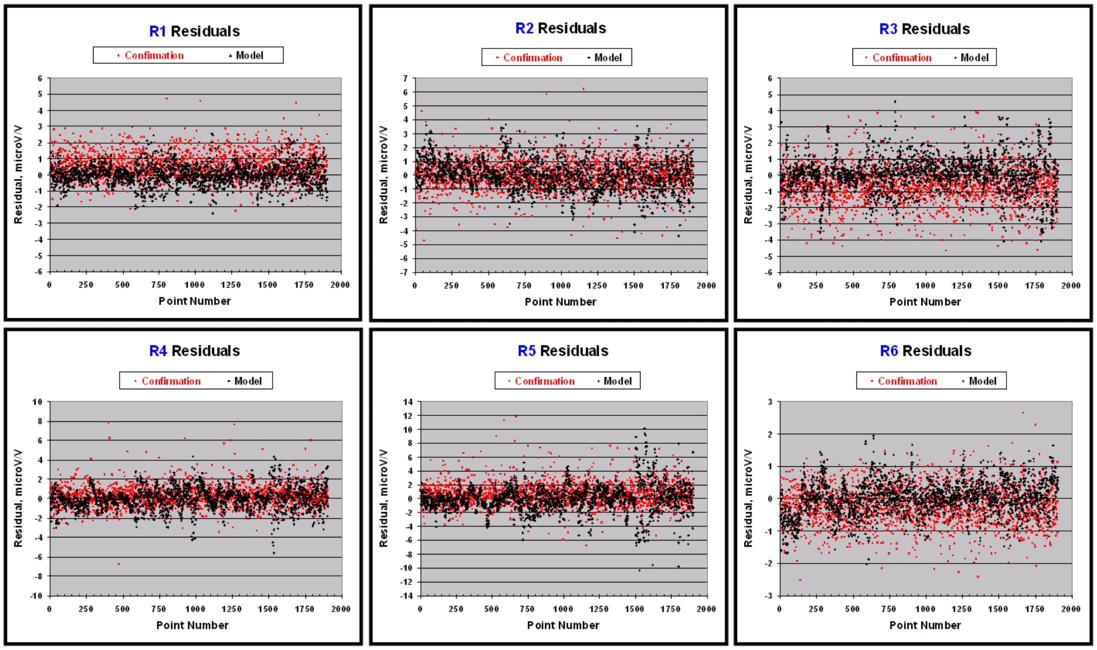


Figure 6. Model and confirmation-point residuals for response models in a balance calibration performed with an automated balance calibration machine. Bias shifts in confirmation-point residuals are evident in a number of responses.

To quantify the mean shift in confirmation point residuals with respect to zero, all 1906 residuals for each of the six responses were averaged. A tolerance level was computed for each response equal to 0.25% of the full scale output for that response, and the bias shifts were expressed as a percentage of that tolerance. Fig. 7 displays the results.

Fig. 7 suggests that the mean confirmation point residuals differ from zero by an amount that is a substantial fraction of the calibration error budget. This shift is benign if it is coincidental, due only to the selection of an unlucky combination of confirmation point sites for which ordinary random errors happened to have been mostly of the same sign. But if it is evidence of some systematic error, then that error should be taken into account in estimating the uncertainty of the balance calibration.

Figure 8 compares the mean of the 1906 confirmation point residuals displayed in Fig. 7 for the automated balance calibration machine, with the mean of the 25 confirmation point residuals acquired in the dead weight calibration described in the previous section. The mean shift for the automated balance calibration machine appears to be generally greater than for the hand-loaded calibration. This is probably not attributable to differences in the balances and does not reflect on balance precision, per se. Those factors would impact the *variance* in the residuals, not their means. A shift in mean confirmation point residuals suggests a block effect, attributable to a systematic difference between the mean of measurements made in one block of time and the mean of measurements made in another.

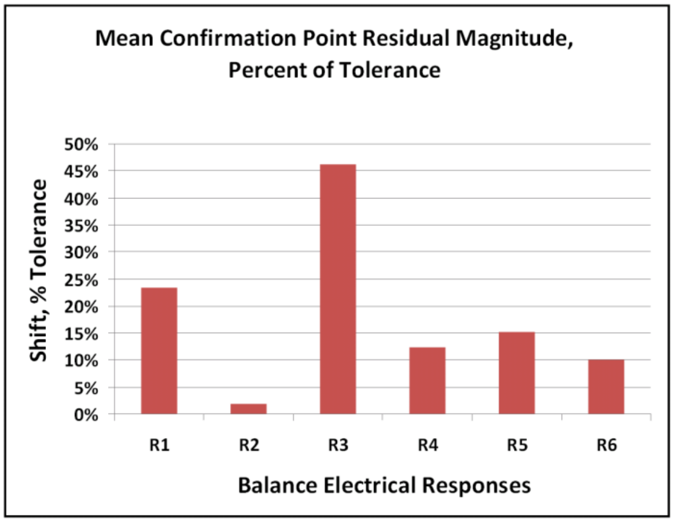


Figure 7. Magnitude of average confirmation-point residual for an automated balance calibration machine, as a percent of a tolerance level of 0.25% of full scale output.

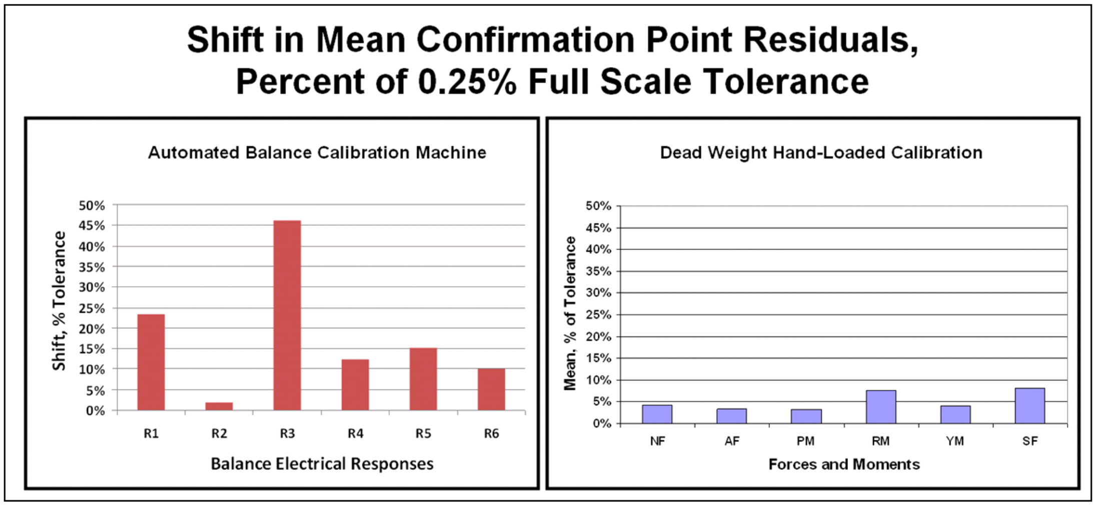


Figure 8. Comparison of confirmation-point residual means for an automated balance calibration machine and a dead weight, hand-loaded calibration.

One possible explanation for these bias shifts is that in an automated balance calibration, the check loads are acquired in runs consisting of a large number of individual data points that share a common set-up and alignment. If there is any alignment error, this error does not cancel from point to point, but impacts all of the check loads in that run.

In any case, the use of check-load confirmation-point residuals to test the predictive power of a candidate calibration model can reveal such shifts if the check loads are acquired in different runs for which different machine setups are used. These block effects provide additional insight into the uncertainty that is likely to be associated with a given balance machine calibration performed at other times and with other setups and alignments. These results suggest that it would be a good practice to block the calibration experiment into a number of runs for which the balance is realigned independently. This would permit any realignment errors to be easily detected and also easily quantified, resulting in a more realistic estimate of the true calibration uncertainty.

It has not been the intent of this analysis to reach general conclusions about the relative merits of automated and manual balance calibration methods, and indeed such conclusions cannot be reliably inferred from a comparison of only two calibrations, each on a different balance. Each calibration method has its strengths and its weaknesses. The intent here has been to simply show how independent confirmation point residuals can provide greater insights into the true uncertainty in a response modeling experiment than is afforded by design-point model residuals alone.

III. The Impact of Multicollinearity

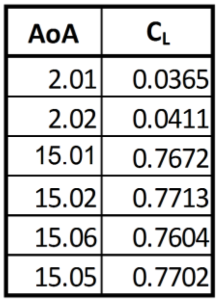
There is an especially pathological situation that can occur in which design-point residuals present a particularly erroneous assurance of the quality of a response model's general predictive capability. This is the case in which significant multicollinearity is in play. Under these circumstances, confirmation point residuals have an especially useful role to play.

A. A pedagogical illustration

We illustrate the multicollinearity case with an elementary example based on a subset of lift data acquired in a recent wind tunnel test and extracted from a larger sample of data to facilitate this illustration. The central points that are illustrated in this simple example will then be revisited in a more realistic case in the next subsection.

Consider the lift data in Table 1, which displays a number of lift coefficients (C_L) and the angles of attack (AoA) at which they were acquired.

Table 1. A Sample of Lift Data



AoA	C_L
2.01	0.0365
2.02	0.0411
15.01	0.7672
15.02	0.7713
15.06	0.7604
15.05	0.7702

These data are from a pre-stall lift polar. We know from experience that such a polar can often be represented adequately as a quadratic function of angle of attack, assuming as in this case that the highest angle of attack to be fitted is sufficiently below the stall angle. (The closer we encroach upon the stall angle, the higher the order of polynomial that is generally required to achieve a good fit.) The data in Table 1 can be fitted with a quadratic model, as follows:

$$Y = -0.4750 + 0.2817x - 0.0132x^2 \tag{3}$$

The adjusted R^2 statistic for this curve fit, indicating what fraction of the total variance is explained by the model, has a value of 0.9999. A value of 1 indicates a perfect model, suggesting that a quadratic model fits the data well. However, a simple plot of the data as in Fig. 9 is enough to raise suspicions.

While in principle there are a sufficient number of unique angle of attack levels in Table 1 to support as high as a 5^{th}-order polynomial, these angles actually cluster in only two rather tightly packed groups around nominally 2° and 15° angle of attack. From a simple inspection of Fig. 9 it is clear that a first order model would therefore fit the measured data essentially as well as the quadratic model, providing an equivalent level of predictive capability for the angles of attack that were fitted. A first-order model would also avoid the excessively elaborate mid-range structure of the quadratic model on display in Fig. 9, which is unsupported by any data in the range between 2° and 15°. Figure 10 displays such a first-order fit.

The inadequacy of the quadratic model for fitting the data of Table 1 is even more clearly revealed by displaying its 95% confidence interval, as in Fig. 11, prepared with the aid of commercially available analysis software that generates such intervals automatically[15]. Clearly there are many quadratic curves that could fit within these limits. Since the width of the confidence interval contracts significantly in the region where there is actual data, quadratic response models that fall within the confidence interval of Fig. 11 would generally pass close to the plotted data in this figure. This includes the special case in which the coefficient of the quadratic term is zero and the response is first order, as in Fig. 10. For this reason, model residuals obtained at design-point sites where fitted data were acquired will be small, providing a false sense of model adequacy in this instance.

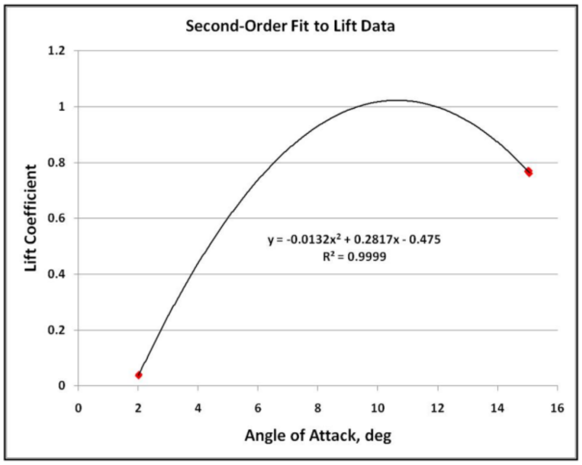


Figure 9. Quadratic fit to sample of lift data.

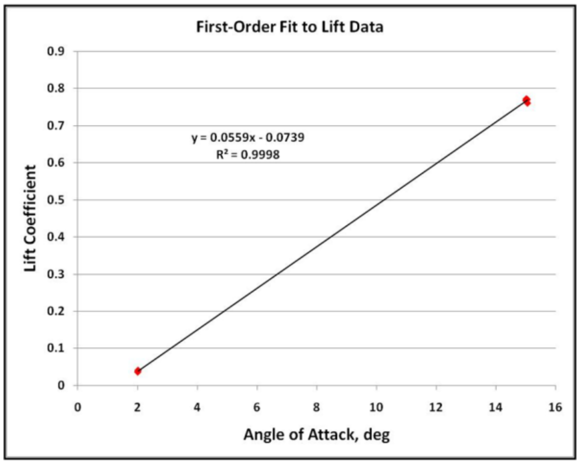


Figure 10. First-order fit to sample of lift data.

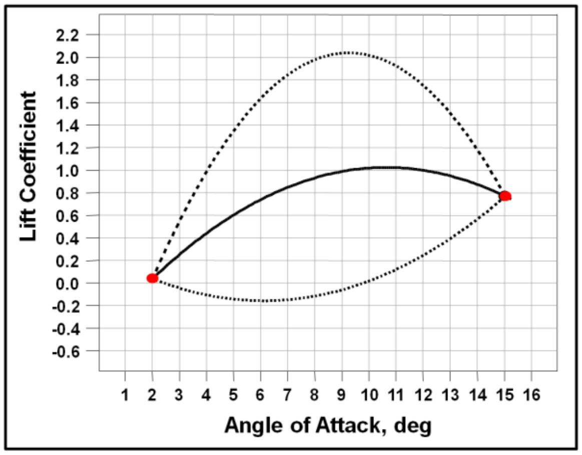


Figure 11. Quadratic response model (solid line) with 95% confidence interval (dashed lines).

The confidence interval for the quadratic model is large because the uncertainties in the model coefficients are large. To understand why this is so requires a brief review of the mathematics of regression analysis.

The lift coefficient numbers in the right-hand column of Table 1 comprise a response vector, **Y**, and if we also represent the three regression coefficients in Eq. (1) as a vector, **b**, we can then compute **b**, given **Y**, by this formula:

$$\mathbf{b} = \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{Y} \qquad (4)$$

where **X** is the *design matrix*. The design matrix has *p* columns and *n* rows, where *p* is the number of terms in the model to be fitted (including the intercept) and *n* is the number of data points to be fitted to the model. For the case of a quadratic model in one variable being fitted to the data of Table 1, **X** has six rows and three columns.

There is one row in **X** for each data point that is being fitted and one column for each term in the model. In the present case of a polynomial in one variable, *x*, the elements in each column are computed by raising *x* to an exponent corresponding to the order of the term represented by that column, where *x* in this case is the angle of attack. The column corresponding to the first-order term in the quadratic model simply contains values equal to x^1, or the angles of attack themselves, as displayed in the left-hand column of Table 1. The column corresponding to the second-order term in the quadratic model contains values equal to x^2, the square of the angles of attack. The column corresponding to the intercept, or "0th-order term" of the model, contains the angles of attack each raised to the 0th power, and therefore simply consists of a column of 1s. The design matrix, **X**, for a quadratic polynomial in one variable fitted to the data of Table 1 is therefore as follows:

```
X =

    1.0000    2.0100    4.0401
    1.0000    2.0200    4.0804
    1.0000   15.0100  225.3001
    1.0000   15.0200  225.6004
    1.0000   15.0600  226.8036
    1.0000   15.0500  226.5025
```

The **X′X** matrix that must be inverted in Eq. (4) is then

```
X′X = 1.0e+005 *

    0.0001    0.0006    0.0091
    0.0006    0.0091    0.1361
    0.0091    0.1361    2.0443
```

Note that every element in the third column of the **X′X** matrix is approximately a constant (15.0 ± 0.2) times the corresponding element in the second column. This means that columns 2 and 3 are very nearly linearly dependent. We would say in this case that there is a high degree of multicollinearity among the columns. This makes the **X′X** matrix very nearly singular, so that inverting it to obtain reliable estimates of the regression coefficients in Eq. (4) is problematical.

The **X′X** matrix would be exactly singular if there was a perfectly linear dependence involving two or more columns, in which case its inverse would not exist and there would be an infinite number of solutions to Eq. (4) rather than a single, unique solution. In the present case in which the **X′X** matrix is merely ill-conditioned but not perfectly singular, a large number of solutions exist for Eq. (4) as reflected by the wide prediction interval displayed in Fig. 11 that can accommodate so many such solutions, and there is no way to identify which one represents the true dependence of lift coefficient on angle of attack. Putting this a different way, multicollinearity introduces uncertainty in the regression coefficients and therefore in the response predictions. We say that the variance associated with estimates of each regression coefficient is *inflated* as a result of multicollinearity, causing the higher uncertainty.

We can relieve the multicollinearity in this example by dropping the quadratic term from the model. This minimizes the range of unique coefficient combinations that satisfy the equation for **b**, and therefore restricts the number of response models that can fit the data. The reduced design matrix produces an **X′X** matrix that is not ill-conditioned. Absent multicollinearity effects, the 95% confidence interval upon which the resulting first-order curve fit is centered (Fig. 10) collapses to such a narrow range that it cannot be resolved graphically. That is, within the resolution limits of the figure, the upper and lower limits of the 95% confidence interval for the response model plotted in Fig. 10 cannot be distinguished from the model itself.

This modeling uncertainty associated with the quadratic model of Fig. 11 is difficult to detect by an examination of design-point residuals only. The contracting confidence intervals make it clear that a wide range of poor response models would display relatively small residuals near the model design points. The practical consequence of multicollinearity is that while a large number of response models will fit the design-site data, only the "true" combination of coefficients will also predict off-design points adequately. This is the essential rationale for using confirmation points to assess the adequacy of a candidate response model, rather than relying entirely upon residuals obtained at model design points.

We inject the usual warning that extrapolation outside the range of fitted data is not recommended, even when multicollinearity is not in play. We also note that the data in Table 1, even as fitted in Fig. 10, do not present a compelling case for a true first-order response model. A higher density of data within the design space would be needed to explore the true functional dependence of lift coefficient on angle of attack over these ranges. However, while a higher density of data would in this case be necessary to assure an adequate response model, it would not be sufficient to do so. As we will show in the next subsection, a higher density of data points is not a guaranteed defense against multicollinearity effects, and it is still possible with a relatively high volume of data to fit an inadequate response model for which small residuals are nonetheless obtained at the model design points. In such cases, confirmation-point residuals are especially important to realistically assess the adequacy of a response model.

B. Confirmation points in the presence of multicollinearity

Section II described a comparative evaluation of two software systems in which the author recently participated[14], each using different algorithms to objectively identify insignificant terms in a balance calibration response model. As noted in that section, this is useful because the quality of a balance calibration can be enhanced substantially, and at essentially no extra cost, by eliminating terms that contribute more to the unexplained variance than to the explained variance in a sample of calibration data. The same is true for response models describing any other type of system response; whether they are forces, moments, and pressures in a wind tunnel test, loads in a structural dynamics test, temperatures in a propulsion system exhaust test, or any other type of system response that is to be modeled as a function of specified independent variables.

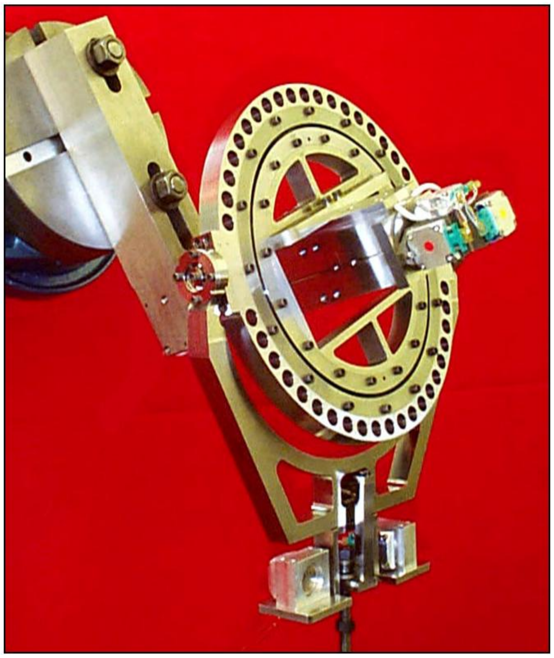


Figure 12. Single Vector System for balance calibration.

In the software evaluation exercise noted above, residuals at 25 randomly selected confirmation-point sites were used to evaluate response models produced when balance calibration data were analyzed by each of the two software systems under study. Standard deviations of these residuals were compared for a number of data sets representing different noise environments and different experiment designs. To focus on the comparison of design-point and confirmation-point residuals in Section II of this paper, all results presented in that section were obtained with the same software system, eliminating this as a confounding factor in the comparison of residuals that is of interest in this paper. We now revisit the original software comparison study to highlight an additional facet of that investigation that sheds light on the relative effectiveness of design-point and confirmation-point residuals in evaluating the adequacy of response models.

One of the balance calibration data sets used for comparison was acquired with advanced dead weight loading hardware known as a Single Vector System (SVS)[16]. The Single Vector System, shown in Fig. 12, allows complex combinations of forces and moments to be applied to a balance under calibration, which facilitates the use of formally designed experiments for balance calibration. This results in significant improvements in both quality and productivity.

When the Single Vector System is used to calibrate a balance, a certified mass is hung for each data point from a prescribed offset point and at a prescribed three-dimensional angle with respect to the coordinate system of the balance. This imparts three force components and three moment components to the balance, allowing a greater variety of simultaneous loads than a conventional manual loading system and facilitating very efficient response surface modeling calibration experiments. By applying all six loads simultaneously on every data point rather than one or two at a time as is typical with conventional manual loading systems, the SVS system forces each data point to "work harder" by conveying more information per data point. The result is a dramatic reduction in the number of loads required to calibrate a balance and a corresponding reduction in time and direct operating cost. There are fewer sources of systematic error in play with the SVS system, and the system facilitates a number of quality assurance tactics (randomization, replication, blocking) that are not as practical to implement with conventional systems[4].

While the SVS balance calibration system has a number of significant advantages, there is one complication that must be taken into account in analyzing the data. The total force vector is always orthogonal to the total moment vector in a SVS loading[17]. As a consequence, the dot product of the two vectors is zero, leading to the following constraint among three interaction terms that could each be retained in a candidate balance calibration equation:

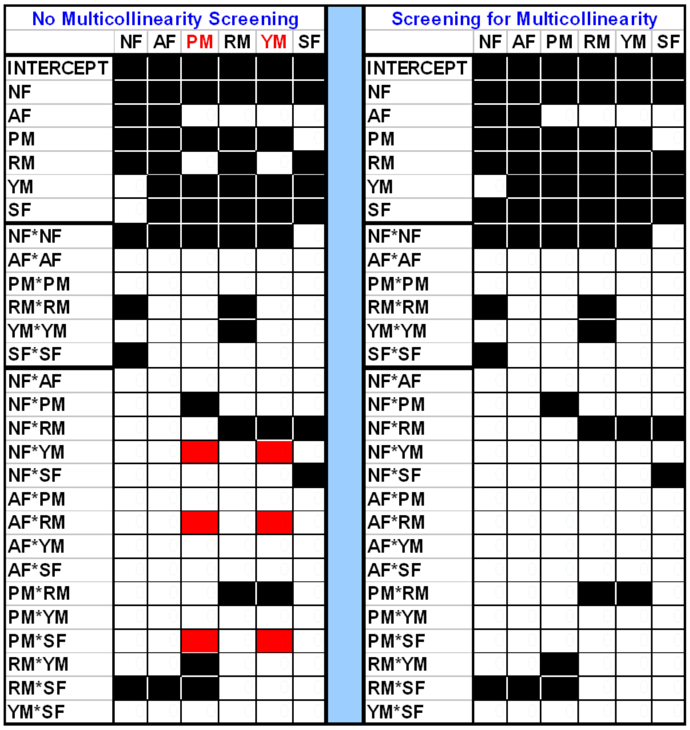
$$F_X M_X + F_Y M_Y + F_Z M_Z = 0 \tag{5}$$

where F_i is the ith component of the force vector and M_i is the moment component about that force axis. In a common coordinate system in which F_x, F_y, and F_z are the normal, axial, and side force components of the total force vector, M_x, M_y, and M_z would correspond to yawing moment, rolling moment, and pitching moment, respectively.

If all three force-moment interaction terms are retained in a candidate response model, only two of the three can be independent. The third term must automatically satisfy the constraint equation, which introduces severe multicollinearity (perfect multicollinearity, except for experimental error) if all three interaction terms are judged to be significant. Model predictions will be poor in this case unless the multicollinearity is resolved by dropping one of the terms from the model. This is easy enough to do when the analyst is aware of the potential problem; dropping any one of the three terms resolves the multicollinearity issue with no significant loss of useful information, since in any model featuring two of the terms, the third term conveys no unique information. The decision as to which term to drop in those instances in which all three terms are candidates is often facilitated by a knowledge of the design characteristics of the specific balance being calibrated.

This particular quirk of the SVS system provided a serendipitous opportunity to demonstrate the impact of multicollinearity on the assessment of response model results when only design-point residuals are available, compared to the case in which independent confirmation points can be used. One of the software systems under evaluation originally had no means to detect multicollinearity, and produced recommended response models for pitching moment and yawing moment that featured all three correlated force-moment interaction terms. The other software system displayed multicollinearity metrics for each candidate term in the model, allowing the analyst to identify the correlated terms, manually drop one of them from the model, and re-compute the regression coefficients for a new model for which the multicollinearity issue was thus resolved.

Figure 13 displays terms that were retained and those that were rejected for the six force and moment calibration response models that were developed by each software system. Filled cells indicate terms that were retained and empty (white) cells indicate rejected terms. The axial force, pitching moment, and side force models developed by the two software systems were identical, and the normal force models differed by only a single regressor. (There was a first-order side force term in the normal force model developed by one software system that was not in the other, but the coefficients were distributed such that there was no significant difference in response predictions.)



	No Multicollinearity Screening						Screening for Multicollinearity					
	NF	AF	PM	RM	YM	SF	NF	AF	PM	RM	YM	SF
INTERCEPT	■	■	■	■	■	■	■	■	■	■	■	■
NF	■	■	■	■	■	■	■	■	■	■	■	■
AF	■	■	□	□	□	□	■	■	□	□	□	□
PM	■	■	■	■	■	□	■	■	■	■	■	□
RM	■	■	□	■	□	■	■	■	■	■	■	■
YM	□	■	■	■	■	■	□	■	■	■	■	■
SF	□	■	■	■	■	■	□	■	■	■	■	■
NF*NF	■	■	■	■	■	□	■	■	■	■	■	□
AF*AF	□	□	□	□	□	□	□	□	□	□	□	□
PM*PM	□	□	□	□	□	□	□	□	□	□	□	□
RM*RM	■	□	□	■	□	□	■	□	□	■	□	□
YM*YM	□	□	□	■	□	□	□	□	□	■	□	□
SF*SF	■	□	□	□	□	□	■	□	□	□	□	□
NF*AF	□	□	□	□	□	□	□	□	□	□	□	□
NF*PM	□	□	■	□	□	□	□	□	■	□	□	□
NF*RM	□	□	□	■	■	■	□	□	□	■	■	■
NF*YM	□	□	🟥	□	🟥	□	□	□	□	□	□	□
NF*SF	□	□	□	□	□	■	□	□	□	□	□	■
AF*PM	□	□	□	□	□	□	□	□	□	□	□	□
AF*RM	□	□	🟥	□	🟥	□	□	□	□	□	□	□
AF*YM	□	□	□	□	□	□	□	□	□	□	□	□
AF*SF	□	□	□	□	□	□	□	□	□	□	□	□
PM*RM	□	□	□	■	■	□	□	□	□	■	■	□
PM*YM	□	□	□	□	□	□	□	□	□	□	□	□
PM*SF	□	□	🟥	□	🟥	□	□	□	□	□	□	□
RM*YM	□	□	■	□	□	□	□	□	■	□	□	□
RM*SF	■	■	■	□	□	□	■	■	■	□	□	□
YM*SF	□	□	□	□	□	□	□	□	□	□	□	□

Figure 13. Response model terms retained (filled squares) and rejected (white squares) by the software system with no multicollinearity screening (on the left) and the one that screened for multicollinearity (on the right). Red squares indicate correlated regressors.

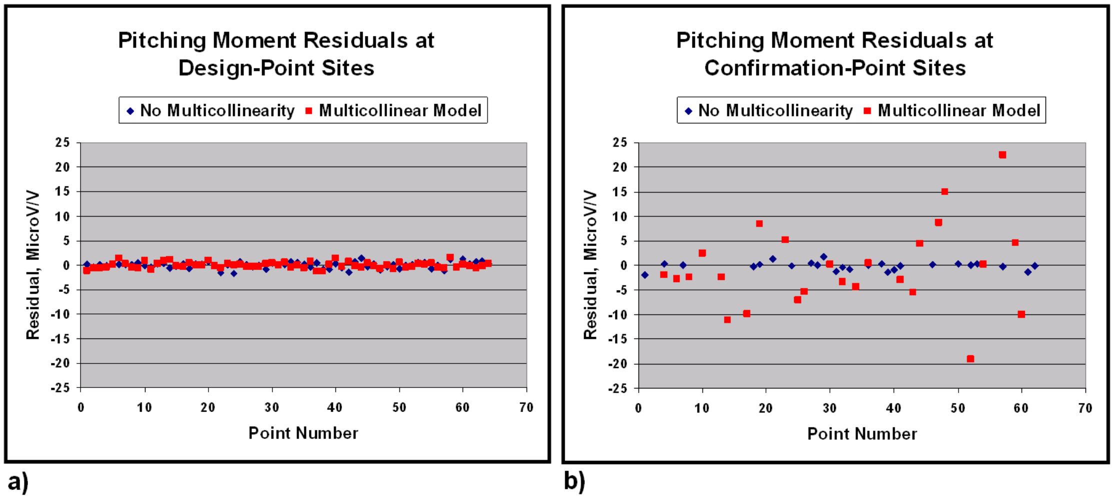


Figure 14. Pitching moment residuals with and without collinear regressors. a) No significant difference in conventional *design-point* residuals; b) Significant difference at off-design *confirmation-point* sites

The pitching moment and yawing moment models developed by the two software systems differed substantially. The software system with no multicollinearity screening retained all three of the correlated regressors in each model (red cells in Fig. 13), and failed in each model to detect as significant a first-order rolling moment term that was identified as significant by the other software system. (Note that significant first-order rolling moment terms were identified by both software systems for all other models). We now examine the impact of these model differences on response predictions made at the regression model design sites, and again at independent, randomly selected confirmation-point sites.

Figure 14 illustrates how multicollinearity affected the pitching moment response models. Figure 14a reveals that when the residuals are obtained at model design-point sites, no significant difference can be detected between the model that retained correlated regressors and the model that rejected them. Fig. 14b tells an entirely different story, however, when the residuals are obtained at random confirmation points. The confirmation-point residuals are quite large for the model that retained the correlated regressors, but they remain quite small for the model with no significant multicollinearity. The effect of multicollinearity is clear for the pitching moment response models. When multicollinearity is in play, good fits are apparently only achieved for the data used to fit the model, and not for general combinations of independent variable levels.

Figure 15 displays the effect of retaining correlated regressors in the yawing moment response model. We see the same effect as in the pitching moment model. The pitching moment response model with correlated regressors seem capable of predicting pitching moment reasonably well only at the design-space sites where data were acquired to fit the model. It predicts relatively poorly for more general combinations of independent variables. The pitching moment model without correlated regressors appears to predict responses relatively well at factor combinations for which the regression data were acquired, but also at other sites within the design space. This is, of course, the kind of robust predictive behavior that we seek.

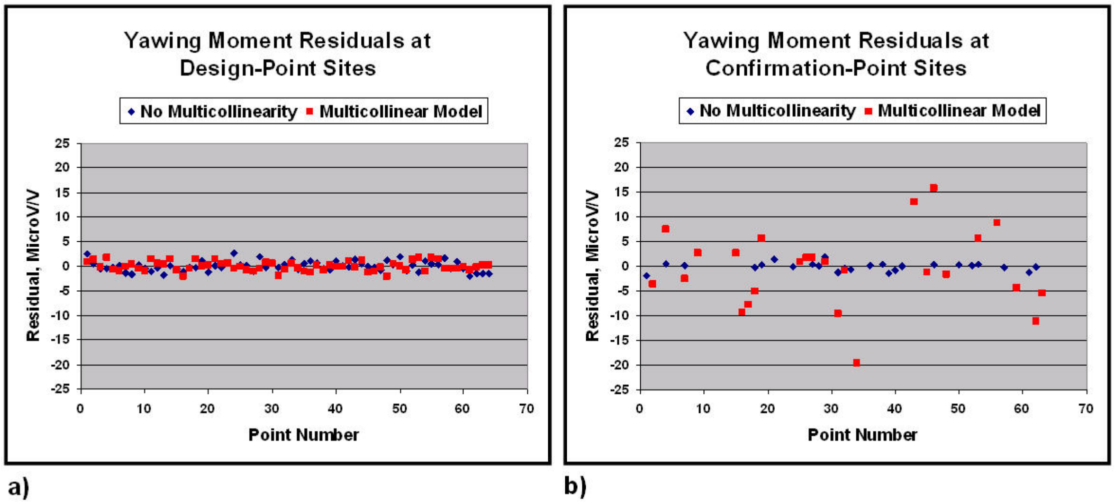


Figure 15. Yawing moment residuals with and without collinear regressors. a) No significant difference in conventional *design-point* residuals; b) Significant difference at off-design *confirmation-point* sites.

The simplified example used earlier to describe the impact of multicollinearity on response surface model construction was intentionally contrived for pedagogical reasons, but the current example is one in which undetected multicollinearity could in fact have a severe negative impact in an actual experimental situation. As in the simple example presented earlier, the root of the problem is that correlated regressors inflate the variance associated with estimates of the regression coefficients, increasing the uncertainty in estimating them. The result is a specific model that is simply drawn from a large family of candidate models, many differing substantially from each other. Substantial prediction uncertainty can result when such a model is applied to predict system responses for a specific combination of factor levels.

IV. Critical Binomial Analysis of Residuals

This paper has focused on the importance of confirmation-point residuals for assessing the adequacy of candidate response models. Regardless of where within the design space the residuals are obtained, it is useful to have some objectively quantitative process for determining when the residuals are constrained within acceptable limits. On the face of it, this does not seem to be a terribly difficult idea to implement. There is a prediction interval centered on the response level predicted by the model for a given combination of independent variables, within which a measured value is expected to fall with a prescribed probability. We can specify the combination of independent variables for which we wish to test the model, compute the upper and lower limits of a prediction interval at this point, make a measurement there, and see if the measurement falls within the prediction interval or not. If it does, this would provide evidence of an adequate model, and if it does not, this would seem to be evidence of an inadequate model.

The problem with this idea is that, as noted, there is a *probability* associated with any prediction interval. We speak of the “95% prediction interval,” for example, which describes a range of predicted response levels within an individual measurement is expected to fall 95% of the time. There is thus an expectation of a failed test for one trial in twenty on average, even for an adequate model. Failures may suggest some imperfection in the model, but they may also simply reflect experimental error in the measured confirmation points. To account for the stochastic nature of the problem, we may therefore simply resolve to define model adequacy in terms of the percentage of residuals that lie within their corresponding prediction intervals, allowing for some realistic number of failed tests, but with some expectation of a specified fraction of successes if the model is indeed capable of making reliable response predictions. This is in fact the general procedure that is advocated in this paper. We simply have to ask how many successes are necessary for a given number of trials to pronounce the model adequate.

The answer to this question is rather more elusive than it would seem it ought to be at first glance. After all, if we are using 95% prediction intervals to test the model, within which our claim is that there is a 95% probability that a measured point will fall, can we not simply define model adequacy in terms of a 95% success rate? That is, if we employ 100 confirmation points to test the model, are we not justified in demanding that 95 of these tests be successful? The answer to this question, regrettably, is “no.”

The problem with demanding 95 successes out of 100 when success is defined as a measurement that falls within a 95% prediction interval is that in this instance “95” is merely the *expectation value* for the number of successes. That is, if we conducted a large number of similar evaluations of an adequate model, we would expect to find 95 successes out of 100 more often than any other number of successes. But we would find in such a series of evaluations that 94 successes occurred almost as often as 95 successes, even when the model is perfectly adequate. Would we seriously reject a response model that predicted response levels within a prescribed tolerance level 94 times out of 100 but not 95 times out of 100? As a practical matter, a relatively small nudge to the limits of a 95% prediction interval would be sufficient to expand it incrementally to a 94% prediction interval, which would then allow the model to “pass the test” rigorously. In many circumstances, such an incremental increase in the prediction interval limits would have no practical impact on the quality of the model prediction.

This question of the minimum number of successful trials that are necessary to assert the adequacy of model can be grasped rather easily by asking how many times out of 100 trials a coin has to land on “heads” to be declared “fair” (that is, not weighted). The expectation value is 50 (neglecting pedantic appeals to account for the probability that it will land on its edge), but surely a fair coin could easily land on heads 49 times out of 100, or 48, or 47. If in never landed on heads, or did so only one or two times, then clearly we would say it is weighted. So something in between is a rational criterion. We can appeal to the binomial probability distribution for such a criterion.

The coin toss problem and the prediction interval test are both examples of a binomial process that is characterized by a specified number of trials, an assumed probability of success that is the same for each trial, and, and an observed number of successes. There is a probability distribution that describes this process, which has in common with other probability distributions a family of Critical Values corresponding to specified probability levels. In the case of the coin toss problem, the critical value associated with 95% confidence is 42 when there are 100 trials for which the probability of success on any one trial is 0.5. This means that in 100 tosses of a fair coin, there is a 95% probability that not less than 42 (and by symmetry, not more that 58) of the tosses will be “heads.” If you actually did flip a coin 100 times and found that there were fewer than 42 “heads” (or, again by symmetry, fewer than 42 “tails”), you would be entitled to infer that the coin did not have a 50% probability of landing on “heads,” with no more than a 5% probability of an inference error. We call “42” in this example the “Critical Binomial Number.” Critical Binomial Numbers are tabulated in standard statistical references, or they can be computed with readily available software. The CRITBINOM workbook function in Excel returns the Critical Binomial Number, for example (=CRITBINOM(100,0.5,0.05) returns the value 42).

To apply a critical binomial analysis in response model adequacy testing, we simply require a critical binomial number of successes in a specified number of trials, where “success” is defined as a confirmation point that falls within a given prediction interval. To evaluate the software systems described in this paper, we used 25 residuals (25 total trials) in conjunction with a 95% prediction interval. Therefore our claim (for an adequate model) was that the probability of success on any one trial was 95%. We wished to make inferences about the adequacy of the models we tested with which we could be 99% confident. The corresponding Critical Binomial Number is 21. The process, then, was to count how many confirmation points out of a total of 25 fell within the 95% prediction interval that was computed individually for each combination of confirmation-point factor levels. (This interval varies with location in the design space, tending to be somewhat larger than its average near the design space boundaries and somewhat smaller near the center of the design space.) Any model with fewer than 21 successes could be declared inadequate, in the sense that there was less than a 1% chance that the true 95% prediction interval was as narrow (precise) as we claimed.

This begs a further question, which is one about how much precision is actually required to meet the objectives of an experimental investigation. We address this question in the following section.

V.Aleatory and Epistemological Uncertainty in Response Model Adequacy Assessment

There is always some uncertainty as to the true dispersion in an experimental result since estimates are inevitably made from finite samples of measurements, each carrying some degree of experimental error. This issue has been addressed on some level in the prior section, but there is in fact another element of uncertainty that impacts (or *should* impact) the assessment of response model adequacy. That is the fact that even for the most unambiguously documented precision requirements, there necessarily remains some uncertainty in the statement of the requirements themselves. We simply acknowledge here the distinction between *aleatory* uncertainty, dealing with the kinds of chance variations that characterize individual measurements in an experiment and about which the bulk of this paper has been concerned, and *epistemological* uncertainty, which has been defined in an engineering context[18] as a "lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a specific application."

It has been the author's experience for over forty years that experimentalists often display an almost congenital inability to articulate specific precision requirements for a test they plan to undertake. When asked how good a result is required, in at least half of the encounters the author has experienced in his career, the experimentalist does not even seem to understand the question. The most common responses range from a blank stare to an aggressively vacuous assertion that NO uncertainty is acceptable in a test as important as this one. On those occasions when the experimentalist does realistically comprehend the basic concept, specific assertions of precision requirements remain hard to extract. One is likely to hear that the experimentalist requires "the best that can be done," or that response models should predict responses "within experimental error," with no indication of the level of error that would be acceptable. These vague assertions make it difficult to implement an exit strategy based on achieving specific objectives.

It often seems that the experimentalist, for whatever reason, simply does not have a well-formed idea of what his specific precision requirements actually are. This epistemological uncertainty should be taken into account, along with the aleatory uncertainty associated with ordinary experimental error, in assessing the adequacy of a response model. It is not especially useful to assess whether a response model is capable of predicting drag coefficient in a wind tunnel test to within one count of drag with 95% confidence, if the "one drag count" requirement is at best vague, and at worst, entirely arbitrary.

Figure 16 presents a set of residuals associated with a pitching moment response model developed by the author from data acquired in a recent wind tunnel test. The red lines represent tolerance levels specified by the principal investigator, who required that the response model be able to predict pitching moment "within typical experimental error." Further discussion lead to a consensus decision that the "experimental error" tolerance level would be defined as twice the standard deviation in a series of pitching moment replicates that had been acquired in a different wind tunnel test, and on a test article that was not the same as, but was generally similar to, the one for which the data in Fig. 16 were acquired. (This reliance on data from another test was necessitated by the lack of comparable replicates in the test being analyzed.)

The Critical Binomial Number for these 41 confirmation-point residuals was 35, assuming as a success criterion a 95% probability that a given residual was within tolerance, and that a 99% confidence level was required to infer an inadequate fit. Eight of the residuals in Fig. 16 are outside the tolerance limits, so the 33 successes achieved in 41 trials was insufficient to declare the model adequate in this test.

The green lines in Fig. 16 represent adjusted tolerance limits for which the critical binomial number test would have been successful and the model declared adequate, as these lines do encompass the Critical Binomial Number of 35 residuals. Moving the tolerance limits from the red lines to the green lines represents a very small adjustment, and as a practical matter it is unlikely that an experimenter who would be satisfied by a precision interval represented by the red lines would find the green lines totally unsatisfactory. This would be true in any case, but especially in the current not-uncommon situation in which the tolerance limits were so vaguely specified ("within typical experimental error") and so imprecisely quantified (based on replicates from a different test article in a different wind tunnel test). Figure 16 illustrates the role of epistemological uncertainty in the adequacy assessment of a response model; adequacy decisions must often be made when we cannot rationally identify a make-or-break acceptance criterion.

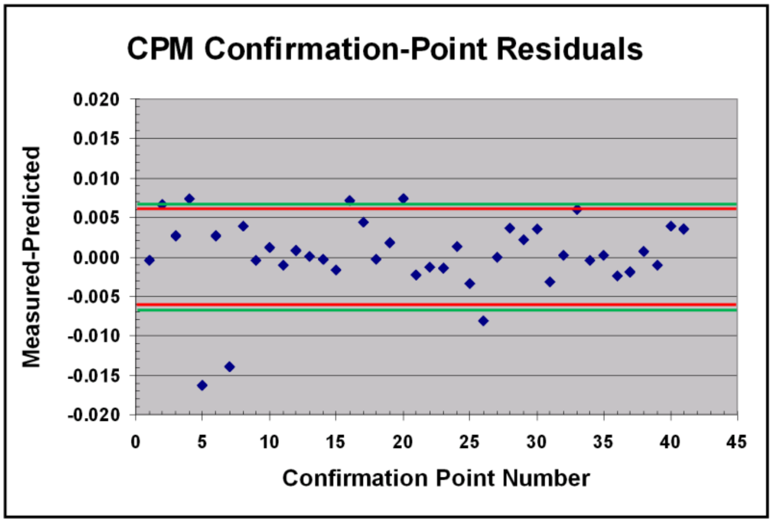


Figure 16. Confirmation-point residuals for wind-tunnel pitching moment response model. Red line: "Experimental Error" tolerance level. Green line: Critical Binomial Number limits.

Sophisticated methods have been proposed to account for epistemological uncertainty in experimental research[18], with much of it based on the Dempster-Schafer evidence theory extension of Bayesian subjective probability[19-20]. However, a considerably less sophisticated analysis may be sufficient for the specific case of response model adequacy testing through residual analysis, as illustrated in Fig. 16. We have in this instance a specified variance criterion for which we have some uncertainty in the specification. Again we emphasize that this uncertainty is epistemological and not aleatory; it is not that there is uncertainty in the *estimate of the variance* that is of concern here (although surely there is such uncertainty), but rather, it is that there is uncertainty in the *specification of the level of acceptable variance*.

Since we are dealing with uncertainty in a specification of acceptable variance, it is useful to consider a reference distribution that governs how the variance is distributed under a null hypothesis that the true variance specification is what we assume it to be. In this instance, the null hypothesis is that the red lines in Fig. 16 truly do represent an interval with a half-width equal to twice the pitching moment coefficient's population standard error, σ_0, for this tunnel entry. The chi-square statistic is useful for testing this hypothesis:

$$\chi^2 = \frac{n-1\ s^2}{\sigma_0^2} \tag{6}$$

where s is a sample standard deviation based on n-1 degrees of freedom. This statistic has the usual chi-square probability distribution. The standard error in pitching moment upon which the red line interval in Fig. 16 was based was computed from a sample of n=10 replicates, which corresponds to the mean of the 9-degree-of-freedom chi-square probability distribution shown in Fig. 17.

Critical values for the 0.05 significance level are marked with dashed lines in this figure, and the value of χ^2 for the pitching moment variance that just passes the Critical Binomial Number Test (green lines in Fig. 16) is labeled "CPM." The critical values in Fig. 17 encompass an interval within which χ^2 values cannot be distinguished from the mean of this distribution with at least 95% confidence. We therefore infer that there is no significant difference between the standard error in pitching moment initially specified as a tolerance level for the Critical Binomial Number analysis (mean of the distribution), and the value needed to pass the Critical Binomial Number test (χ^2 value labeled "CPM"). This objective inference is in harmony with the intuitive expectation articulated earlier, that there could be no practical difference in the red and green precision intervals of Fig. 16.

The critical χ^2 values are tabulated in standard statistical references for specified significance levels and number of degrees of freedom. By inserting these values for χ^2 in Eq. (6) along with σ_0, the standard error in the n=10 pitching moment replicates used to estimate typical experimental error initially, one can compute a range within which standard error values are statistically indistinguishable from σ_0. This range comprises an uncertainty band about the upper and lower tolerance limits that reflects the uncertainty in estimating those limits, as Fig. 18 illustrates. Note that the red and green tolerance limits both fall within the uncertainty bands, suggesting one of those limits is as good as the other. That is, the limits for which the residuals pass the Critical Binomial Number test (green lines) are statistically indistinguishable from the original estimates of confidence interval limits (red lines).

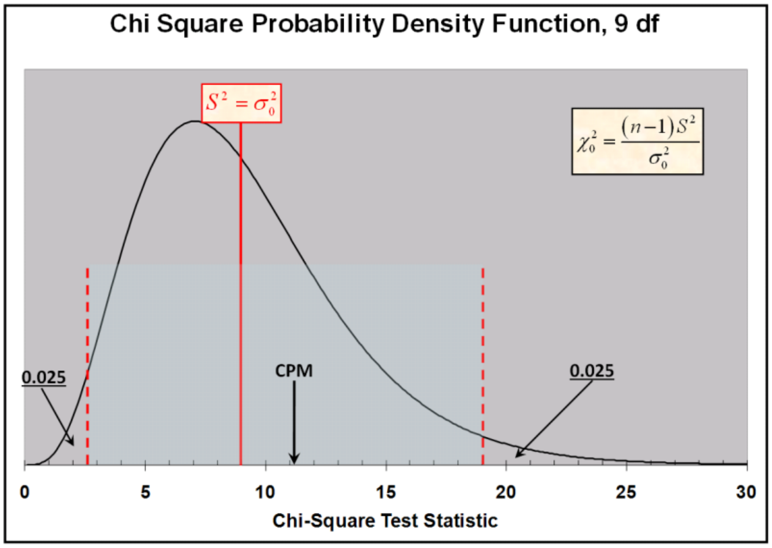


Figure 17. Nine degree of freedom chi-square distribution showing critical values for the 0.05 significance level and value of χ^2 for pitching moment variance that passes the Critical Binomial Number Test

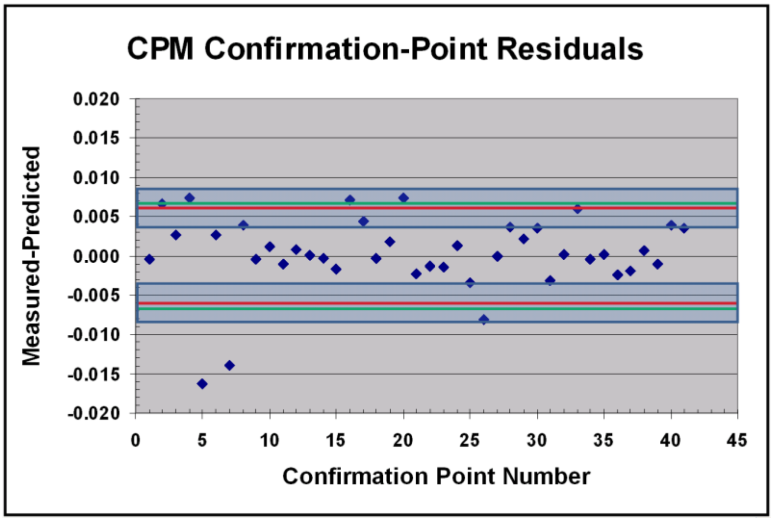


Figure 18. Confirmation-point residuals for wind-tunnel pitching moment response model. Bands reflecting epistemological uncertainty in specifications of tolerance limits

VI. Discussion

A number of miscellaneous discussion points will be touched upon briefly in this section. See the earlier text for more detail on these points.

This paper has focused on the distinction between confirmation point residuals and residuals obtained from regression model design sites. The general conclusion is that confirmation point residuals provide a more stringent test of the adequacy of a response model. They are generally as large as or larger than design-point residuals and can be substantially larger, especially when regressors are correlated. They also provide a more realistic test of the general function of the response model, which is to adequately predict responses at arbitrary combinations of factor levels within the design space and not just at the sites where the data were acquired that produced the models.

The experimenter should resist the temptation to offer his response models only in the most favorable light, by limiting residual analysis to the design space sites where such an analysis is likely to minimize indicated prediction uncertainty. By the same token, it is not especially useful to limit residual analysis only to the most challenging factor combinations—the extreme corners of the design space, for example. Such a test would roundly overstate the general prediction uncertainty of the model. The author recommends a policy of selecting confirmation-point residuals at random, which ensures a mixture of challenging and easier tests, and eliminates the natural human tendency to want to portray the model in either an especially good or an especially critical light.

In the analysis of balance calibration residuals obtained with an automated balance calibration machine, one advantage of confirmation point residuals over design point residuals that was more or less serendipitously discovered was that bias shifts could be detected in confirmation point runs acquired either before or after the calibration runs. Calibration residuals obtained through linear regression are guaranteed to feature a mean of zero but confirmation residuals are unconstrained and can feature non-zero means. For small samples of confirmation-point residuals a non-zero mean can be attributed to design-space site selection, but when nearly two thousand confirmation points are more or less symmetrically distributed over the design space as in this study, relatively large non-zero means are more likely to be associated with systematic biases in the confirmation-point data relative to the calibration data.

Bias shifts in an automated balance calibration may be due to small alignment differences from run to run. Alignment errors are not expected to introduce within-run variance and are therefore difficult to detect in a single calibration run. But if data acquired in other runs with the same balance are used to obtain confirmation point residuals, a component of variance is in fact introduced if the alignments are not identical across runs, and this can be detected and quantified. Such bias shifts highlight the disadvantages of repeat points over genuine replicates. Any set point errors due to slight misalignments become fossilized in all points acquired for a given machine setup, with no opportunities to cancel.

This characteristic of automated balance machine calibration does not represent a particularly serious problem if it is recognized, and if quality assurance tactics are adopted to compensate for this effect. The calibration test matrix should be organized in a series of blocks, with the balance realigned between blocks to intentionally induce any bias shifts attributable to any alignment changes. The number of blocks should be great enough to provide a statistically significant number of degrees of freedom to assess contributions of block effects to the unexplained variance. Loadings should be balanced within blocks as well, and randomized. Replicates consisting of identical loading combinations in different blocks can provide a direct estimate of block effects.

In an illustration used in an earlier section, correlated design matrix components were intentionally induced by improperly matching a proposed response model to the data acquired to fit it. While this example was intentionally exaggerated for pedagogical purposes, the same potential ambiguity always accompanies any mathematical representation of a finite sample of data. The true response behavior between measured data points is always open to question, no matter how large or small the intervening interval. This is another argument in favor of randomly selected confirmation points to test the adequacy of response models. Random confirmation points falling between regression design points can reveal lack of fit errors that could go undetected if residuals were limited to the design point sites.

There is a sense in which the distinction between aleatory and epistemological uncertainty with respect to response model adequacy assessment can seem redundant when the original tolerance limits for confirmation-point residuals are specified as “within experimental error.” In such a case the experimental error is commonly expressed as the product of a standard deviation of replicates and some coverage factor. The coverage factor depends on the number of degrees of freedom available to estimate the standard deviation, and reflects the fact that uncertainty in an estimate of variance is a function of the volume of data used to estimate it. The use of a coverage factor that depends on degrees of freedom may therefore seem to already account for the uncertainty in a specification of “typical experimental error,” in which case it may seem as if a further provision for “epistemological uncertainty” is unjustified.

The coverage factor reflects the uncertainty in *estimating* the experimental error. There is a further element of uncertainty, however, that represents how sure the principal investigator is that this is in fact the proper tolerance criterion in the first place. In the example treated in this paper, we computed tolerance levels by estimating the standard deviation in replicates acquired on a different model from a separate wind tunnel test. Even if that standard deviation estimate had been made with negligible uncertainty (if the calculation had been based on a very large volume of independent replicates, for example, so that the coverage factor approached an asymptotic limit), the question remains as to whether standard errors based on data from a different wind tunnel test comprise an appropriate tolerance level for the present test. In this paper we address this uncertainty by identifying a range of statistically indistinguishable variance estimates drawn from a distribution whose mean corresponds to our best estimate of the tolerance level. But even if that estimate is perfect, there can still be some ambiguity as to how appropriate it is, which is reflected in the uneasiness that many experimentalists feel when asked to formally declare their inference error risk tolerance levels.

The accounting for epistemological uncertainty was linked in this study with aleatory uncertainty in a way that, upon reflection, is probably not entirely appropriate. The variance distribution used to characterize epistemological uncertainty was defined to have the same number of degrees of freedom as the best estimate of the standard deviation in the residual variance. It is probably more appropriate to assess the epistemological uncertainty independently, in terms of a number of degrees of freedom that reflects our confidence ("belief in") the appropriateness of the specified tolerance level. If we assume that we know the tolerance level perfectly; that is, if our tolerance declaration rests effectively on an infinite number of degrees of freedom, then there is no epistemological uncertainty and the model adequacy assessment would be based entirely on the ordinary aleatory uncertainty that is reflected in the residual variance. This would be the case if the tolerance level was based on some specific certification limit, for example. In that case, the uncertainty bands in Fig. 18 could be quite narrow, approach a width of zero in the limit of absolute conviction as to the appropriateness of the tolerance declaration.

At the other extreme, for practical purposes the experimenter may have virtually no idea what tolerance limit to specify, in which case whatever tolerance he did declare could be assumed to rest on the equivalent of rather few degrees of freedom. The uncertainty bands in Fig. 18 could in that case be rather broad, and forgiving of fairly substantial deviations from a tolerance level specified with little conviction.

The specification of an equivalent number of degrees of freedom corresponding to some level of belief is beyond the author's ability, although an expert in Dempster-Shafer evidence theory, for example, may not find the task so daunting. This excludes the author, unfortunately, who at this stage is more comfortable recognizing and articulating this aspect of the problem than offering a rigorously defensible solution. The basic idea, however, is easy to grasp. When a given model adequacy tolerance level is specified with some doubt as to its appropriateness, then other tolerance levels that are statistically indistinguishable should be considered at least as suitable. In a time-honored tradition of scientific inquiry, the author identifies this as a topic for future consideration.

VII. Concluding Remarks

This paper has considered various aspects of response surface model adequacy assessment, highlighting the following broad themes:

1) The fact that both adequate and inadequate models can predict well at design sites while only adequate models can consistently predict well at off-design sites suggests that confirmation-point residuals are a more reliable indicator of model adequacy than design-site residuals.
2) Since especially poor off-design predictions can result from correlated regressors that substantially inflate the variance of regression coefficients and therefore add significant prediction error, considerable effort should be expended to reduce multicollinearity as much as possible.
3) The assessment of response model adequacy by an analysis of residuals can be treated as a Bernoulli process in which each residual is regarded as a trial for which success is defined by the size of the residual relative to some prescribed tolerance level. A Critical Binomial Number of successes is evidence that response predictions are within tolerance.
4) Experimenters often do not have a firm idea as to what constitutes adequate performance in a response model. The assessment of response model adequacy must account for this epistemological uncertainty as well as ordinary aleatory uncertainty that derives from the random nature of independent experimental error.

Acknowledgments

This work was supported by the Aeronautics Test Program Office of the National Aeronautics and Space Administration. Balance calibration data acquired with dead-weight loading were supplied by Dr. Peter Parker of NASA Langley Research Center. Balance calibration data acquired with an automated balance calibration machine were supplied by Dr. Norbert Ulbrich of Jacobs Technology, Inc, at NASA Ames Research Center. Dr. Ulbrich also supplied the calibration experiment design used in the analysis of these data. Certain wind tunnel data were supplied by Dr. Mark Kammeyer of the Boeing Company. The contributions of these individuals are gratefully acknowledged.

References

[1]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.

[2]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.

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

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

[5]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.

[6]DeLoach, R. “Formal Experiment Design as a Tool to Automate Aerospace Ground Testing (Invited),” Tenth Annual Spring Research Conference on Statistics in Industry and Technology, University of Dayton, June 4–6, 2003.

[7]Dowgwillo, R. M., and DeLoach, R., “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.

[8]Erickson, G. E., and DeLoach, R., “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.

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

[10]Montgomery, D. C., *Design and Analysis of Experiments,* 7th Ed., John Wiley & Sons, New York, 2009.

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

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

[13]Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, 2nd ed., John Wiley and Sons, New York, 2002.

[14]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.

[15]Design Expert, Software Package, Ver. 7.03, StatEase, Inc., Minneapolis, Minnesota, 2006.

[16]Parker, P .A., Morton, M., Draper, N., and Line, W., “A Single-Vector Force Calibration Method Featuring the Modern Design of Experiments,” *AIAA 2001-0170,* 39th Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[17]Beer, P. P., and Johnston, E. R. Jr., *Vector Mechanics for Engineers,* McGraw-Hill, 1962.

[18]Swiler, L.vP., and Giunta, A. A., “Aleatory and Epistemic Uncertainty Quantification for Engineering Applications,” Sandia Technical Report SAND 2007-2670C, Joint Statistical Meetings, Salt Lake City, Utah, July29–August 2, 2007.

[19]Dempster, A. P., “A generalization of Bayesian inference,” *Journal of the Royal Statistical Society, Series B* **30** 205-247. 1968.

[20]Shafer, Glenn, *A Mathematical Theory of Evidence,* Princeton University Press, 1976.